Gene Expression in *P. falciparum*: Statistical Patterns and Molecular Determinants

Jacob E. Lemieux

September 7, 2012

Contents

List of Abbreviations

3C — chromosome conformation capture
APC — allophycocyanin
BSA — bovine serum albumin
BSD — blasticidin S deaminase
cDNA — complementary deoxyribonucleic acid
ChIP — chromatin immunoprecipitation
ChIPseq — chromatin immunoprecipitation sequencing
DNA — deoxyribonucleic acid
EDM — euclidean distance matrix
eQTL — expression quantitative trait locus
FISH — fluorescence in situ hybridization
GCC — genome conformation capture
GFP — green fluorescent protein
H2A.Z — variant Z of histone H2A
H3K4Ac — acetylation at lysine 4 of histone H3
H3K4me3 — trimethylation at lysine 4 of histone H3
H3K9me1 — monomethylation at lysine 9 of histone H3
H3K9me3 — trimethylation at lysine 9 of histone H3
HP1 — heterochromatin protein 1
HPI — hours post invasion
ICAM-1 — inter-cellular adhesion molecule 1
IDC — intraerythrocytic Development Cycle
MDR1 — multidrug resistance protein 1
MLE — maximum likelihood estimate
mRNA — messenger ribonucleic acid
PBS — phosphate buffered saline
PCA — principal components analysis
PCR — polymerase chain reaction
PfEMP1 — *Plasmodium falciparum* erythrocyte membrane protein 1
PolII — RNA polymerase II
PSU – pathogen sequencing unit
PfSIR2 — *Plasmodium falciparum* silent information regular 2
qPCR — quantitative polymerase chain reaction
qRT-PCR — quantitative reverse-transcription polymerase chain reaction
QTL — quantitative trait locus
rDNA — ribosomal DNA locus
RMA — robust multiarray analysis
RT-PCR — reverse transcription polymerase chain reaction
SVD — singular value decomposition
TPE — telomere position effect

Acknowledgments

Many people have helped me in the course of pursuing a D.Phil. Without them, I could not have done the work described in these pages. To all who have assisted, encouraged, and supported me along the way—thank you, I am deeply grateful. In this short space I could not possibly thank everyone, but there are a few individuals whose contributions deserve to be singled out, and I would like to mention them here.

I would like to thank the members of the Newbold and Su laboratories for their assistance in lab, encouragement, and mentorship. In particular, I am grateful to Dr. Sue Kyes, Bob Pinches and Zoe Christodolou. Bob taught me how to grow parasites and assisted with all of the parasite culture work that I did at Oxford. Zoe explained and demonstrated many molecular biology techniques in addition to graciously fielding a four-year stream of questions on how to do things in the lab. Sue has been an exceptionally generous mentor and friend throughout my years at Oxford and NIH. She has offered expert advice and guidance in all aspects of my work. At the NIH, Dr. Richard Eastman and Dr. Maria Jose Lopez-Barragan were great teachers, friends, and valued collaborators.

The work on temporal estimation of gene expression was done in collaboration with Avi Feller and Professor Chris Holmes in the Oxford Department of Statistics and Dr. Natalia Escobar-Gomez and Professor David Conway of the MRC research station in Gambia. Their good nature, expertise, and enthusiasm for engaging with a new topic made the collaboration an extremely enjoyable one on both personal and professional levels. In addition to a collaborator, Avi has also been a close and valued personal friend throughout my D.Phil.

The studies of chromosome folding were performed in collaboration with the Pathogen Sequencing Unit at the Sanger Center. I would like to thank Dr. Thomas Otto, Dr. Matthew Berriman, and Dr. Michael Quail for their numerous contributions in the areas of sequencing including genome assembly, short-read mapping, and library preparation and sequencing. I would also like acknowledge

with gratitude Professor Artur Scherf and members of Scherf laboratory, particularly Dr. Jose-Juan Lopez-Rubio, for teaching me FISH and for hosting me at the Institut Pasteur in Paris.

Many thanks are due to Dr. Benedikt Kessler and Dr. Cynthia Wright of the Oxford Proteomics facility, who provided support and assistance with proteomics early in my D.Phil. Because of time and space constraints, not all of that work was not able to be included in this thesis in completed form; however both Benedikt and Cynthia devoted many hours of assistance on my behalf and I am grateful for their efforts.

I would also like to thank my parents for their constant love, support and encouragement throughout my D.Phil.

Finally, and most of all, I would like to thank my supervisors, Professor Chris Newbold and Dr. Xin-zhuan Su, for their constant encouragement, guidance, and mentorship. Attempting to pursue doctoral research split between two labs at institutions separated by the Atlantic ocean is not the type of experiment which is guaranteed to work. Nevertheless, I think it did work, and it was great fun in the process, thanks in large part to the patience, dedication, and generosity of Chris and Su. For the past four years, they have supported my research and my life in ways which are too numerous to count or list here. It has been a privilege to work under their joint supervision, and I thank them for their great effort and kindness in shaping this work and and my education.

To reflect the contributions of these individuals, I use the pronoun 'we' in the remainder of the thesis.

Abstract

This thesis investigates patterns and mechanisms of gene expression in *P. falciparum*. The rapidly cycling patterns of genes during the asexual stages confounds the analysis of gene expression in culture and in patients. In order to overcome this problem, we develop statistical models to estimate the temporal progression of malaria parasites by using the observed gene expression values and known reference sets. We extend this framework to account for lineage commitment, and show that, similar to asexual development, it is also possible to recover information about parasite sexual differentiation given observed gene expression values. Using datasets from our own lab as well as those available in the literature, we establish that the patterns of expression in patients are similar to those observed in culture but in additive mixtures whose proportions can vary between patients.

We then investigate epigenetic and spatial factors that are important in regulating gene expression. Using second-generation DNA sequencing, we generate genomic maps of chromatin state and chromosome interactions. These maps provide the first global view of chromosome folding and locus-specific affinities in malaria. They also highlight the importance of epigenetic modifications in imposing structure on the spatial organization of the genome. After generating an initial set of genomic maps, we apply these tools to study chromatin reconfigurations during *var* gene switching. We show that when the active *var* gene changes, reconfigurations can be seen in the local three-dimensional chromatin structure of the activated locus.

Overall, our results contribute new methods for analyzing malaria microarray data, highlight the importance of lineage mixtures in patient infections, generate an atlas of spatial interactions in the nucleus at a resolution of approximately 5 kilobases, and establish a link in malaria between the spatial configuration of active loci and the local chromatin environment.

Chapter 1

Introduction

Malaria is an infectious disease caused by parasites of the genus *Plasmodium*. It has been called "the most important parasitic disease of man," [19], a reputation earned because its of ubiquity and lethality. The parasite is widespread in tropical climates, causing destruction on the scale of 1 million deaths and 200 - 500 million cases per year [107, 138]. The prevalence and severity of malaria can be explained at least in part by its biology. Adapted to generate chronic infection of vertebrate hosts in order to allow for transmission via mosquito vectors, many *Plasmodium* species remain virulent enough to severely sicken or kill their hosts.

As mentioned in the abstract, this thesis investigates the biology of *Plasmodium falciparum* at the level of gene expression. During its complex lifecycle, the *P. falciparum* cell undergoes dramatic changes in its form and function, including antigenic variation to evade the host immune response, yet the cell retains the same DNA during this process. How is this possible? While the same molecular instructions are always present, they are read or "expressed" differently at various points in the lifecycle. Only some genes are active during the asexual stages, while others are turned on in gametocytogenesis. How and why are individual genes expressed? We work on those questions in this thesis, focusing first on statistically describing gene expression, and then on the molecular mechanisms that control gene expression, particularly chromatin state and chromosome folding.

In order to put our research in context, we begin this thesis with comments about history of the disease. From there, we proceed to the molecular biology of gene expression and introduce some of the problems studied in this thesis.

1.1 Background

1.1.1 History

Descriptions of malaria go back well into antiquity. Hippocrates describes the disease in his work *Epidemics* [25]:

> *"When the paroxysms fall on even days, the crises will be on even days; and when the paroxysms fall on odd days, the crises will be on odd days. Thus, the first interval of those with crises on even days is on the fourth day, the sixth day, the eighth day, the tenth day, the fourteenth day, the twentieth day, the twenty-fourth day, the thirtieth day, the fortieth day, the sixtieth day, the eightieth day, and the one hundred and twentieth day. While those with crises on odd days, the first interval is on the third day, the fifth day, the seventh day, the ninth day, the eleventh day, the seventeenth day, the twenty-first day, the twenty-seventh day, and the thirty-first day. Furthermore, it is necessary that one know that if crises fall on days other than those mentioned above, there will be a relapse, and this may be deadly. But it is essential to pay attention and know at which times the crises will lead to death and in which to recovery, or during which is there tendency to fare better or worse. The intervals when crises occur in irregular fevers, quartans, quintans, septans and nonanes, should also be considered.*
> – Hippocrates (*Epidemics*) (quoted in [25]).

The periodic fevers characteristic of malaria were known into the modern times and went under the name of 'agues' in England, 'paludisme' in France and 'malaria' in Italy. The latter name—meaning literally "bad air"—was coined because malaria was attributed to the stale air that permeated marsh areas [30]. Following the work of Pasteur and Koch in outlining the microbial theory of disease and identifying organismal causes of several infectious diseases, the *Bacillus malariae* was discovered in 1879 [136]. Klebs and Tommasi-Crudelli postulated that *B. malariae* was the cause of malaria; the rod-shaped bacillus was later identified in the bodies of deceased malaria patients. The bacterial cause of malaria was rapidly accepted and when Laveran, a student of Pasteur working in Algeria as a physician, reported in 1880 that he observed live parasites in the blood of malaria patients, the scientific community reacted to his discovery with skepticism.

Controversy ensued [136], but by the end of the 1880's, with repeated failure of the *B. malariae* to cause malaria in animal models, consensus began to build around Laveran's discovery. Nevertheless, if a parasite was responsible for causing malaria, how was it transmitted? At the encouragement of Manson, Ross went to India in 1897 to study potential routes of mosquito transmission. In a series of experiments that took place in the years 1897 - 1901, Ross demonstrated that the mosquito was the vector of transmission between hosts. Manson soon confirmed these findings himself using a cohort of human volunteers which somewhat astonishingly included his son. In recognition of this work, Ross was awarded the Nobel prize in 1902.

With the turn of the twentieth century, understanding of the lifecycle was mistakenly believed to be complete. The question of how a sporozoite went from the mosquito salivary glands into erythrocytes had been ostensibly settled by Shaudinn's discovery of a sporozoite invading an infected red blood cell. Nevertheless, the timing of the onset of blood stage parasitaemia in volunteers, and the crucial fact that the propensity to relapse depended on whether volunteers were inoculated with blood stage parasites or via mosquito sporozoites, maintained the possibility of exo-erythrocytic stages [54]. These were later identified by James and Tate in avian malaria in 1937 and Schortt and Garnham in primate malaria in 1948 [133].

The modern age of malaria research began with two major events. The first was the discovery of continuous culture of malaria parasites by Trager and Jensen in 1976 [145]. The second, completed in 2002, was the decoding of the malaria genome [53]. While the function of most genes remains obscure, the base sequence of all the genes of the parasites is now known, and having this information adds greatly to researchers' ability understand the molecular details of parasite functions.

In recent years, scientists have been pursuing two convergent paths toward deciphering the function of parasite genes. The first is directed studies to identify individual genes that underlie particular phenotypes, often using a combination of classical genetics and molecular cloning. Through this approach, the genetic basis of many of the most important traits has been identified, such as genes responsible for antigenic variation [6, 137, 141], drug resistence [147], and invasion [118]. This approach is accurate, reliable, and reproducible, but can be slow. Tracking down the function of a single gene can take years or even decades. A second approach, made possible by the completion of the malaria genome project, is to measure a large number of variables about all of the genes in the genome and attempt to guess at their function through an inferential process. This approach often uses technologies such as microarrays, bioinformatic analyses, and high-throughput screens to measure expression or the response of the parasite under various chemical or genetic perturbations. In some ways, the field remains in flux, because it is not always clear what to do with the information generated in a high-throughput experiment. Converting the results of high throughput experiments into biological insights remains a major challenge for biomedical research, and we struggle with it in this thesis.

1.1.2 Prevalence, Distribution, and Clinical Disease

Malaria is widespread throughout the developing world and remains endemic in many countries. *P. falciparum* is by far the best studied and is estimated to be responsible for approximately 515 million cases [138] and one million deaths each year. The top map in Figure 1.1 gives 2010 figures from the World Health Organization showing the spatial distribution of malaria cases around the world. The vast majority of these deaths occur in young children living in endemic regions, particularly sub-Saharan Africa. The number of deaths per 1000 populations is

shown in the bottom part of the Figure 1.1; that map reflects the great burden of malaria mortality in Africa. Severe, life-threatening *P. falciparum* frequently occurs in younger children who have not developed sufficient immune response to control the infection.

Most cases of *P. falciparum* malaria are "uncomplicated" or "mild". These individuals present with periodic fevers and flu-like symptoms accompanied by parasites in the blood. Symptoms typically resolve after antimalarial chemotherapy. In all cases, malaria requires urgent response. Recommended treatment is with artemisinin combination therapy or quinine, and chloroquine in areas where resistance to that drug is not a problem. In a minority of cases, the disease is severe and requires hospitalization with close clinical management. These cases of "complicated" or "severe" malaria include one or more of severe anemia, cerebral malaria, and respiratory distress. Particularly in mild cases, the disease is treatable and good clinical outcomes are often achieved with antiparasitic chemotherapy. Nevertheless, in severe malaria due to the *P. falciparum* parasite, complications such as coma and respiratory distress are associated with high mortality from malaria; greater than 10% of individuals who present with these symptoms ultimately die [97].

Perhaps surprisingly, while malarial disease implies a circulating asexual parasite population, not all individuals with detectable parasitemia have symptoms. In hyperendemic areas, where exposure to the parasite is frequent and previously exposed individuals have a high degree of acquired immunity, there are parasite-positive individuals without symptoms.

Many of the factors which underlie this broad spectrum of disease phenotypes, and which determine the severity of infection, remain unknown [100]. There is strong evidence from genetic and animal studies that host and parasite genetics play an important role. Genes which confer risk and protection from malaria are known in both animal models and humans. For example Hemoglobin S, C,

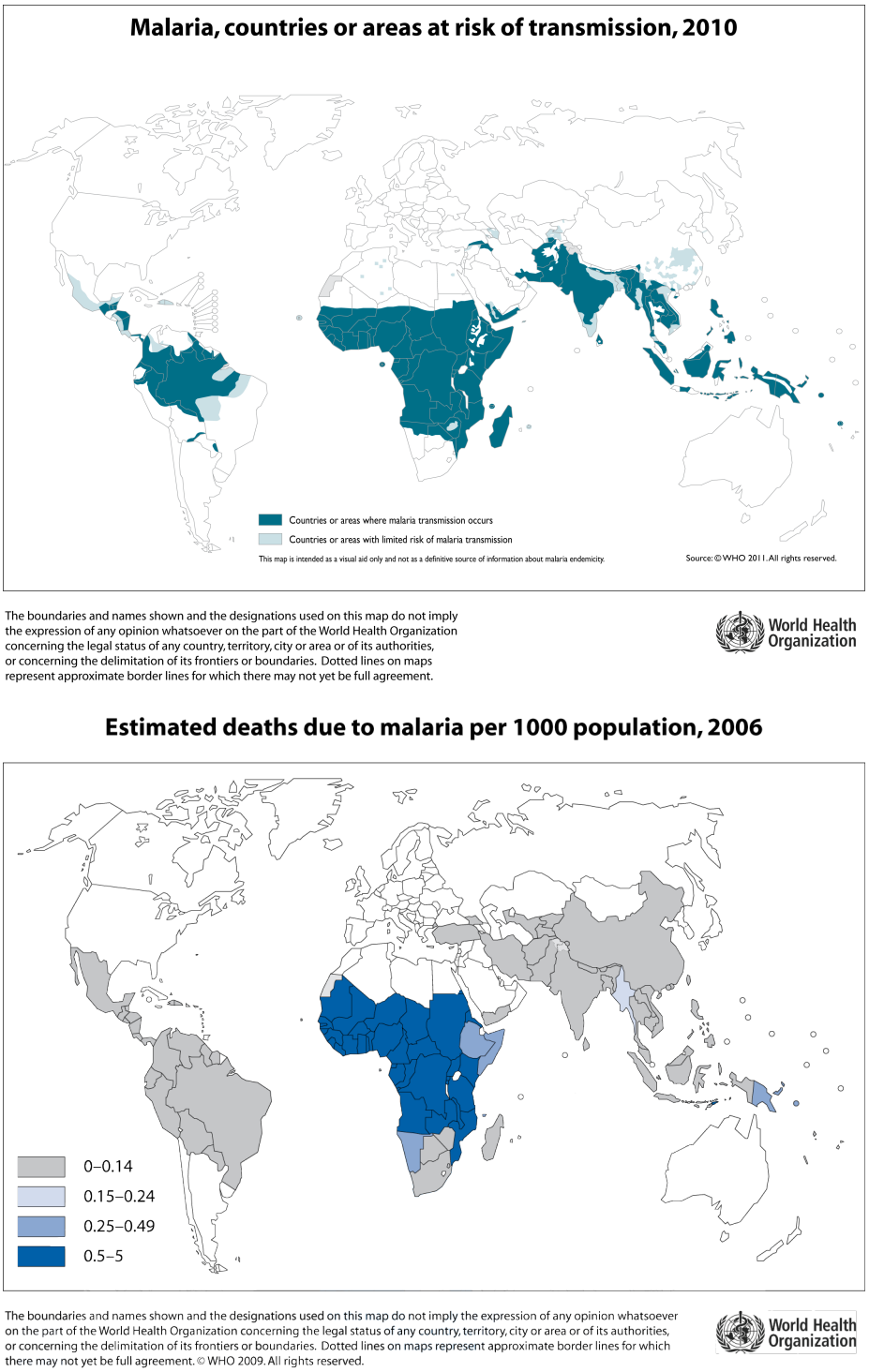

Figure 1.1: Data from the World Health Organization showing the geographic distribution of malaria transmission (top panel) and deaths (bottom panel). Approximately 2.2 billion people are at risk for Malaria [138]. The greatest number of deaths is observed in sub-Saharan Africa, where the disease is endemic, but substantial numbers of cases and deaths are present in Asia, the Pacific Islands, and Cenral and South America.

and E variants, Glucose-6-phosphate dehydrogenase deficiency (G6PD deficiency), and α^+ thalassemia yield malaria protective traits to their carriers [74]. Yet much of the heritable variation in protection remains unexplained by these genes, implying that a number of disease severity genes remain to be discovered [92], a scenario which also holds for parasite genes. A major goal of the microarray studies conducted in this thesis, discussed in Chapter 2, is to identify parasite genes whose expression was associated with severe malaria.

1.1.3 The Biology of *P. falciparum*

Parasites of the genus *Plasmodium* possess a complex lifecycle that involves a human host and a mosquito vector, as well as sexual and asexual stages. Morphological and biological changes occur as the parasite transitions from one life cycle stage to another. The lifecycle of *P. falciparum* is shown in Figure 1.2. Briefly, a mosquito injects sporozoites into the vertebrate host during a blood meal; the sporozoites then migrate to the liver where they infect hepatocytes. During this period, the infection is asymptomatic and the parasite undergoes numerous rounds of cell division. The hepatocyte ultimately ruptures and asexual forms, known as merozoites, are released into the circulating blood. The merozoites invade red blood cells, and once inside, go through a process of asexual cell division. Replication within the red blood cells takes approximately 48 hours and maturation is marked by morphological progression through a series of forms known as rings, trophozoites, and schizonts. Once the asexual division process is complete, the erythrocyte ruptures and releases free merozoites into the blood stream. The cycle repeats, leading to an exponentially amplifying parasitaemia.

While the majority of merozoites released by a rupturing schizont return to the asexual replication cycle, a fraction enter the sexual development pathway. These parasites initially look like rings when they reinvade red blood cells, but

then mature over a period of approximately 1 week to become gametocytes. Progression toward mature gametocytes is tracked through a stage numbering system consisting of stages I - V. Because this staging system is based on morphologically identifiable changes during gametocyte development, it does not include committed gametocytes prior to stage I which resemble asexual rings. Gametocytes are taken up by the mosquito during a blood meal, and once in the mosquito midgut they differentiate into male and female gametes. These gametes fertilize one another, forming a motile ookinete. The ookinete migrates through the mosquito peritrophic matrix and midgut epithelium, where it differentiates into an oocyst. The oocyst undergoes successive rounds of replication and ultimately ruptures, releasing thousands of sporozoites, completing the cycle.

Other parasites in the *Plasmodium* genus share nearly all the features of the *P. falciparum* lifecycle, but there are some minor differences. The simian parasite *P. knowlesi*, for example, possesses a 24-hour cycle while *P. malariae* and *P. ovalae* take 72 hours to mature in the red cell. Other malaria parasites, such as *P. vivax* and *P. cynomolgi* have exo-erythrocytic stages known as hypnozoites.

Of all five *Plasmodium* parasites which routinely infect humans, *Plasmodium falciparum* is the most virulent. This increased virulence relative to the other parasites that infect humans can be attributed to the high parasitaemias that are typically achieved by this organism as well as tissue-specific sequestration [7, 75, 99, 101], which is discussed in greater detail below.

1.2 Genes and Gene Expression in Malaria

In order to achieve chronic blood-stage infections, *P. falciparum* alters its dominant surface phenotype over timed by changing parasite-encoded proteins at the infected red cell surface, a process known as antigenic variation [75, 121]. The most important of these molecules is a family of proteins known collectively as

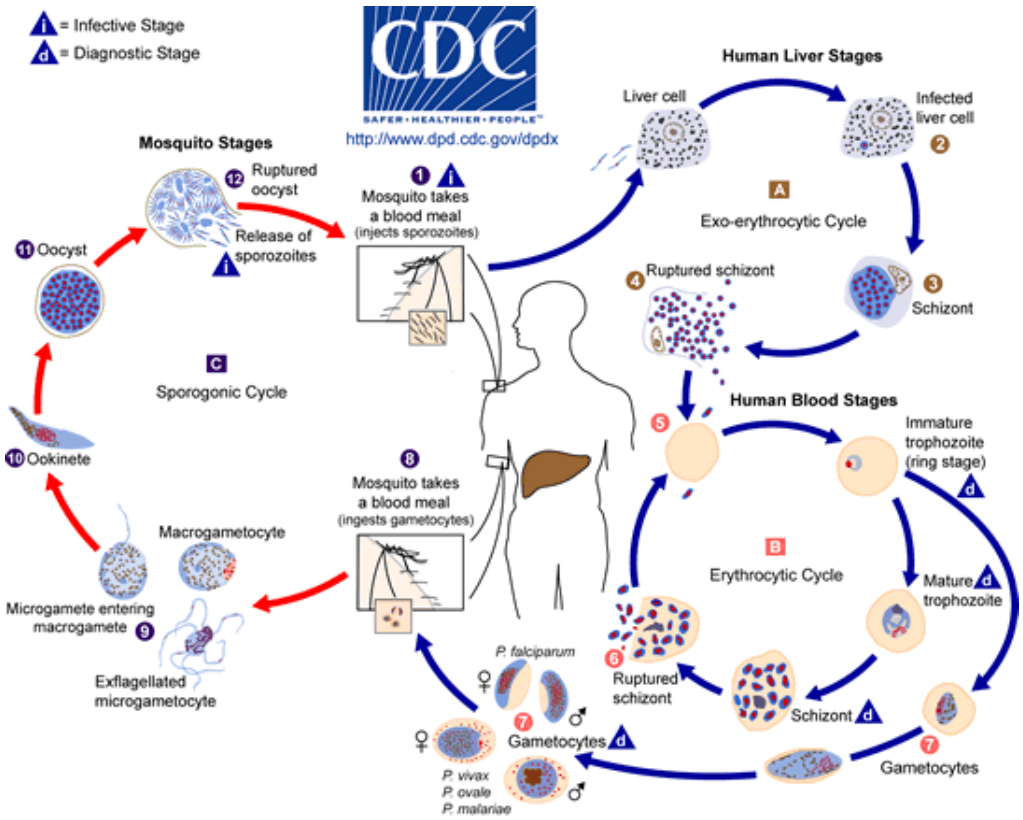

Figure 1.2: A schematic from the CDC depicting the various stages of the complex lifecycle of *P. falciparum*. The parasite posses a complex lifecycle that includes both sexual and asexual stages in the mosquito vector and human host, respectively. An infected mosquito injects sporozoites into a human. These sporozoites then rapidly migrate to the liver, where they infect hepatocytes and establish an asymptomatic liver infection. After successive rounds of cell division, infected hepatocytes rupture, releasing tens of thousands of merozoites into the blood stream, beginning the asexual replication cycle and causing the symptoms of malaria. During subsequent asexual replication, a portion of parasites becomes gametocytes, beginning development along a sexual lineage. Gametocytes are taken up by the mosquito and differentiate into male and female gametes, which mate, forming an oocyst. Sporozoites are again formed in the mosquito, completing the lifecycle. Lifecycle stages are summarized in the text.

Plasmodium falciparum Erythrocyte Membrane Protein 1 (PfEMP1) [83]. Single members of this family are exported by the parasite to the host red cell surface 18 - 24 hours post invasion (HPI) [93], where they mediate adherence to a variety of host molecules including CD36, thrombospondin, Inter-Cellular Adhesion Molecule-1 (ICAM-1), and chondroitin sulfate [5, 51, 52], thereby defining the tissue tropism of the parasite in the post-capillary venules (Figure 1.3). In addition to their function in vessel adhesion, PfEMP1 proteins are important targets of host immunity [13]. PfEMP1 is encoded by a large, diverse, multi-copy family of genes termed *var* [6, 141], and switches in the expression of these genes account for changes in surface phenotype [137].

Consequently, defining the mechanisms by which genes are turned on and off is essential to understanding malaria pathogenesis and the acquisition of blood-stage immunity to *P. falciparum*, and is a major focus of this thesis, particularly Chapters 4 and 5. Many of the details of this process have been worked out, yet others remain elusive (see below, Section 1.3). Current models highlight the function of epigenetic mechanisms, nuclear positioning, and paired-promoter silencing in maintaining the coordinated expression of a single *var* gene.

The genes encoding the surface protein PfEMP1, *var* genes, are among the best-studied genes in the genome, and it is not known to what extent the mechanisms of *var* gene regulation apply to other *P. falciparum* genes. The first genome-wide measurements of transcription in *P. falciparum* revealed a very strong stage-dependence in the expression patterns of individual genes [10, 61, 95]. Messenger RNA isolated from ring, trophozoite, and schizont forms showed substantial differences in abundance between these lifecycle stages. Further exploration of this pattern, with 1-hour time resolution by Bozdech, Llinas and colleagues [9, 87], showed, somewhat surprisingly, that this stage-dependence was underlaid more finely by a time-dependence in which the expression patterns of most genes is sinusoidal through the cell cycle. The period of most such genes is equal to the

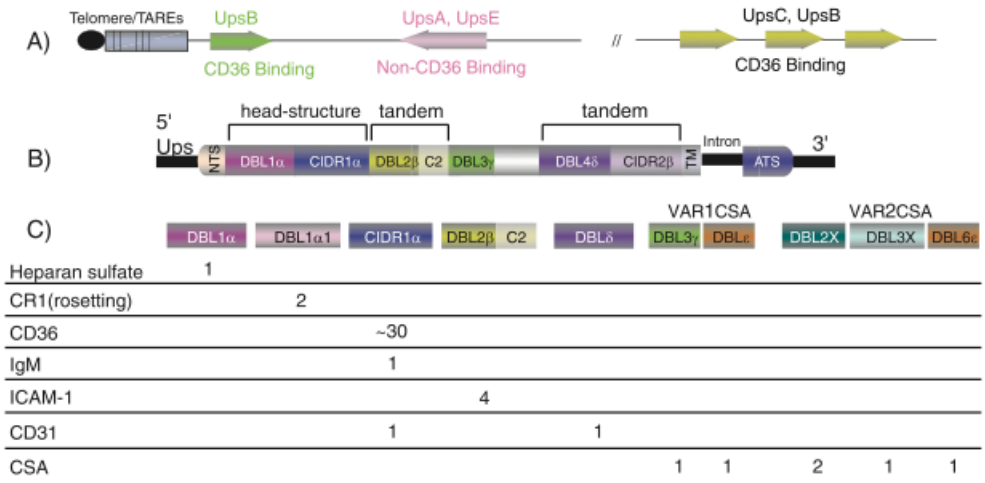

Figure 1.3: Figure 1 from Kyes et al. [78] showing the genomic organization and domain structure of *var* genes. The top panel shows the organization of a typical *P. falciparum* chromosome. A pair of *var* genes resides just distal to the telomere and are known as the subtelomeric *var* genes. Clusters of *var* genes can also be found in internal parts of chromosomes, often in repeats, as shown in the right part of the top panel. These are referred to as internal *var* genes. The proteins encoded by *var* genes, known as PfEMP1, have a complex domain structure that confers their binding specificity. These binding specificities mediate the cytoadherence pattern of mature *P. falciparum* parasites in post-capillary venule endothelium. The domain structure of a typical *var* is shown in the middle panel, and the known binding specificities of some domains are shown in the bottom panel.

length of the asexual replication cycle (Figure 1.4). It is not known what regulatory mechanisms account for this unusual program of gene activation, although a set of sequentially expressed specific transcription factors has been suggested to play a role [134].

In their seminal report [9], Bozdech and colleagues likened the transcriptional program of the malaria parasite to a "just-in-time" manufacturing process in which a component is fabricated just prior to its incorporation or use. Similarly, the authors were able to show by considering the expression pattern of groups of genes involved in particular biological processes that the malaria parasite appears to upregulate mRNAs shortly before the protein is required, and then to turn off transcription quickly.

The strong time-dependence of expression profiles presents unique analytical challenges, both technical and theoretical, for identifying differences in these patterns. Among the technical challenges, the primary one is that parasites must be perfectly synchronized between treatment and control groups if only a single time point is measured. Otherwise, differences in temporal progression between the experimental groups will confound the identification of true differences in gene expression. One way to compensate for this is to measure enough time points so that the underlying expression function through the lifecycle can be estimated reliably; however, this requires larger and more expensive experiments.

Furthermore, the number of time points which must be measured in order to reliably reproduce the underlying expression function, i.e. the appropriate sampling rate, is not clear. Under-sampling yields insufficient data to reconstruct underlying expression profiles and call differential expression, whereas oversampling is resource inefficient. There are few theoretical guidelines which can be applied in this case because the underlying signal depends strongly on the measurement device used, so except for the case of DeRisi-type [9,34] glass slide arrays, it is not known. Empirically, 6-8 timepoints seems to be the minimum to sufficiently re-

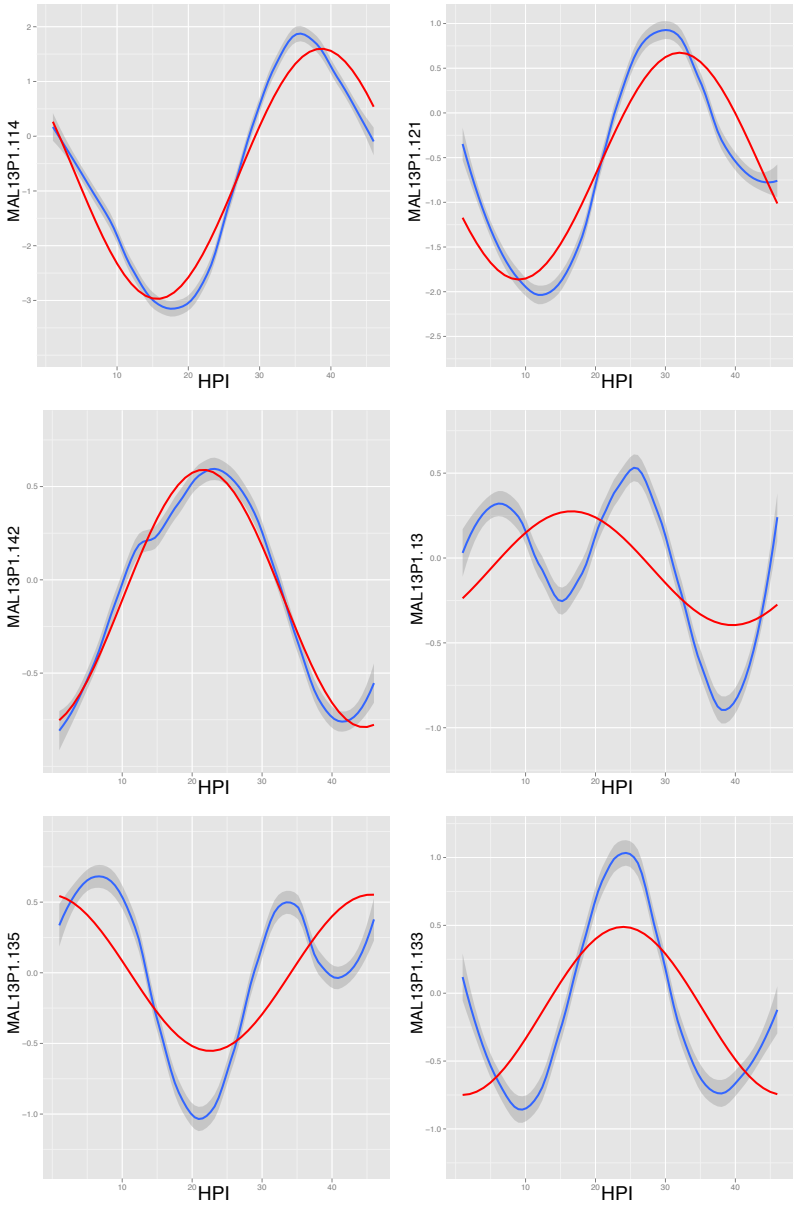

Figure 1.4: Periodic expression profiles of the majority of *P. falciparum* genes measured at hourly intervals by Bozdech et al. [9]. Raw measurement values are plotted as black points; a smoother (locally weighted regression line) is shown in blue, and a sine-fit is in red. The expression patterns of more than 70% genes oscillate during the course of the asexual development cycle. Six representative genes are shown. Most of the genes have a period of 48 hrs with a single maximum and single minimum throughout the cycle (*top row*). These genes are often well-described by a sinusoid with frequency $\omega = 1/48$ hrs (red curve). However, some of the genes oscillate with different frequencies, e.g. MAL13P.13 *(middle row, right panel)* and MAL13P.135 *(bottom left)* appear to have a 24 hour period. Other genes are described well by the sinusoid's frequency but not its amplitude *(bottom right)*.

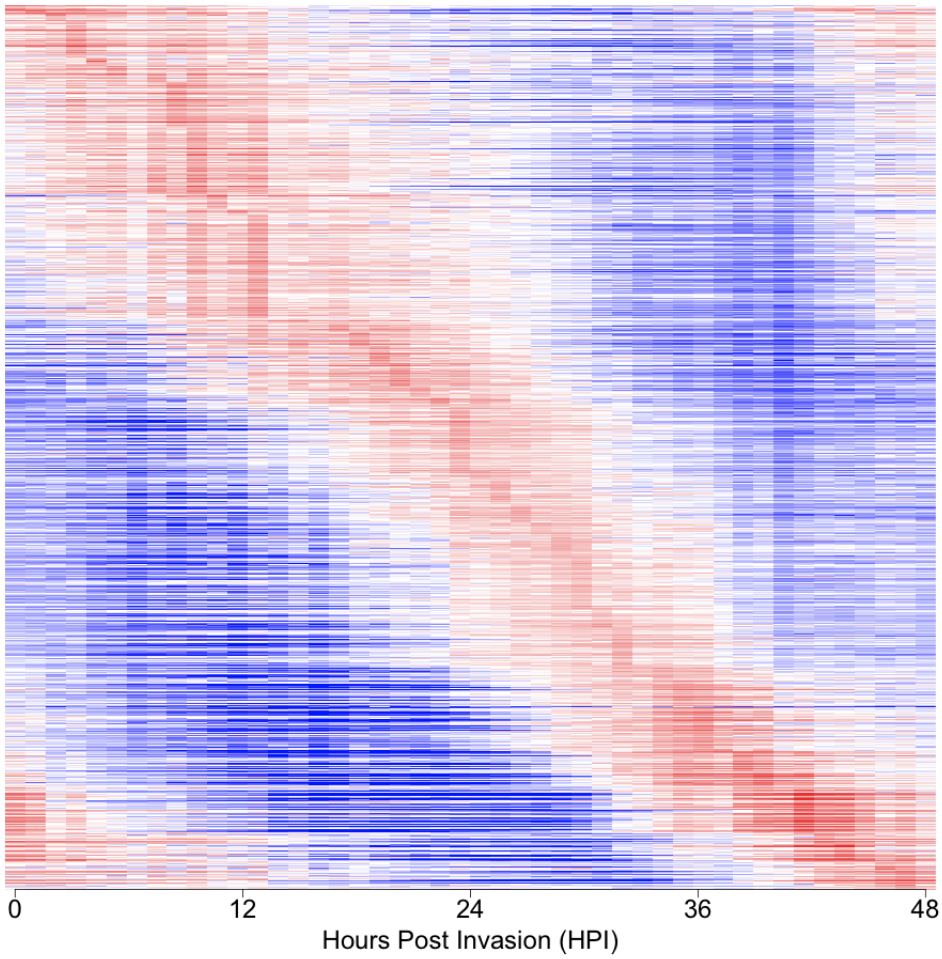

Figure 1.5: The intraerythrocytic development cycle shown as a "heatmap". A heatmap is a data display format in which each row is a gene, each column is a sample, and the expression level is represented by a color. In this case, red is high expression and blue is low expression. Data are from reference [9]. The rows are equivalent to the genes from Figure 1.4, but the heatmap has the advantage that it enables representation of a larger collection of genes than plotting them individually. The rows are ordered by maximal expression time. A set of 3532 genes that met quality control metrics as measured at 46 timepoints were used to create the figure. The heatmap is closely related to how the expression set is stored in the computer because each element of the expression matrix is assigned a color and then plotted. As a mathematical object one can consider the expression matrix, A to be an element in a vector space of 3532×46 matrices, written $A \in \mathbf{R}^{3532\times 46}$. Adopting that viewpoint leads to useful decompositions and manipulations as discussed later, e.g. in Section 2.5.

produce the underlying time series of gene expression values in a manner that can be evaluated using existing statistical methods [4,140]. The methods developed in reference [42,85] and Chapter 3 and Appendix C provide some alternatives which enable the computational synchronization of samples.

Overall, gene expression and its control in *P. falciparum* is a complex topic of which we have only skimmed the surface at this point. We give a short introduction to the open questions and major challenges in *var* gene regulation in the next section. A more substantial introduction to the regulation of gene expression, with a focus on *var* genes, in Chapter 4.

1.3 Antigenic Variation: Questions and Challenges

1.3.1 Molecular Mechanisms of Antigenic Variation

Malaria evades the host immune system by altering the surface proteins of the infected red cell over time, a process known as antigenic variation. This phenomenon was first described in malaria in the now-classic experiments of Brown and Brown [11], in which the authors demonstrated the recrudescence of antigenically distinct parasites in the same host after chloroquine treatment. While the absence of a cloning step prior to inoculation left open the possibility that a selection of a distinct clone present at low levels was occurring, similar results were soon documented for other *Plasmodium* species, including *P. falciparum*, *P. fragile*, and *P. chabaudi* [60, 62, 98], suggesting antigenic variation was a genuine feature of many types of malaria.

In the 1990's, the underlying molecular details were worked out by several groups. Rather than try to give a complete account (the reader is referred to the reviews [75, 127] for more details) we emphasize here the main points which are

relevant for gene regulation. Working with cloned parasite lines of *P. falciparum* in 1992, Roberts and colleagues demonstrated the disappearance and reemergence of an antigenic phenotype over time in culture, confirming that *P. falciparum* used antigenic variation. Three years later, Su and colleagues [141] showed that *P. falciparum* contained a large family of highly variable, two-exon genes that were spaced throughout the genome. Several other groups, collaborating on two additional reports which were published contemporaneously [6,137], demonstrated that the expression of these genes switched over time, and that these switches correlated with alterations in the surface molecular phenotype of parasites.

In following years, several groups established many of the regulatory mechanisms by which these genes are controlled. Scherf and colleagues [126] observed *in situ* switches among *var* gene transcripts and, along with careful studies by Kyes and colleagues [77], established that each cell encodes a single, full-length *var* transcript. Horrocks and colleagues [63] studied the switch rates of different *var* genes and concluded that the on- and off-rates of different *var* loci were different. Deitsch and Wellems [31] identified a role for the *var* intron in maintaining the silenced state of a *var* gene. These seminal studies outlined some of the most important molecular details of mutually exclusive expression and switching among *var* genes.

Nevertheless, *var* genes have multiple layers of regulation, because expression of surface antigens requires a delicate balancing act between variability, which promotes immune evasion, and the establishment of a chronic infection, which requires that new variants not be exposed to the immune system too quickly. These evolutionary constraints have built a system which is tightly regulated in order to behave optimally under these circumstances, which is a complex task [119].

The stable inheritance of *var* gene expression status without obvious sequence reconfiguration [64, 65] seems to implicate epigenetic memory in gene expression. A major breakthrough came in 2005 when two groups established that epigenetic

control played a role in establishing mutually exclusive expression of malaria *var* genes. Freitas-Junior et al. and Duraisingh et al. described mechanisms similar to the telomere position effect (TPE) in yeast were in effect in silencing *P. falciparum* *var* genes. Telomeric heterochromatin, residing at silenced clusters at the nuclear periphery, contained silenced *var* genes in the off state. These foci marked a distinct nuclear sub-compartment for genes in the 'off' state, and activation of *var* genes involved transition away from these sites toward permissive nuclear zones.

The two studies also established a role for the histone deacetylase PfSIR2 in maintaining these foci of heterochromatin; SIR2 is an enzyme, highly conserved from yeast to mammals, that removes the acetyl groups from histone octamers, contributing to the silencing of adjacent genes and the overall compaction of chromatin regionally. The active *var* gene was seen to move away from these silenced clusters by fluorescence *in situ* hybridization. Silenced clusters of heterochromatin at telomeres have been shown in other organisms to be marked by methylation (specially, three methyl groups or trimethylation) at lysine 9 of Histone H3 and held together by heterochromatin protein 1; the functional importance of these proteins for mediating the silenced state was later shown by Chookajorn and colleagues [17], Lopez-Rubio and colleagues [89, 90] (H3K9me3) and Perez-Toledo et al. [116] and Flueck et al. [46] (HP1).

Collectively, these discoveries have generated a model for *var* gene regulation which is fairly complete. Active *var* genes occupy a transcriptionally permissive zone in the nucleus, while other, silenced members of the gene family reside in a compacted state in heterochromatin clusters at the nuclear periphery. Their silencing is maintained by their sub-nuclear localization, chromatin condensation mediated by histone modifications such as H3K9me3 and histone modifying enzymes such as SIR2, and other molecular factors which bind to the intron and sense paired promoters. This model is remarkably satisfying but still leaves some

unanswered questions such as what other molecules are involved in removing or adding specific histone marks, and how these feedback loops propagate, there are larger questions unaddressed by this model. How do transitions occur between *var* genes? This model treats all *var* genes equally, but what factors are responsible for the major observed differences between switch rate and chromosomal position [47]? What brings heterochromatin clusters to the nuclear periphery, and what keeps them there?

The importance of locus repositioning upon activation of individual *var* genes is a part of a broader appreciation in recent years of the physical properties of the nucleus in gene expression. The archetype for this model of transcription is the nucleolus, in which rDNA genes cluster in a spatially restricted region containing high concentrations of RNA PolI to facilitate the transcription of rRNA. In mammals and in yeast, tRNA genes–the targets of polIII–are spatially clustered as well [59,112]. RNA PolII appears not to homogenously distributed throughout the nucleus; instead it is localized into discrete foci or transcription "factories", both in mammals [67], as well as in *P. falciparum* [96]. These factories consist of > 8 molecules of RNA PolII assembled into macromolecular complexes of 45 - 100 nm in size [142].

While these factories remain controversial, in part because they are usually observed with electron microscopy or fluorescence in situ hybridization (FISH) and are not typically seen in GFP-labelling experiments, other evidence also underscores the importance of three-dimensional genomic space in transcription. Numerous studies have reported that actively transcribed genes are more likely to colocalize in the nucleus [108, 129]. Studies using chromosome conformation capture have also revealed that the physical structure of globin loci changes between silenced and activated states [144], and that chromosomes appear to be divided into two compartments which favor inter-compartmental interactions [86].

In malaria, the main spatial features of the nucleus have not historically been a

focus of investigation except for particular classes of genes. Telomere ends cluster together at the edge of the *P. falciparum* nucleus [49], and spatial repositioning of the *var* genes is well documented [36, 39, 50, 117]. Mancio-Silva recently observed preferential association between chromosomes as well as the spatial clustering of rDNA genes in the nucleolus, but aside from the *var* genes and rDNA genes, the role of spatial positioning of different loci remains to be determined.

1.3.2 Experimental Challenges

Because of its relationship to disease pathogenesis, acquisition of immunity, and chronicity of infection, antigenic variation is at the heart of many of the most important aspects of malaria. Yet, as the previous section made clear, the mechanisms of *var* gene expression are incompletely understood. Indeed, instead of becoming simpler over time, the picture and molecular details seem to become more convoluted. The trees are becoming clearer, but perhaps the same is not true of the forest. Why?

Perhaps the most important factor that hinders progress is the difficulty of making genetic manipulations to the parasite. Technical factors which limit the efficiency of transfection, as well as the inability to rapidly grow up a population of cells from an individual cell, make modifications to *P. falciparum* genetic loci difficult and time-consuming [149, 150]. Much of what has been learned about *var* gene expression comes from studies using artificial constructs maintained as episomes. While such studies have offered unique insight into the cis-acting factors and activation sequences necessary for *var* gene activation, it is not clear whether the results found in these artificial constructs would hold true under natural activation conditions (i.e. without artificial selectable markers) and in the proper chromosomal context. At present, the poor genetic manipulability of the parasite and the limited toolkit of genetic techniques available in this non-model organisms

makes testing such hypotheses difficult.

A second major obstacle involved in studying *var* gene expression is a limitation in the available techniques to study multi-gene families. Quantitative real-time PCR and microarray technologies have ushered in an era in which it is possible to measure the expression of large numbers of genes cheaply and easily. Yet *var* gene expression is in large part spatially coordinated, through the nuclear clustering of chromosome ends and the generation of a *var*-specific sub-nuclear site. Techniques to study the spatial localization of gene families have not kept pace with techniques to study the transcript levels or the chromatin state. This is an issue that we directly address in this thesis, by creating methods that facilitate mapping of the spatial characteristics of the entire genome.

1.4 Structure of Thesis

We investigate gene regulation from two complementary perspectives in this thesis. The first part, which spans Chapters 2, 3, and Appendix C, addresses statistical issues encountered in analyzing observed gene expression patterns from malaria parasites. These studies were undertaken at the beginning of thesis research because of a question that recurred in interpreting malaria expression profiles: In a parasite whose gene expression varies dramatically throughout the course of a lifecycle, can the observed expression differences be due to subtle differences in lifecycle progression or age? Put differently, can the results of a single timepoint comparison even be taken seriously in an expression profile study? If no, then how many timepoints should researchers be performing before the results can be considered valid?

We address this question by first defining the problem by presenting an example from a study in our own lab, performed in collaboration with researchers in Gambia and at the Sanger Institute. The problem is fundamentally a statis-

tical one, and we begin to address it by exploring the mathematical structure of expression profiles in developing malaria parasites–over time and across the sexual/asexual lineage transition. That work is presented in Chapter 2. Building on those results, we then go on to present several ways to estimate stage and lineage commitment. Parameters describing temporal progression and lineage commitment can be learned from observed expression profiles, and doing so yields improvements in the analysis of expression data. Most of the relevant biological results are presented in Chapter 3, while Appendix C contains a more thorough description of the algorithms and comparison between methods.

The second part, which includes Chapters 4 and 5, takes up the underlying question of what determines the observed expression patterns. Inspired by the recent work of the Scherf, Deitsch, and Cowman labs in establishing a role for epigenetic regulation of gene expression [17, 36, 39, 50], we began to investigate the role of histone marks and the reversible silencing of *var* genes. Chapter 4 contains an introduction to the phenomenon of epigenetic regulation in malaria, results pertaining to the changes in mark distribution at individual loci in the genome, as well as the distribution of marks genome-wide in the context of *var* gene switching. These studies were conducted at a time of rapidly evolving new technologies including second-generation sequencing methods. A major focus of the chapter is on pushing the technical boundaries of epigenetic measurements that can be performed. In particular, we focus on epigenetic changes in the A4 parasite after *var* gene switching, and our findings are consistent with those of other studies such as [17, 89, 90]. Chapter 4 also serves to set the stage for the subsequent investigations of Chapter 5.

The final data chapter, chapter 5, investigates the spatial properties of the genome in the context of other epigenetic changes that occur as a part of *var* switching. In this chapter, we leverage next-generation sequencing to build spatial maps of the genome. We then use these maps to study the reconfiguration that

occurs in genomic folding when the active *var* gene switches. The final two data chapters remain as works in progress. Because the time between performing an experiment, obtaining second-generation sequencing, and analyzing the enormous quantity of data produced can sometimes be long as six months or more, it is perhaps worth noting the disclaimer that we are still in the process of obtaining the sequence and generating the final analysis for some of our datasets.

A theme throughout the thesis is the use of large, multivariate datasets to answer questions in malaria molecular and cellular biology. The power of these methods to build a complete picture of the cell, for example by studying the expression of all genes instead of just a handful, is truly awe-inspiring. Yet by nature this type of data entails a Faustian bargain in which we sacrifice ease-of-interpretability and introduce a susceptibility to subtle artifacts in exchange for an unprecedented quantity of information. In the succeeding chapters, we focus on extracting meaningful biological insights from such datasets while trying to develop methods that protect against some of their inherent pitfalls.

Chapter 2

Patterns of Gene Expression *Ex Vivo*

2.1 Introduction

Clinical malaria is a highly variable condition in which individuals can present with symptoms ranging from simple fever to severe disease involving coma and potentially progressing to death. The determinants of infection severity are largely unknown, and we hypothesized that expression differences in parasite genes influenced the disease phenotype. Here, we use microarrays to quantify the expression of all protein-coding mRNA transcripts in mature stages in malaria parasites from cases of severe and mild disease. We were able to identify some differences between groups, but none passed the threshold for significance after correcting for multiple tests. Given that this was a pilot study aimed primarily at establishing the feasibility of the approach and working through some of the statistical issues which arise in the analysis, the lack of strong associations is not surprising. While this may indicate a lack of genuine differences in parasite transcript expression between the phenotypes, it is more likely to be due to a lack of study power.

As we have discussed in the introduction, the analysis of gene expression from a single time point is potentially fraught with difficulties involving synchrony, the time-dependent nature of most genes, and the commitment of parasites to gametocytogenesis. A major goal of the analysis in this section is to address some of the issues that arise in the context of a genuine experimental situation. In this context, we discuss statistical challenges arising from the study of expression patterns, and we investigate molecular patterns that correspond to asexual development and commitment to the gametocyte lineage. We show how treating expression datasets as matrices facilitates the identification of these patterns, and sets the stage for a more precise treatment of the stage and lineage estimation problems of the next chapter.

2.2 Searching for Disease Severity Genes

In collaboration with Professor David Conway and Dr. Natalia Escobar-Gomez, we sought to identify parasite genes whose expression differed between malaria cases of varying severity. Identifying genes which modulate disease virulence is a major goal of basic research because of obvious therapeutic applications. Furthermore, the ability to separate cases of malaria into severe or mild from their expression would have prognostic value because clinicians could predict the course of disease based on an expression signature alone.

We enrolled patients in a cross-sectional study to investigate changes in parasite gene expression in relation to clinical disease phenotype. Standard clinical laboratory tests were measured on all patients. In an effort to identify virulence genes, which we hypothesized were likely to be expressed in the mature stages during cytoadherence, parasites were cultured *ex vivo* until schizogony.

2.2.1 Study Design

Patients were enrolled in the trial by consent upon admission to the MRC Gambia research station. A total of 21 patients were included in the study. Blood was drawn from these individuals and grown in *ex vivo* culture until parasites reached maturity. Once parasites reached the schizont stage as measured by Giemsa staining, infected red cells were placed in TriZol and RNA extracted by standard procedures. Later, labeled cDNA was transcribed and microarray hybridization performed on Affymetrix microarrays using the custom-designed chip known as the *pfsanger* array, a high-density tiling-like array. Clinical variables were recorded for all patients in the study, including age, parasitaemia, clonality of infection, lactate, hemoglobin, and Blantyre coma score. Clinical data for patients included in the study are available in Table 2.1.

ID	Culture	Severity	Age	Pt.	Hb	Lact.	Blantyre	Diagnosis
17	24	mild	1	6.4	9.6			mild
22	23	mild	4	3.3	11.1			mild
29	23	mild	5	8.1	11.0			mild
38	29	mild	2	2.1	6.4			mild
43	48	severe	4	12.3	12.3		4	prostration
53	35	severe	2	5.0	8.4	6.0	2	cerebral malaria
61	41	severe	3	10.0	7.4	6.0	2	cerebral malaria
62	24	severe	11	6.3	9.8	0.9	5	prostration
64	43	severe	3	19.0	6.8	6.6	2	cerebral malaria
65	23	severe	1	12.4	2.8	4.7	5	severe anemia
67	33	severe	2	9.4	10.2	2.0	5	prostration
73	36	severe	4	3.2	6.0	8.3	5	prostration
83	40	severe	1	13.0	7.7	7.7	2	cerebral malaria
802	44	mild	13	3.3	14.0			mild
804	24	mild	5	7.0	11.7			mild
807	23	mild	7	10.0	9.0			mild
808	42	mild	5	6.0	12.6			mild

Table 2.1: Clinical variables and patient characteristics from individuals included in the study. Culture gives number of hours the isolates were grown *ex vivo* until they reached maturity. Parasitaemia is abbreviated Pt. and Lactate is abbreviated as Lact. The Blantyre column gives the Blantyre Coma score (lower is more severe).

Data Analysis

After hybridization, raw fluorescence data were acquired and stored in files with the .cel suffix. Information about probe sets was available in a .cdf (chip definition) file available from the pathogen microarray group at the Sanger Institute. Using the statistical computing environment `R`, and the associated project `Bioconductor`, microarray data were normalized and expression levels were calculated for each gene using a suite of algorithms known as Robust Multiarray Analysis (RMA) [68]. Quality control metrics, including checking for degraded RNA by examining the expression levels along the length of the transcript, and assessing the overall intensity distribution of fluorescence values for outliers, were performed. Four of the samples included in the study were found to have abnormally low distributions of fluorescence intensity and were removed from the study. The estimated density of this distribution is shown in Figure 2.1; samples

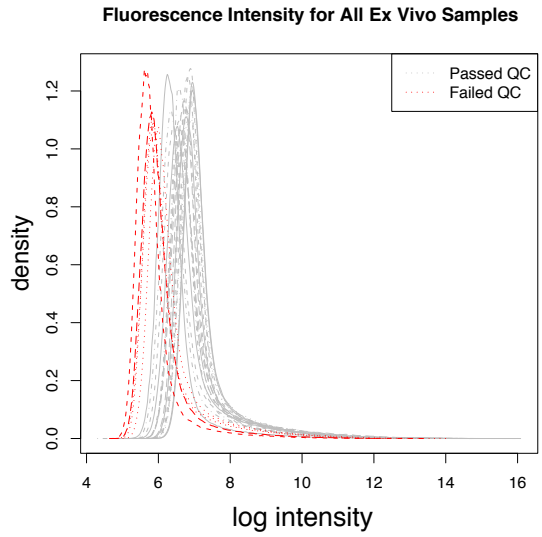

Figure 2.1: The distribution of fluorescence intensities for samples in the study. The curves in the plot are smooth kernel density estimates from the histograms of probe log intensities on the chip. Differences in fluorescence intensities can be reduced but not eliminated during microarray normalization, and for reliable comparison, these curves should look the same. Curves marked with a red line did not pass quality control for this study and were removed from downstream analysis, reducing the sample size from 21 to 17.

are marked with red denote those removed from downstream analysis.

The patients in the study were coded into two groups – those presenting with uncomplicated malaria were labeled "mild", while those presenting with coma, prostration, or severe anemia were coded as "severe". Genes were tested for differential expression by means of a two-sample t-test, which tests for null hypothesis that the means of the two groups are equal. A "moderated" version of the t-test, designed to increase the power of the test by putting a prior, learned from the array as a whole, on the individual test variances was also used along with correction of p-values for multiple testing.

2.2.2 Results

Initial comparison revealed no genes with significant changes between two groups after correction for multiple testing. We attributed this to the small sample size ($n = 17$) and the large number of comparisons (≈ 6000). We therefore used the uncorrected p-value as a measure of trend toward significance, but with the understanding that the genes highlighted cannot be considered *bona fide* associations.

A heatmap (first introduced in Figure 1.5) is a display of gene expression data, now relatively common in microarray analyses, which plots a number of genes side-by-side. Individual genes are plotted across a row, and each sample is represented by a column; intensity of expression of a particular gene in a given sample is given as a color. Two heatmaps are shown in Figure 2.2. Genes with $p < 0.01$ are shown in the left map while genes with $p < 0.1$ are shown in the right map. The genes have been colored by z-score, and the color scheme is given in the figure legend.

The results are striking for several reasons. First, as discussed above, there do not seem to be many genes that are substantially differentially expressed between the conditions. Second, most of the hits are not what one might have expected. For example, while the gene in the top of the list, PFA0015c, is a *var* gene, and another gene on the list, PF11_0520, is a *rif*, the other genes do not stand out based on their annotation. Finally, the expression profiles across the group is heterogenous, even for the most differentially expressed genes. A strong possibility is that there may be unexplained sources of variation in the dataset, and we investigated what these might be.

There are two likely sources of variation in the data. The first is that severe malaria is a heterogenous outcome with substantial substructure; the second is that there is background temporal heterogeneity in the developmental stage of the measured parasites. We can likely improve the quality of the analysis by

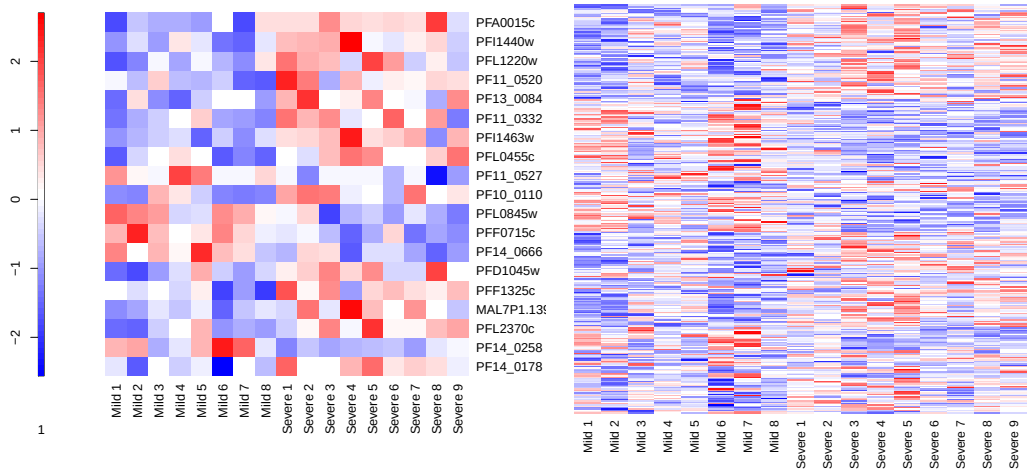

Figure 2.2: Heatmap of genes differentially expressed in patients with severe malaria (defined as the phenotype of cerebral malaria, severe anemia, or prostration from Table 2.1). Individual patients were coded as Mild or Severe and a t-test was applied to each gene to test for differential expression between the two groups. Genes with $p < 0.01$ are shown in the left panel. Genes with $p < 0.1$ are in the right panel.

Gene ID	**Product Description**
PFA0015c	*var*-like erythrocyte membrane protein 1
PFD1075w	serpentine receptor, putative
PFF0715c	endonuclease III homologue, putative
PFF1325c	c3h4-type ring finger protein, putative
MAL7P1.139	mago nashi protein homolog, putative
PFI1440w	conserved Plasmodium protein, unknown function
PFI1463w	conserved Plasmodium protein, unknown function
PF10_0110	conserved Plasmodium membrane protein, unknown function
PF11_0527	conserved Plasmodium protein, unknown function
PF11_0332	nucleic acid binding protein, putative
PF11_0520	rifin
PFL0455c	conserved Plasmodium protein, unknown function
PFL0845w	conserved Plasmodium protein, unknown function
PFL1220w	conserved Plasmodium protein, unknown function
PFL2370c	conserved Plasmodium protein, unknown function
PF13_0084	ubiquitin-like protein, putative
PF14_0178	ubiquitin fusion degradation protein UFD1, putative
PF14_0666	conserved Plasmodium protein, unknown function

Table 2.2: Annotated function of differentially expressed genes from left panel of Figure 2.2 ($p < 0.01$). One *var* gene and one *rif* appear in the list, along with two genes from the ubiquin-proteasome pathway.

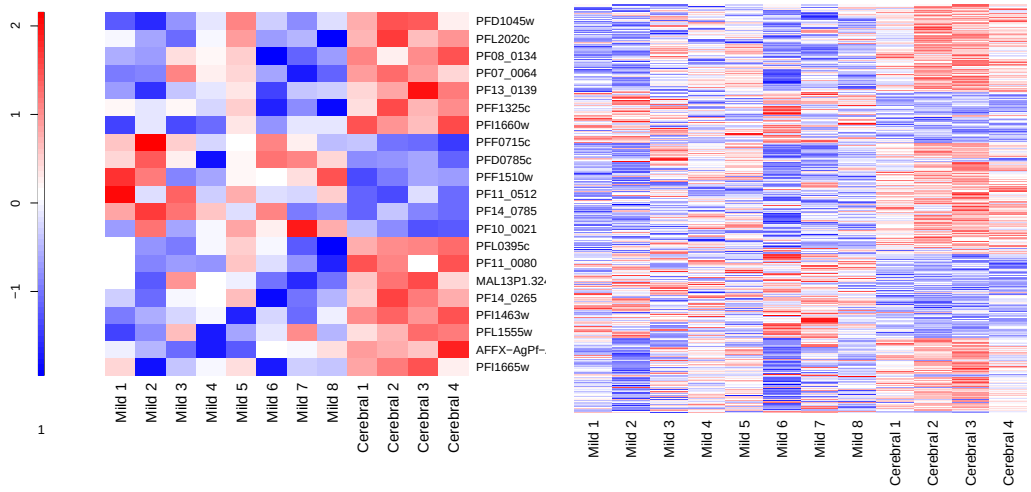

Figure 2.3: The severe disease phenotype is a heterogenous group of clinical presentations including cerebral malaria, severe anemia, and prostration. It is reasonable to assume that different genes predispose to cerebral malaria than to severe anemia. The heatmap above lists genes which are differentially expressed (significantly up- or down-regulated at a significance threshold of $p = 0.01$ in a t-test) for cerebral vs. mild malaria as a separate comparison.

accounting for these sources of variation. While this study is hampered by a small number of samples, we can disaggregate the data by the type of severe malaria and consider cerebral malaria as a unique outcome.

We applied the same methods (Student's t-test) using labeling of cerebral vs. mild malaria. The results are presented in the Figure 2.3; the figures there are analogous to those above but with cerebral vs. mild as opposed to severe vs. mild. Again, we were unable to detect any positive associations after correction for all the multiple tests performed; however, the uncorrected associations did seem to improve slightly, evident in the size of the p values as well as the increased contrast in the heatmaps. While the findings at the level of individual genes remain unremarkable, it is interesting to speculate that studies of this type may lose may lose power because of the phenotypic heterogeneity of severe malaria.

The second potential source of variation in the study is the variable temporal progression of individual samples, discussed in the next section.

2.3 Temporal progression

The striking visual pattern depicted in Figure 1.4 derives from the periodic expression of most parasite genes during the asexual lifecycle. Changes in mRNA concentration during the cell cycle evolve rapidly over time. While these changes presumably allow the parasite to accomplish the varied tasks of invasion, growth, and replication in a short amount of time, the oscillations in expression levels significantly complicate the interpretation of experiments meant to understand changes in gene expression. Time of development will always act as a confounding variable in the analysis.

An example of this is given in Figure 2.4. Consider that an experimenter has grown two populations of cells. Population A is treated with a drug and population B is left untreated as a control. The cultures are grown up, synchronized with sorbitol, and assayed for gene expression of gene X. Unbeknownst to the experimenter, the drug has a modest effect on growth rate of the culture, and therefore population A grows slightly slower than population B. Expression is measured for population A yielding a value represented by the red curve in figure 2.4. The expression of gene X in population B yields a value given by the green curve. The experimenter concludes that the drug acts specifically to repress the transcription of gene X.

The effect is not real but due to the confounding nature of temporal progression. Repeating the experiment will not help because the drug will reproducibly delay growth and yield the reduced expression phenotype in treated populations. The best way to address this issue would be to simultaneously measure temporal progression in the experiment, but since the effect is small, the conventional approach of using a Giemsa stain will not work here. Having an accurate and precise estimate of temporal progression would be very useful, which is the focus of Chapter 3 and Appendix C.

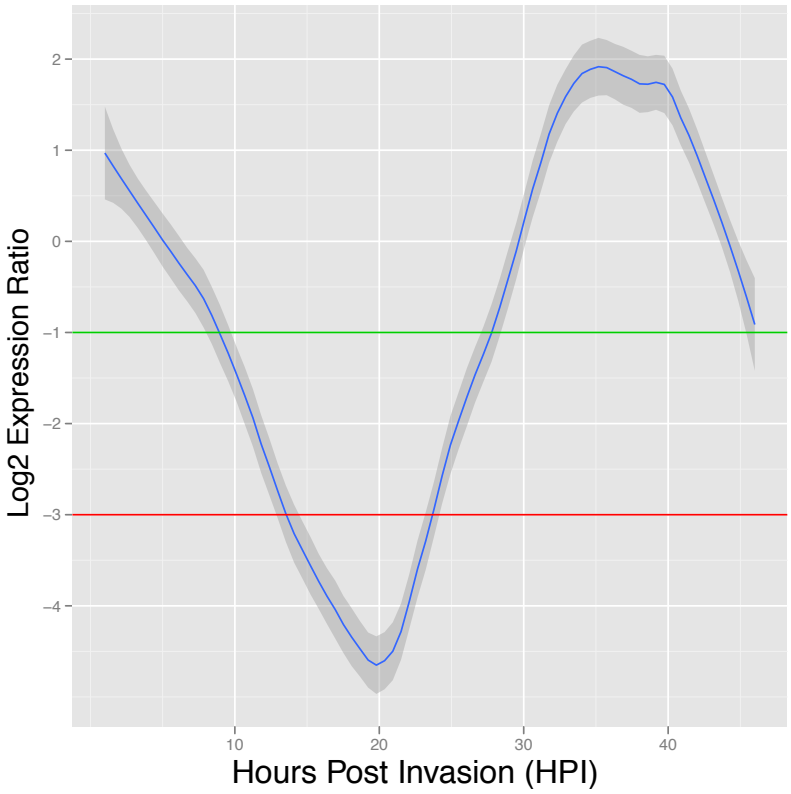

Figure 2.4: The confounding effect of time. The expression profile above is a representative expression pattern (from [9]) of a single gene during the 48-hr intraerythrocytic cycle. The gene reaches its minimum expression level at 20 hours post invasion, and rapidly rises to its maximum 14 hours later. The changes in expression level occur so swiftly that small differences in synchrony between cultures can easily be mistaken for differential expression. For example, consider an experimenter who wishes to measure two cultures at approximately 25 hours after invasion at the trophozoite stage. One culture is slightly faster than another and when measured yields a log expression ratio of -1 (this is represented by the green line). The second culture is 2 hours slower, and when measured yields a log ratio of -3, represented by the red line. Because dramatic changes in expression level occur around 25 hours post invasion, differences of ± 2 hours would yield apparent differences of 2 log2 ratios, or a four-fold difference.

Before progressing to that problem, there is one more phenomenon that we should point out. In the example above, we cautioned that the underlying nature of rapid changes in expression level could lead to spurious results about the effect of a drug or experimental condition on a gene, which we called gene X. In fact, the problem is far worse than that, because in a real malaria microarray experiment, we have approximately 6,000 genes, which we might call genes x_1 ... x_{6000}. The danger is that rather than arriving at faulty conclusions about a single gene, we would be led to faulty conclusions about all of the genes, or about the expression profiles as a whole. This is not purely an academic issue; we show later that this was part of the problem in the study by Daily and colleagues [27], though the issue there was a subtle transition to gametocyte-like expression patterns combined with batch effects as we discuss in Section 3.5.

More generally, there is something different about the structure of the experiments that we are performing. The patterns between genes are correlated. There is an "inverted pyramid" of dependencies such that we measure thousands of variables which depend on smaller numbers of underlying biological variables, and those in turn are influenced by technical and environmental conditions under which the experiment is run. In order to properly understand and analyze the results of microarray experiments, we need methods that can consider the dependencies between our measurements, which is the subject of the next section.

2.4 Microarrays, Sequencing, and High Dimensional Data

The quantity of biological data now publicly available and continually being generated is enormous. Microarrays exemplify this pattern, but it extends to a wide variety of data types. The malaria genome project, for example, begun in 1995

and completed in 2002, took 7 years and hundreds of personnel. At present that quantity of data is a fraction of what a small team of a few researchers can generate in a week. Along with microarrays, the use of liquid handling robots, biological mass spectrometry and massively-parallel sequencing have completely altered the landscape of biological investigation.

It is an exciting time for quantitative biological research, but there are also new challenges. The enormous quantity of data is impossible to digest in aggregate. Huge numbers of variables are measured simultaneously. The dependencies between these variables are not always understood. Many datasets are unwieldy or impossible to visualize and may take months or years to analyze even in part.

A theme in fields where large datasets are routinely used is the identification of measurements with points or vectors in an abstract mathematical space, typically a vector space. An experiment in which ten things are measured is said to be a point in a 10-dimensional vector space. In a commonly used terminology the set of 10 measurements would be a point, denoted x, and we might write $x \in \mathbf{R}^{10}$ (c.f. Section B.1) to say that it is a point in a real, 10-dimensional vector space. We use that construction and the associated terminology frequently, although mostly in the appendices. In making such an identification, it becomes possible to use mathematical techniques to understand patterns and relationships among biological variables such as genes.

This is often effective, for example in using the singular value decomposition (SVD – Section 2.5 and B.3) to discover meta-patterns or in quantifying the "distance" of a test sample from a reference sample (Appendix C). Treating the results of a data collection experiment as a collection of points in a vector space is a common theme in multivariate statistics. Interestingly, among the vector spaces usually encountered in data collection experiments, those of current biology are unusual in that they typically possess more measurements than samples. We discuss this circumstance, often called the $p \gg n$ case, in Section B.2 of Appendix B.

In the main text of this chapter, we focus on the relevant biological results that come from this line of thinking.

2.5 Matrix Decompositions

Breaking a matrix into simpler component matrices can be a useful procedure for a number of reasons. For computational and statistical purposes, it can reduce the dimensionality and complexity of the problem and reveal often simpler substructure. Such decompositions often reveal underlying dominant patterns that suggest expression modules or transcriptional programs.

2.5.1 Singular value decomposition

A major difficulty of microarray and other high-throughput studies is that the sheer volume of data generated is difficult to manage. When are two samples similar to each other and when are they different? The SVD can be used to plot samples and genes in lower dimensional subspaces that capture the variation in the data. The extensive dimensionality reduction that is possible results from the strongly correlated patterns of expression between genes.[1]

We factored the intraerythrocytic development cycle expression matrix using the SVD. A large portion of total variation (85% – see left-hand side of figure 2.5) in gene expression over the course of the cycle could be accounted for in linear combinations of two basis vectors, shown in center and on the right in figure 2.5. Another consequence is the ready comparison of samples in this lower dimensional space, including visual clustering and an intuitive understanding of the effect

[1]Note that singular value decomposition (SVD) is essentially identical to principal components analysis (PCA), except that standard PCA has pre-processing steps which set the mean of each row of the data matrix to zero before decomposition. Also, strictly speaking, PCA only applies to the case where there are more measurements than variables, which is not the case for expression datasets. We therefore tend to use SVD interchangeably for PCA because it is easier to speak of the left and right singular vectors. The exact relationship between the two techniques is mostly a matter of terminology and is explained in the Appendix, Section B.3

of experimental conditions such as time. For example, looking at Figure 2.6, we see that the samples move around on a circular, one-dimensional manifold in the principal components space, which in itself has many applications (c.f. Section C.3.5).

The highly correlated patterns of expression among genes is the reason that such substantial dimensionality reduction is possible. For example, Figure 2.7 gives the distribution of correlation coefficients for a randomly chosen gene. Several hundred genes are almost perfectly correlated with this gene, and several hundred more are almost perfectly anti-correlated with this gene. Such relationships arise from shared regulatory patterns; genes exist in pathways and share common regulatory factors which activate and inactive them as groups. It is these correlations that are effectively captured by principal component analysis, and the component vectors themselves, which in one sense are simply directions of maximum variance, are in another sense meta-patterns of shared biological meaning.

2.5.2 Non-Negative Matrix Factorization

The SVD is an attractive decomposition for several reasons, including its existence and uniqueness for almost all real matrices and its low-rank approximation properties. Nevertheless, interpretation of the SVD has the drawback that the factors can have negative elements. What does it mean, for example, for a gene to be composed of positive and negative combinations of an "eigengene"?

Non-negative matrix factorization (NMF) attempts to correct this problem by using positive factors. NMF is an approximate matrix factorization that attempts to address the shortcoming of negative combinations of eigenvectors produced by the SVD. First introduced by Lee and Seung in 1999 for face recognition [82], it has seen application to biological problems in recent years [12]. While NMF

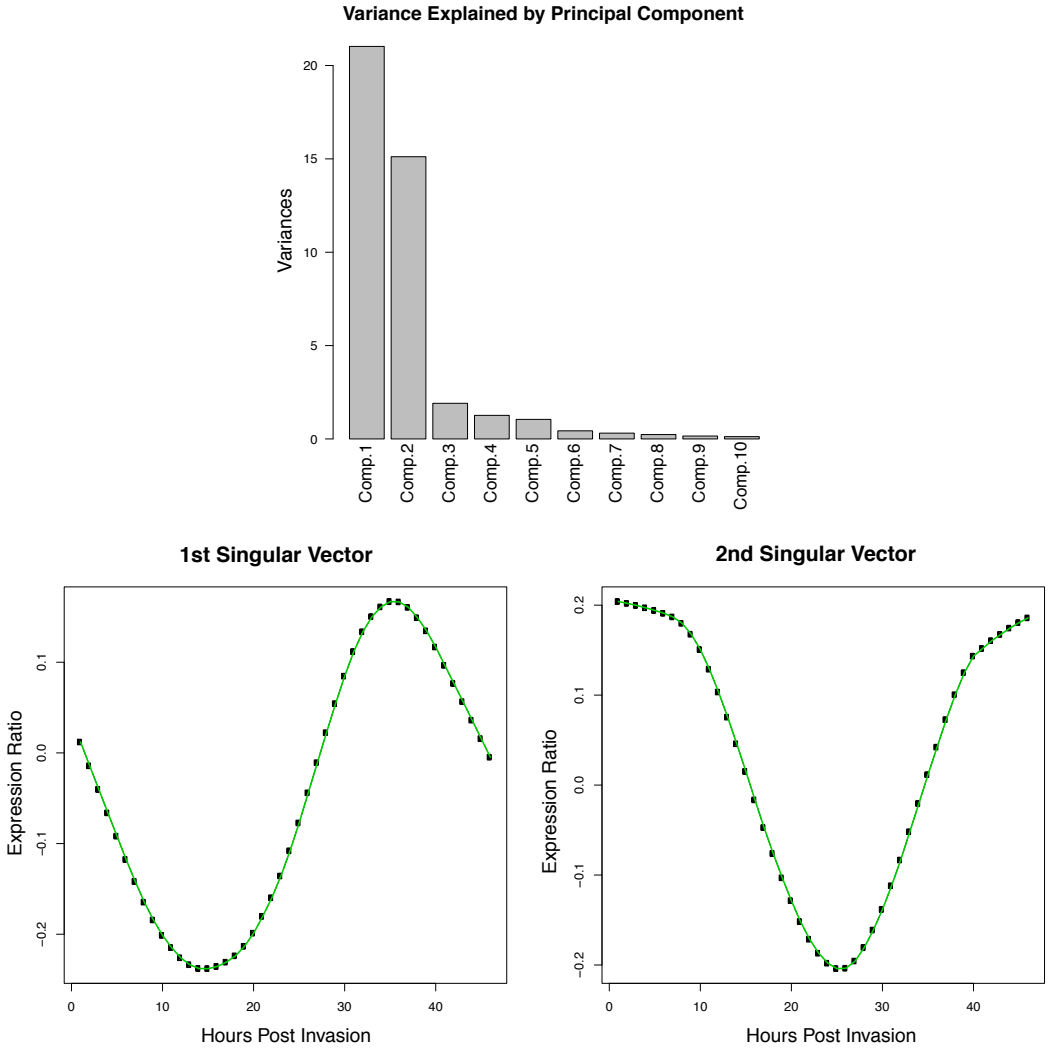

Figure 2.5: The top panel gives the variance in the dataset from [9] explained by each right singular vector, or principal component. The first two right singular vectors, which are shown in the lower panel (left and right, respectively), account for approximately 85% of the variation in the dataset. The individual principal components correspond to overarching patterns and are required to be uncorrelated. The curves in this figure suggest that there are basically two regulatory programs operating during asexual development – one which peaks at invasion and just before schizogeny, and another which peaks during ring stage and during schizogeny. Since the sign of these vectors is arbitrary, it might be more fair to suggest that there are four basic patterns of transcription which include the vectors shown with the sign changed. Overall, the simplicity of these patterns is in some ways contrary to popular belief that the periodic transcriptional program of the malaria parasite is complex and regulated by combinatorial deployment of a large family of transcription factors. The vast majority of total variation in gene expression can be accounted for by linear combinations of the above two expression patterns.

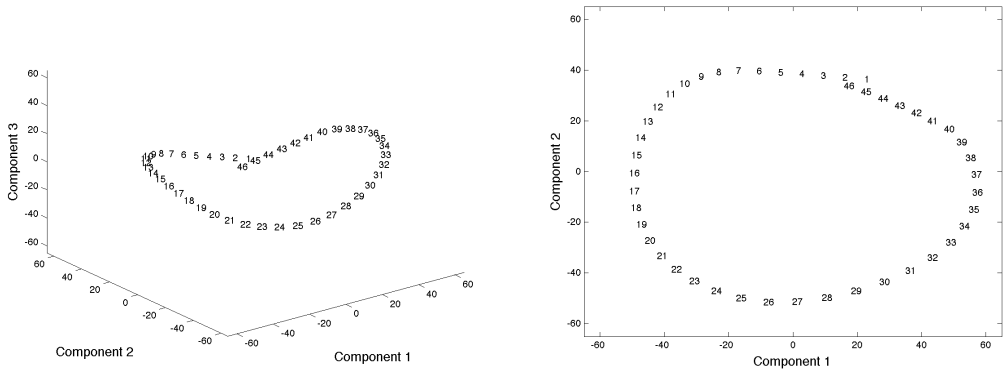

Figure 2.6: The dominant patterns of periodic expression in *P. falciparum* take place in a reduced dimensional space. The rows (i.e. the genes) of the data matrix (from [9] can reproduced by linear combinations of the left singular vectors, and are approximated by truncating the number of singular vectors used to the first k. Similarly, columns (i.e. samples) of the data matrix can be reproduced by linear combinations of the right singular vectors. Once again, taking the first k right singular gives a k-dimensional approximation of the dataset. In this case, we can examine the coordinates of the samples in the basis of the k-dimensional approximation. For $k = 2$ and $k = 3$ this is easily visualized, and is shown above. The periodic nature of the IDC is immediately apparent in both cases, and the temporal progression of samples traces out a nearly circular curve. This is elegant in its own right and also allows for a useful method of comparing samples by finding their coordinates in the space defined by the first k right singular vectors. We return to this as a method of detecting commitment to gametocytogenesis in Section 3.4 and in another method of stage estimation in the Appendix, Section C.3.5.

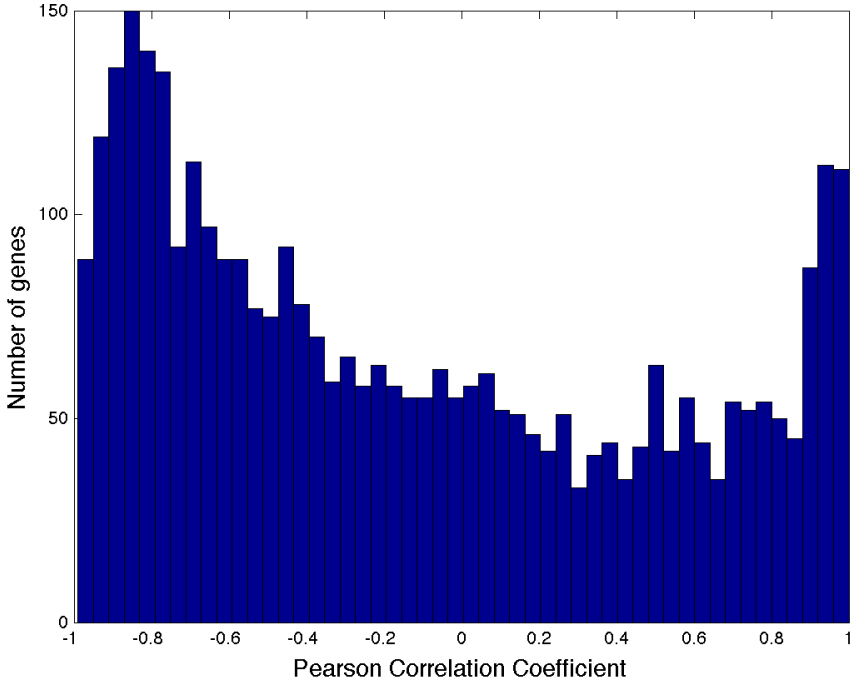

Figure 2.7: Histogram of Pearson correlation coefficients for a randomly selected gene from [9]. In this analysis, a gene from the smoothed intraerythrocytic development cycle was chosen and its Pearson correlation coefficient calculated with every other gene in the genome. The distribution is bimodal and has peaks around 0.9 and -0.9. Hundreds are genes are nearly perfectly correlated with the chosen gene, and hundreds of different genes are nearly perfectly anti-correlated. This analysis, which would be similar regardless of which gene is chosen, demonstrates that the patterns of expression are related to each other at a regulatory level. In many ways this is simply a different way of looking at the approximate low rank (c.f. Figure B.1) of the asexual transcriptome.

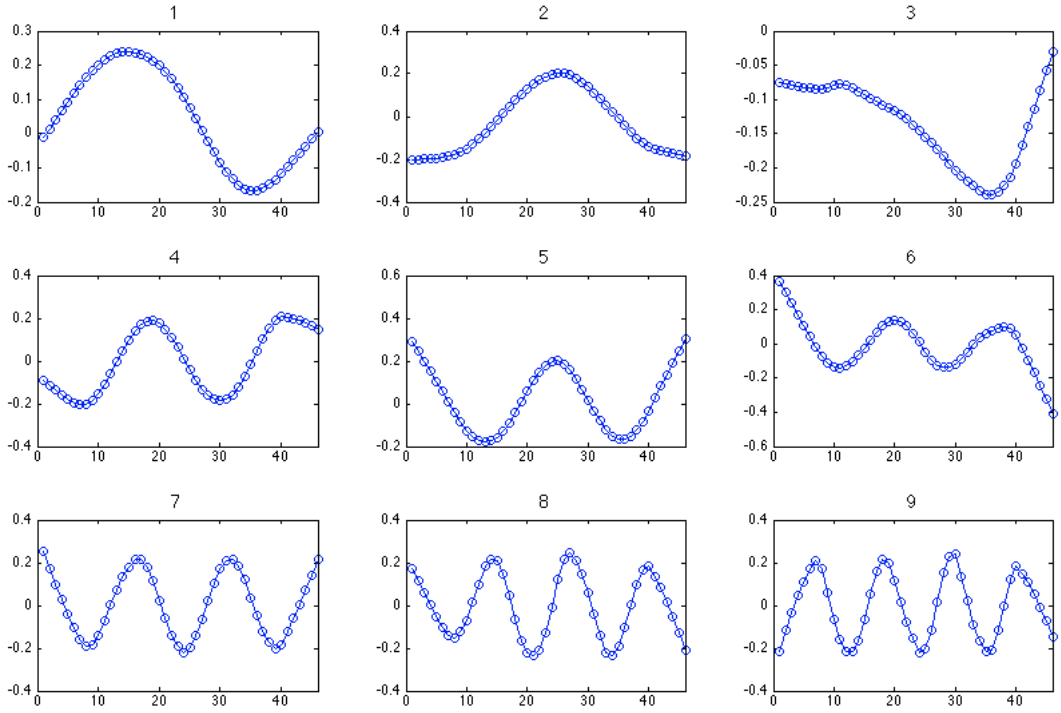

Figure 2.8: The left singular vectors yield insight into consensus regulatory patterns. The first 9 left singular vectors are shown here in order of importance (by the size of their corresponding singular values). Almost 100% of the variance in the data set (c.f. Figure B.1) can be accounted for by linear combinations of these genes. These do not necessarily correspond to transcription factors as the SVD by definition outputs orthogonal basis vectors, and transcription factors regulatory patterns are not required to be orthogonal. Nevertheless they reflect dominant patterns of gene expression shared by multiple genes and it seems likely that they have a close relationship to patterns formed by transcriptional determinants such as soluble factors.

lacks many of the elegant theoretical properties of the SVD, the non-negativity constraint on the factoring matrices aids in interpretability for a matrix such as gene expression values which may not have a clear meaning for negative values.

In NMF, an $N \times p$ data matrix, X, is factored approximately: $X \approx WH$, where W is an $N \times r$ matrix and H is $r \times p$. Typically $r < \text{rank}(X)$ and the factorization gives a compression. All entries of W and H are constrained to be non-negative. The columns of W form a basis for N and are sometimes, in the context of biological applications, referred to as 'metagenes'.

In order to better understand underlying patterns of gene expression in malaria, we factored the expression timecourse using NMF. Unlike the SVD, the basis elements depend on the rank of the factorization required, so a rank must be chosen before selecting the factorization. We present graphical summaries of the columns of W for the case where $k = 4$ in Figure 2.9. While the basis elements ('metagenes') were somewhat sensitive to the choice of k, the results in Figure 2.9 are representative and consistent with the results of SVD: The overall expression patterns can be well approximated by a small number of periodic basis genes that fall into two classes: genes that come on during ring and schizont stages, and genes than peak during trophozoite stages. An interesting result is that the metagenes look like positive and negative versions of the 'eigengenes' from SVD, suggesting that in the context of biological applications a reasonable interpretation is that a negative eigengene is equivalent to a positive contribution from a transcriptional module of the opposite pattern.

The agreement between the two methods gives support to the idea that malaria transcriptional patterns are constructed from a small number of constituent programs. It seems likely that these constituent programs reflect i) the actions of specific transcription factors binding to consensus sequence sites in promoter regions and ii) modifications in chromatin in the promoter region and along the length of the gene. The relative importance of these two factors is not known. Other

silencing mechanisms such as antisense transcription and post-transcriptional silencing may play a role in modulating the level of protein expression but would not likely be detectable at the level of transcription alone.

2.6 Gametocytes and the sexual development transcriptome

The malaria lifecycle is complex and possesses several distinct stages in the vertebrate host as well as the mosquito vector (c.f. Section 1.1.3). Unfortunately, not all stages are equally amenable to laboratory propagation, but the asexual and early sexual stages can be grown by most laboratories without the need for specialized equipment such as an insectary. For this reason, good datasets are available for asexual as well as sexual stages. Fortunately, these are some of the most relevant stages to understand from a research perspective, since the asexual stage causes the symptoms of malaria and the early sexual stages are responsible for mosquito transmission.

At this time, two datasets are available for the sexual stages of the malaria parasite. Silvestrini and colleagues [135] cultured lines that were genetically unable to make gametocytes and compared their expression to the NF54 isolate which readily produces gametocytes; however, the gametocytes only composed a fraction of the culture (approximately 10%) and therefore the power of the study to detect sexual transcripts was limited. Another study by Young and colleagues [152] purified gametocytes by magnetic fractionation after treatment of cultures with N-acetyl-glucosamine, a drug that suppresses asexual stages. Here, we focus on the data from Young and colleagues, because the later time points of their experiment consist of nearly entirely gametocyte stages and represent a relatively pure gametocyte transcriptome with limited asexual contamination. Nevertheless, no

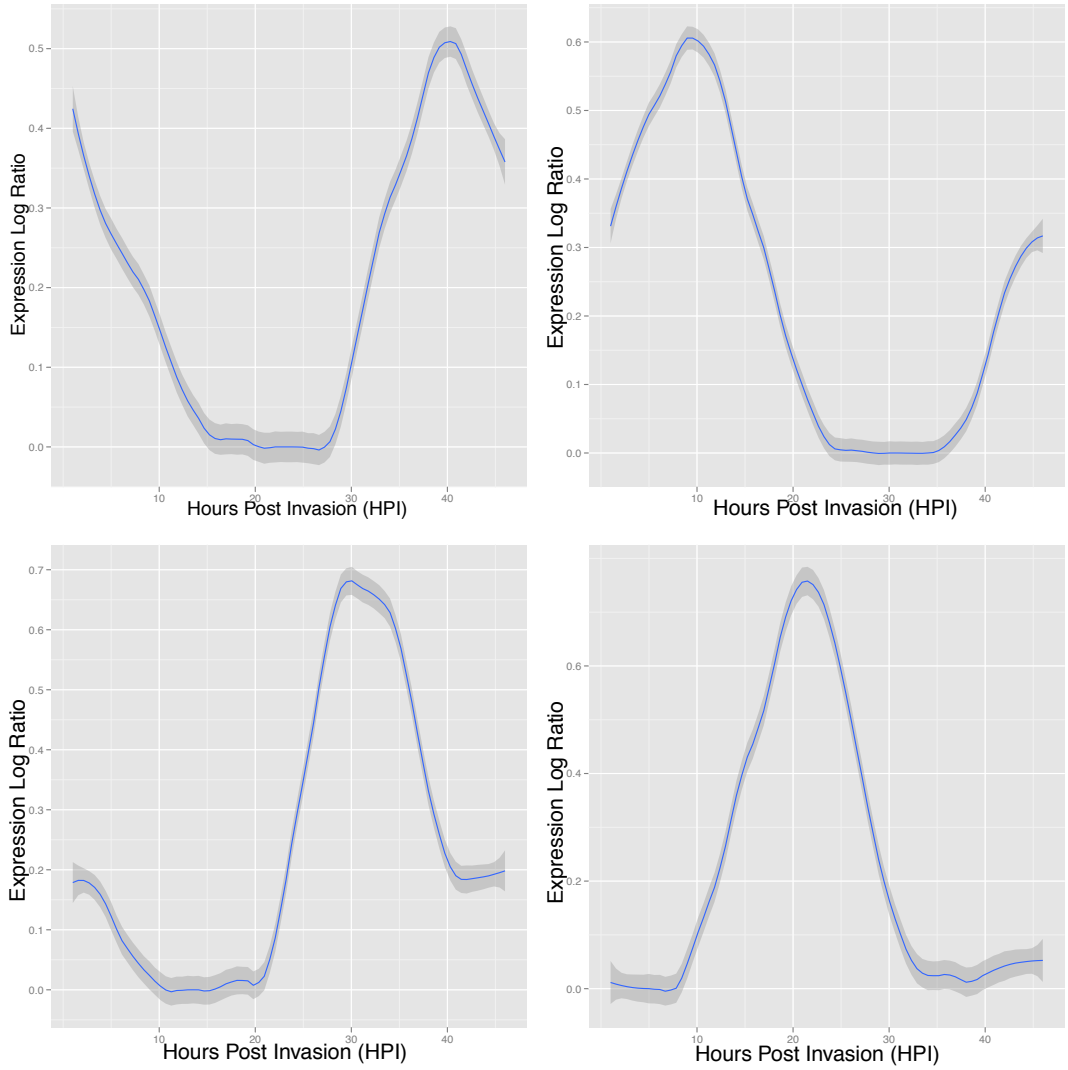

Figure 2.9: Non-negative matrix factorization is an approximate matrix decomposition that attempts to address the difficulty in interpreting negative values in the left and right singular vectors, which do not have an obvious biological interpretation. The four panels in this figure show the columns of the H matrix – the positive "metagenes". The components of each vector are shown as points. A loess smoother is added and the 1.96 standard error region is shaded. The similarity with the results from the singular value decomposition suggests that both approaches are discovering a common and genuine feature of the data, and that the apparently complex cascade of gene expression can be understood as a weighted combination of simpler patterns.

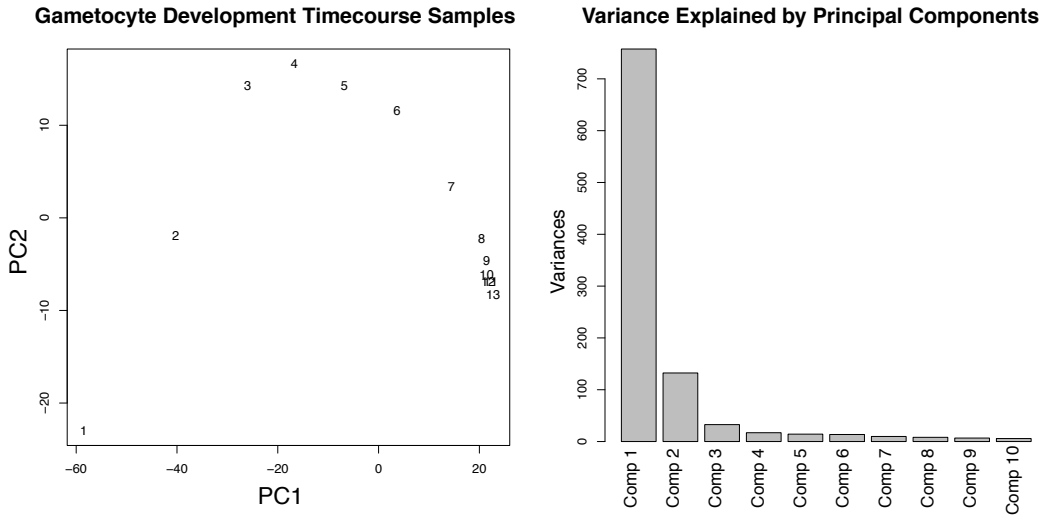

Figure 2.10: Principal component analysis of gametocyte development timecourse. The coordinates of each sample are plotted in the space spanned by the first two right singular vectors (also known as principal components). The variance of the dataset explained by each component is shown at right, demonstrating that nearly all the variation in the dataset can be captured by the first two principal components.

experiment has yet measured the transcription of early stage gametocytes. This is an unfortunate because the molecular features of these cells may distinguish them from asexual rings, something which cannot be done on the basis of morphology.

The singular value decomposition of the gametocyte development time course from Young and colleagues [152] is shown in Figure 2.10, and the singular vectors themselves are shown in Figure 2.11. Similar to the decomposition for asexual stages, most of the variation can be accounted for with the first two or three singular values. Interestingly, for sexual stages, the first singular value is far more important than the others; in other words, the dataset is approximately one-dimensional. In the experiment from Young and colleagues, the first 4 days have significant but decreasing contamination from asexual stages, which is important to keep in mind when analyzing the expression patterns. With that in mind, examination of the singular vectors themselves suggests how the dominant expression patterns might be interpreted: the first right singular vector corre-

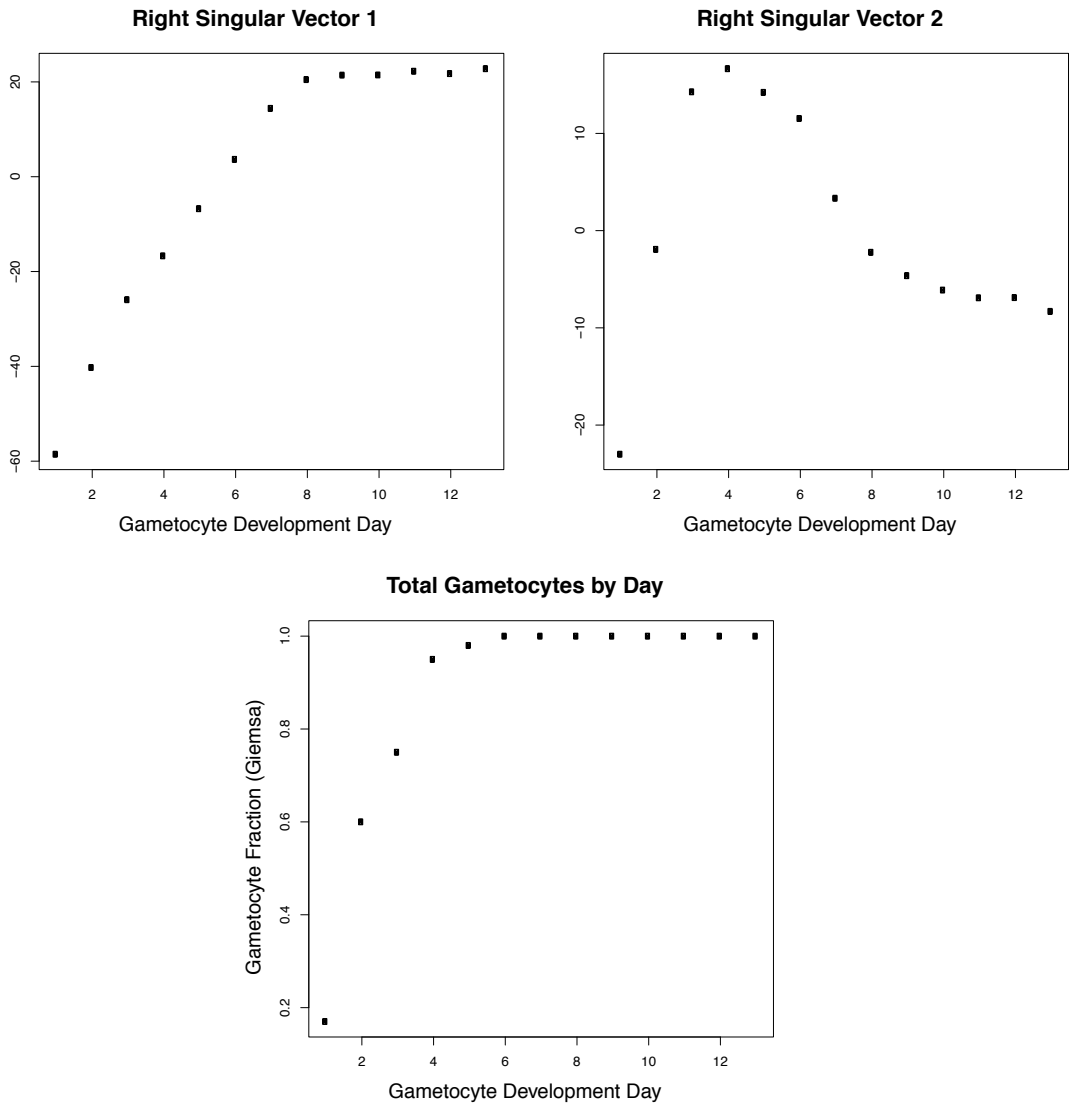

Figure 2.11: The first and second right singular vectors for the gametocyte development time course from [152] are shown above (top left and top right panels, respectively). These are sometimes called eigengenes, and show the patterns of gene expression, chosen to be uncorrelated, that are dominant in the dataset. That the singular value decomposition extracts biological patterns can be shown by plotting the total gametocyte fraction the cultures by day (right panel). The first left singular vector corresponds strongly to the fraction of gametocytes measured in the dataset. This is consistent with a relatively invariant gametocyte expression pattern; as gametocytes development, the most important changes in expression levels derive from changing cellular popluation and not differences between gametocyte stages.

sponds to genes which come on during gametocytogenesis and then stay on as the gametocytes mature from Stage I through Stage V. The second right singular vector corresponds to a set of genes that come on during the early phase of gametocytogenesis and then turn off during the later stages of gametocyte development. From the relative size of the first and second singular values, the first pattern predominates; however, it is interesting to speculate that there may be two transcription factors which control the early and late genes of gametocytogenesis. Furthermore, comparing the first singular vector (left panel of Figure 2.11) to the fraction of gametocytes present in the sample at each day (right panel of Figure 2.11), it becomes clear that the first singular vector is closely associated with gametocytemia in the sample. In other words, the strongest expression signature of the dataset is the loss of asexual signal and the gain of a predominantly invariant gametocyte signal.

A second feature of the gametocyte development time series is the similarity between samples. Gametocyte samples of different stages appear to resemble each other to a greater degree than asexual samples of different stages. Figure 2.12 shows the (symmetric) matrix of pairwise correlation coefficients between different samples. For all gametocyte samples after Day 5, the correlation between samples is typically ≥ 0.95, suggesting that there is little biological variability between samples of different days. On the other hand, asexual samples from the same arrays, processed in the same fashion, are shown on the right of Figure 2.12. Asexual samples of different stages have a Pearson correlation coefficient of ≈ 0.85.

Overall, the sexual development transcriptome appears surprisingly simple. The large quantity of variance that can be accounted for by the first principal component, as well as the strong correlation between time points, indicate that the gametocyte development transcriptome is characterized by stability and consistency over time relative to asexual stages. This approximate invariance of gametocyte transcription will prove useful in developing estimates of temporal

progression and lineage commitment for asexual samples (c.f. Section 3.4)—under the assumption that the approximate invariance extends to earlier time points of gametocytogensis, any of the gametocyte development days can be used to estimate the relative contribution of sexual-stage expression patterns.

Finally, we note that because of its "diagonalization" property, the SVD can be used to identify individual genes which correspond to any of the identified "eigengenes" or right singular vectors. By considering the projection of the expression patterns of individual genes onto the right singular vectors, which in this case is simply the coordinates in the principal components space, we can identify genes whose individual expression patterns are similar or dissimilar to the pattern of the individual principal components. This is useful for identifying genes important in early or late gametocytogenesis (Figure 2.13), and is also a good heuristic for reducing the number of genes used in stage estimation (Section C.4).

2.7 Conclusion

The ability to perform thousands of expression measurements in parallel, for example using a microarray or performing RNA sequencing, offers great power to understand the molecular patterns of cells, but it also brings unique analytical challenges. The rapidly varying, periodic expression signal of most genes in the malaria parasite presents specific difficulties which we refer to as the stage estimation problem. In this chapter, we have discussed the methods used to study gene expression in malaria parasites, and used a dataset from our lab as a case study of the confounding effect of temporal heterogeneity on the interpretation of expression studies. The complexity of malaria expression patterns, both cellular and developmental, necessitates methods which can assess substructure in expression datasets. We have shown that commonly-used matrix factorization methods provide a rapid, straightforward, and effective way to parse the structure of malaria

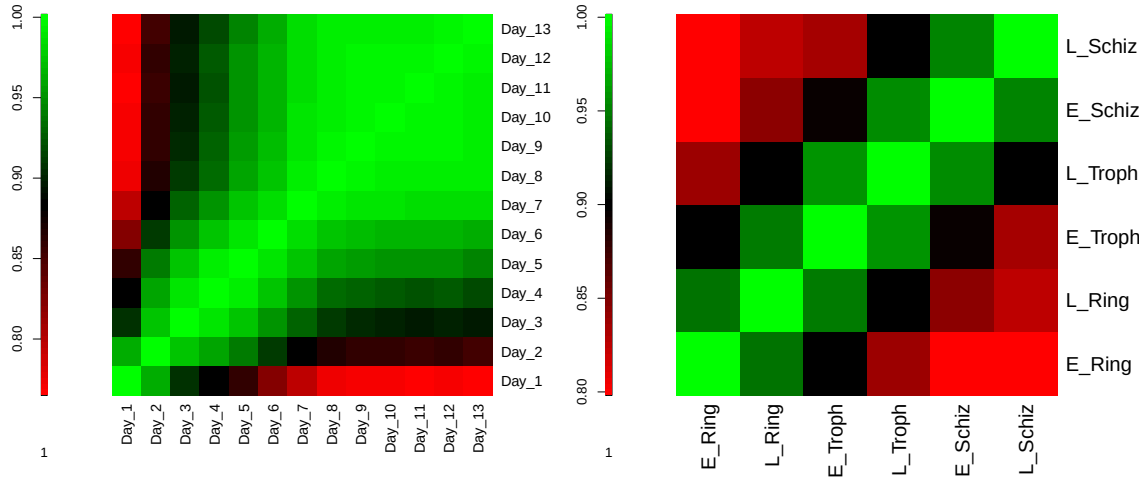

Figure 2.12: Heatmap of correlations between gametocyte development samples (left – from [152]) and asexual samples from array-matched samples (right – from [122]). Gametocytes of various stages correlate more strongly with one another than asexual parasites of different stages, suggesting that the basal transcriptional program changes only modestly during gametocyte development. This phenomenon can also be seen from the presence of a single, dominant principal component from the PCA data in Figure 2.11.

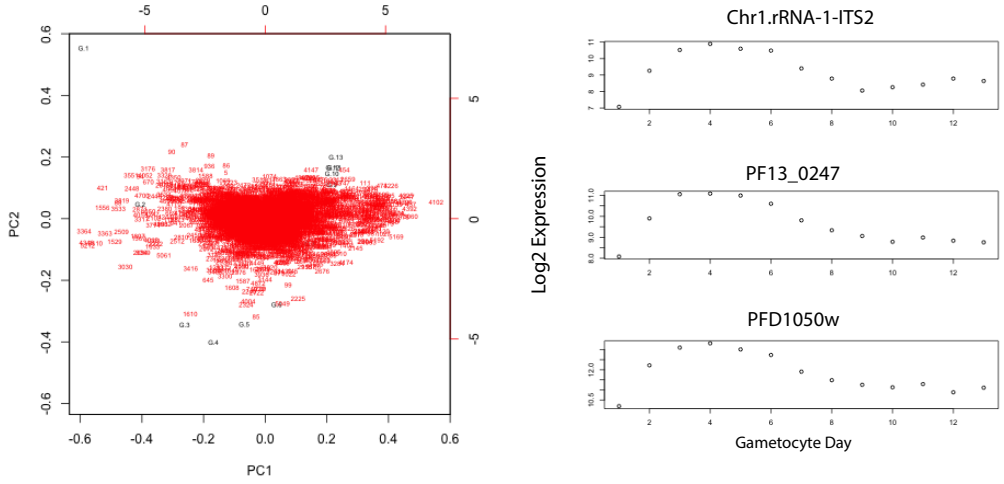

Figure 2.13: Identification of genes important in early gametocytogenesis via the SVD. Biplots of gametocyte samples and genes in the same two-dimensional space. In the singular value decomposition $A = U\Sigma V^T$, the columns of A (i.e. samples) can be expressed as linear combinations of the columns of U whose weights are given by ΣV^T. Similarly, the rows of A can be expressed as linear combinations of the rows of V^T (columns of V) whose weights are given by $U\Sigma$. Approximating A by the closest rank 2 matrix gives each of the samples and genes both as points in 2-dimensional space whose coordinates are scaled by the same values (Σ). A biplot is a method which takes advantage of this property to plot both samples and genes in the same space. This plot shows, for example, individual genes which contribute heavily to the patterns described by the principal components (right singular vectors). The second principal component corresponds to a pattern of induction during early gametocytogenesis. To find individual genes with this pattern, we can consider their coordinates in this space. This corresponds to points which have large or small values on the y-axis. In the right half of the figure, the expression pattern is shown for three genes with a strong early gametocyte expression pattern.

gene expression profiles. Understanding the patterns present in malaria expression data prepares us to tackle the stage and lineage estimation problems in the next section.

Chapter 3

Stage Estimation

3.1 Introduction

The periodic expression pattern of most malaria genes makes the results of single time point experiments difficult to interpret. Here we develop methods to overcome this problem by statistically learning the temporal development of samples. Using these methods, we can address questions in transcriptional biology that have been hindered by our inability to control for stage. For example, when substantial expression differences are observed, it is possible to evaluate the results for the confounding effect of temporally mismatched samples.

The first part of the chapter introduces our approach and tests the performance of stage estimation methods on several datasets of known temporal development. Alternative methods are developed in Appendix C. After validation, we use the information gleaned from stage estimation to improve the analysis of the severe vs. mild disease samples discussed in Chapter 2. To account for the presence of gametocytes, the stage estimation approach is extended to include decomposing malaria expression profiles as additive mixtures of asexual and sexual lineages. Using these methods, the samples from a large study of *in vivo* gene expression profiles [27] are reanalyzed and shown to contain continuous mixtures of sexual and asexual expression profiles. We compare our model of continuous variation to the discrete cluster model of Daily et al. [27], and demonstrate that statistical artifacts resulting from processing of arrays in batches can introduce artificial, discrete variation into expression datasets.

We describe three main algorithms, and several variations, for estimating the temporal progression of a malaria sample in Appendix C. The Appendix is meant as a stand-alone chapter and gives a detailed treatment of the methods and rationale for stage estimation. The method we typically use for stage estimation, and the one used to derive the results in this chapter, is a variant of the nearest centroids of Tibshirani et al. [143] and was suggested by Chris Holmes and

implemented jointly with Avi Feller. As is discussed in detail in Appendix C, the algorithm maximizes the multivariate likelihood with independent Gaussian probability distributions for each gene. Maximizing the likelihood is equivalent to minimizing the distance induced by a certain norm; it is therefore closely related to the method of PlasmoDB which maximizes correlation between test and reference samples. We discuss this further in the Appendix. The problem of reducing the number of genes, and of choosing the most informative subset, is also addressed there. Gene subset selection is a hard combinatorial; therefore, a heuristic is used to give an approximate solution (c.f. C.4). Performance of the methods for stage estimation, and biological results derived from applying it, are treated in this chapter.

3.2 Algorithm Performance

3.2.1 Accuracy

We evaluated stage estimation algorithms first by comparing our estimates to samples of known age, and second by testing whether measurements of parasite area, a proxy for temporal progression, correlate with increasing estimates of parasite age.

We obtained maximum likelihood estimates for the age of several *in vitro* samples and compared them to their known age. Figure 3.1 shows the relationship between estimated and true age for the Dd2 timecourse performed in reference [87]. As the regression line shows, the estimated values are close to the true ages for all time points; the estimates increase at the same rate as the known progression of time. This suggests that stage estimation is an accurate and reliable method to establish the temporal progression of a *P. falciparum* culture on glass-slide arrays under the experimental conditions used.

We sought an independent method to confirm the validity of stage estimation since for some samples, the true age of the culture is not known. Parasites are known to increase in size in the red cell as time passes in the asexual development cycle, and we used parasite area as a predictor of hours post invasion (HPI). We first determined the relationship between parasite size and hours post invasion. This is shown in Figure 3.2. A time course was performed in the 3D7 parasite with a sample taken every 3 hours and stained with Giemsa stain. The area of greater than 500 individual parasites was then measured by microscropy. We estimated the distribution by kernel density methods and plotted the relationship between the mode of the histogram of areas and the age of the culture. The relationship between parasite area and HPI was well fitted by an exponential relationship beginning after $\approx$ 20 HPI. Somewhat to our surprise, we were unable to detect changes in parasite size during the ring stage; however, a strong relationship was seen for mature stages.

We were particularly interested in obtaining estimates of temporal progression for the samples described in Section 2.2. Since we did not know the true age of these samples, we applied both the maximum likelihood approach and the microscopic measurement approach. The left side of Figure 3.4 shows the relationship between the logarithm of parasite size and estimated age of the culture. As parasite size increases in the dataset, so too does the estimated temporal age.

The measurements of parasite area and estimation of stage by maximum likelihood were in good agreement (Figure 3.3). Parasites with larger modal area of the size histogram had increased values of maximum likelihood. The relationship was well modeled by a linear regression, where the variance accounted for by the model (73% of the total) was high and the slope coefficient significantly different from zero.

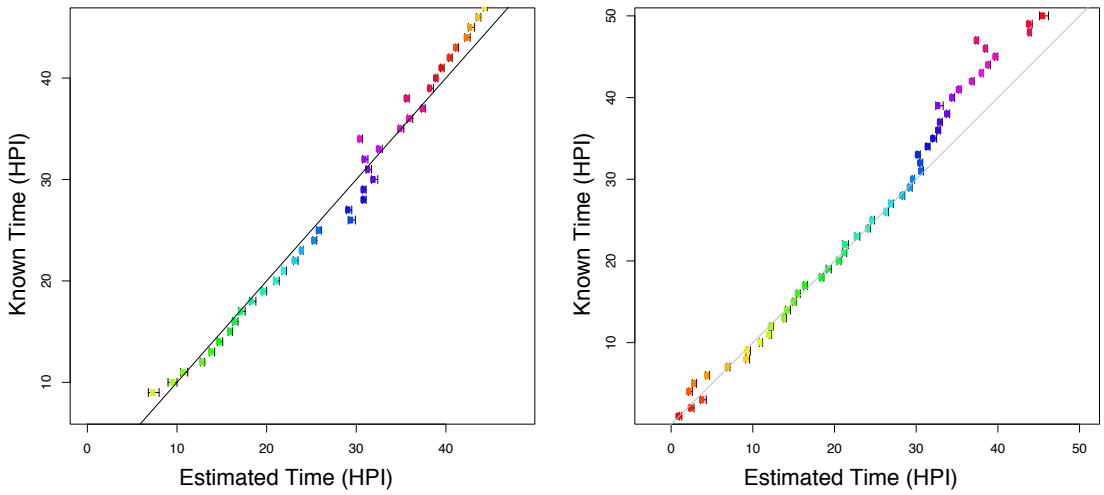

Figure 3.1: The accuracy of stage estimation procedure can be assessed by comparing our estimates to expression profiles from cultures of known age. Fortunately, several of these are available from the high-resolution studies of Bozdech and Llinas et al. [9, 87]. Estimates of Dd2 (left panel) and 3D7 (right panel) estimated HPI vs. known HPI. The known time for hourly samples of the Dd2 and 3D7 isolates from [87] are plotted on the x-axis and the estimated sample age using maximum likelihood estimation is plotted on the y-axis. The observed and estimated samples are in close agreement for nearly all timepoints studied. Deviations are observed > 30 HPI for the 3D7 isolate, possibly reflecting a commitment of a sub-population to gametocytes (c.f. Section 3.4.

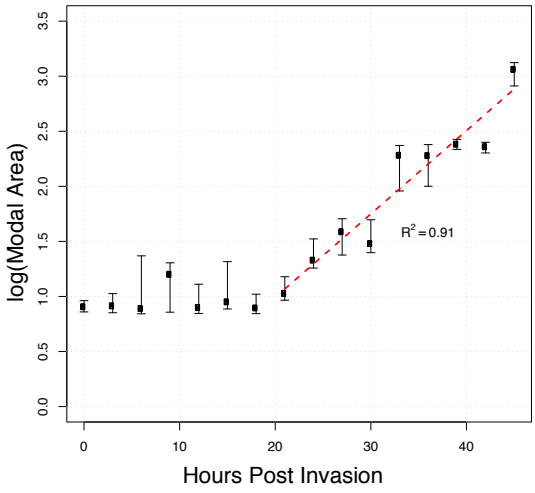

Figure 3.2: Parasite size as a function of HPI. In order to use parasite size as a proxy for temporal progression, we first established the relationship between parasite age and time. Giemsa-stained blood films were obtained from samples every three hours, and the size of > 500 individual parasites was measured. The mean area $\pm$ 95% confidence intervals (obtained via the bootstrap) are plotted above.

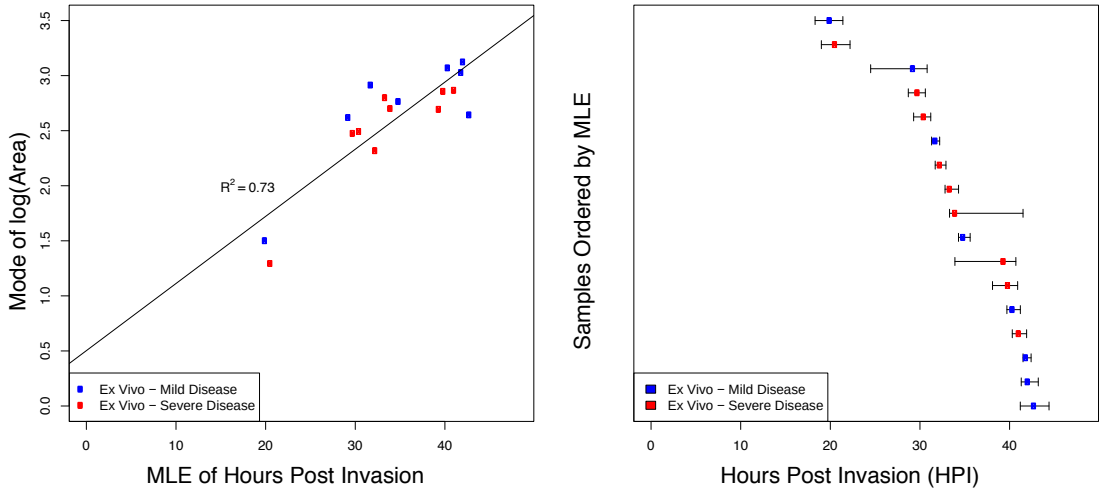

Figure 3.3: Parasite size (left panel) correlates with the maximum likelihood estimate of parasite age (left panel). Maximum likelihood estimates in the *ex vivo* Gambia samples show a range of temporal progression (right panel).

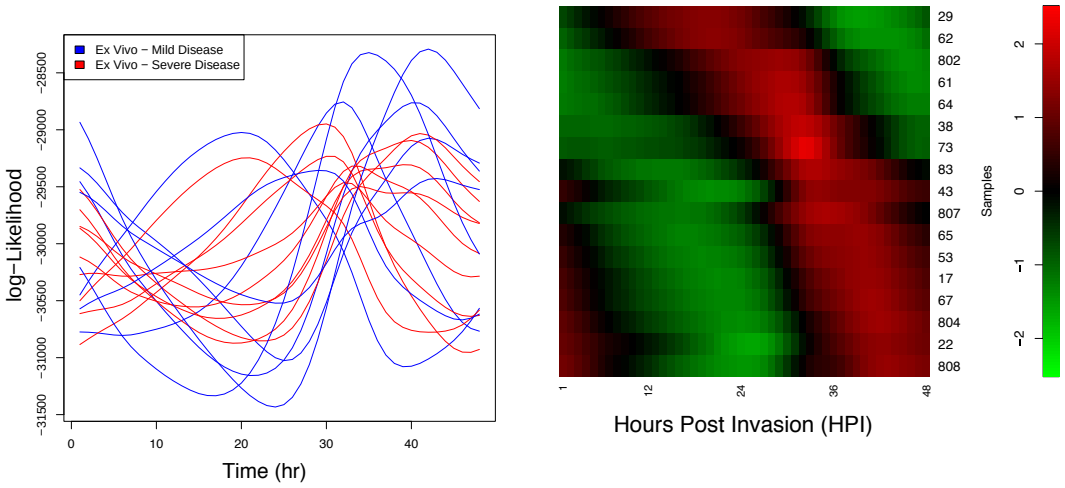

Figure 3.4: The log-likelihood curves for the *ex vivo* samples (left panel) and represented as a heatmap ordered by maximum (right panel). This shows once again the diversity in temporal progression present in the sample and underscores the difficulty of obtaining individual cultures of identical stage and synchrony.

3.2.2 Robustness

The results in Figure 3.1 indicate that the MLEs are accurate, at least in the case of measurements obtained from the same lab on the same array. We did not know if this is a general feature of the algorithm and set out to test the robustness of predictions by maximum likelihood when other types of arrays are used, or when a subset of genes is differentially regulated. Until now, we been using as a reference set the expression profiles obtained by Bozdech and Llinas [9], without offering much justification as to why or considering other reference sets.

The datasets from references [9, 87] were generated from the same lab and correspond to expression timecourses with hourly resolution in the parasite strains HB3, 3D7, and Dd2. Of these, the HB3 dataset in Bozdech [9], is by far the best; it contains the fewest missing values and the expression profiles are nearly exactly periodic, suggesting that the culture was tightly synchronized. For those reasons, we use it as the default reference set where possible. However, not all genes are

represented on the Derisi lab glass slide arrays, both because of high levels of noise inherent to the glass slide preparation and also because each gene is typically represented by a single oligo, and this is a major disadvantage. Furthermore, we used Affymetrix arrays in our experiments, which have many more probes per gene, cover more of the genome, and furthermore provide an estimate of absolute expression rather than the relative expression ratios of glass slide arrays. We therefore wanted to know whether stage estimation was successful for arrays across different platforms. Moreover, with the advent of direct sequencing of cDNA, "RNAseq", we want to be able to use expression profiles generated by that technique as well.

Figure 3.5 assesses the effect of changing array type or altering the expression of a subset of genes on the accuracy of stage estimates. Using Affymetrix arrays of the *scrMalaria* type, we are able to assign accurate ages to parasites sampled from 7 different time points, suggesting that performance of the algorithm is not adversely affected by the array type. Similar results were obtained using RNAseq data from reference [109]. Consistent with theory, the method is general to expression values independent of how they are produced.

The estimates of stage combine the information in individual genes across the sample, with genes weighted according to their biological variability (c.f. Section C.3.1). The method makes a technical assumption that the expression measurements are conditionally independent given time, an assumption that is unlikely to be strictly true in practice. We therefore studied the robustness of the estimation method to changes in a subset of genes by using the expression measurements collected in the presence and absence of doxycycline at multiple time points [26]. Doxycycline specifically represses the genes in the apicoplast. The maximum likelihood estimates for these samples are shown in the right panel of Figure 3.5. The estimates are essentially identical for both treated and untreated samples, suggesting that dysregulation in a subset of genes does not affect the

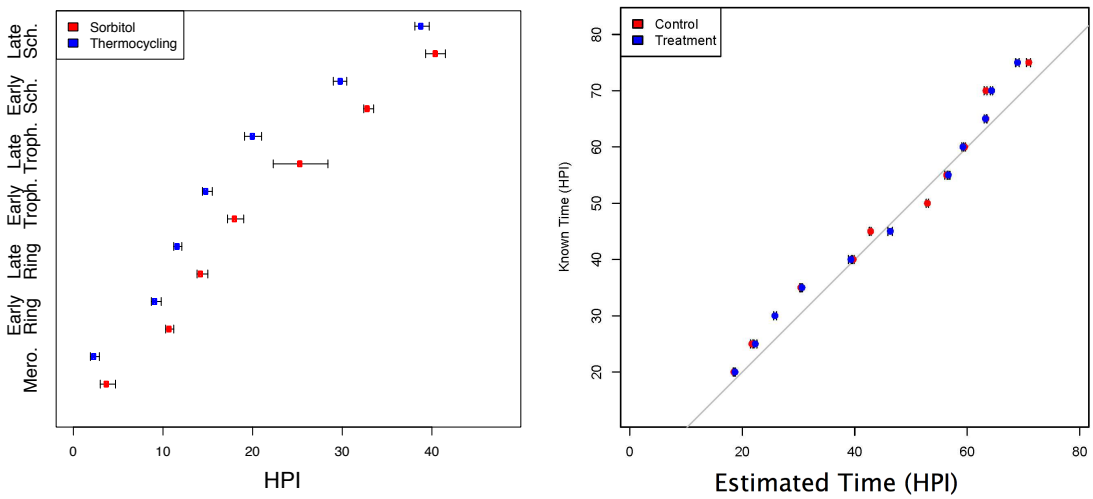

Figure 3.5: Testing the robustness of the stage estimation algorithm to changes in array type and down-regulation of a group of genes. The left panel gives maximum likelihood estimates for samples hybridized using Affymetrix arrays which yield an absolute measurement using a single fluorescence channel. The estimates assign approximately correct ages for samples from [122] for parasites synchronized using Sorbitol and Thermocycling. The right panel shows the maximum likelihood estimates for doxycycline treated and untreated samples from reference [26]. Doxycycline specifically downregulates apicoplast genes but leaves the remainder of the transcriptional profile in tact. The similarity in estimates between the treated and untreated samples indicates that the estimation procedure is robust to changes in a small subset of genes.

overall estimate for the sample.

When we applied stage estimation to gametocyte profiles, we noticed that the estimates of mature gametocytes clustered around ≈ 30 HPI. This is a result of the correlation of gametocytes to trophozoites, and as a result the presence of gametocytemia could act as a confounding variable in the stage estimation. Nevertheless, the gametocyte transcriptome is not exactly the same as the trophozoite transcriptome, which suggests that we should be able to differentiate between the presence of trophozoites and the presence of gametocytes. We develop that in Section 3.4.

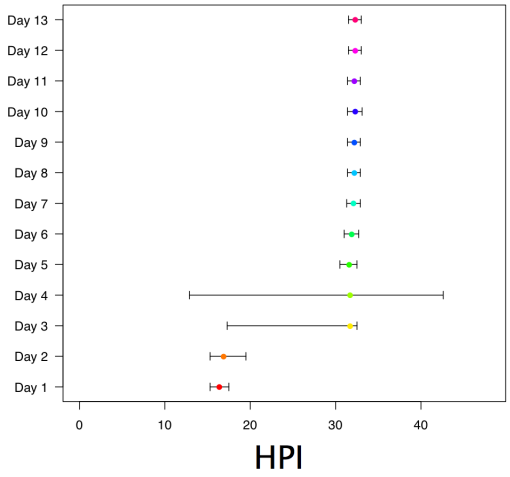

Figure 3.6: Maximum likelihood estimates of the sexual development transcriptome. In attempting to establish an apparent asexual age for the sexual development samples, we noted that mature gametocytes have an increasing peak in their likelihood curves at approximately 30 HPI. The estimated asexual age for mature gametocytes is $\approx$ 30 HPI, suggesting a similarity between the expression program of late trophozoite stage parasites and gametocytes, and potentially pointing to shared regulatory mechanisms between these stages.

3.3 Analysis of Expression Profiles Using Temporal Data

We have shown that stage estimation is possible, as well as that it is accurate and robust. We now use it to solve problems that arise from sampling microarrays from single time points. We would like to understand the observed variation in the expression samples collected from severe and mild malaria cases. Ideally we would do this at a whole-profile level as well as on an individual gene level. Are there expression "signatures" that are associated with disease severity? Are there individual genes that are associated with severe malaria as opposed to mild malaria? One approach we can pursue is to look for clusters in the data and investigate whether these clusters relate to any of the observed clinical variables. This was the approach pursued by Daily et al. [27], who came to the conclusion

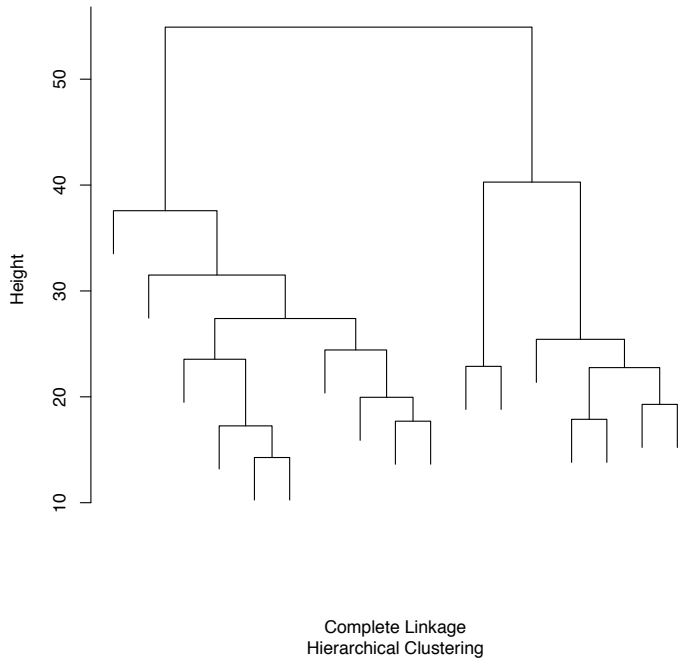

Figure 3.7: Hierarchical clustering of the expression profiles from the cohort of Severe/Mild Samples presented as a dendrogram. Proximity in the resulting tree reflects similarity in observed expression profile. Two clear groups are present; however, these clusters did not associated disease severity of disease.

that there were three distinct expression states in malaria patients.

Clustering of the *ex vivo* samples is shown in Figure 3.7. The dendrogram shows a clear stratification into 2 groups, with possible further subgroups. We sought to identify associations between the clusters observed in the sample and measured clinical variables, but were unable to find any significant associations.

We then tested the possibility that the clusters resulted from an artifact of temporal variation in the age of samples that was not immediately visible by microscopy. Mapping the MLEs onto the samples represented in the dendrogram clearly shows that the basis for the clustering is sample age (left panel of Figure 3.8). The two groups have different distributions of sample age; A t-test reveals that the means of the two groups are not equal (right panel of Figure 3.8).

Further support for the hypothesis that the dominant patterns of transcriptional variation in the dataset are due to temporal heterogeneity comes from comparing correlations to *in vitro* cultured schizonts (Figure 3.9). The *ex vivo* schizonts are ordered by age; as the sample age increases toward that of the *in vitro* schizonts, correlations increase and reach a maximum. As the *ex vivo* samples become older than the *in vitro* schizonts, the similarity between them drops.

Overall, these analyses suggest that the major observed differences in the samples are due to the variable progression into a program of expression which varies over time and asexual development but does not respond to phenotypic changes in the host. The samples closely resemble the expression profiles from *in vivo* cultured samples, and as the differences between *in vitro* schizonts and *ex vivo* mature stages, so too does the estimated temporal distance between them. By the time samples are removed from patients and analyzed 24 - 48 later after maturation in culture, they closely resemble cultured lab isolates. We conclude that there is an absence of large-scale variation present in the mRNA transcription patterns of patient isolates relative to the patterns observed in culture, or if there is, it is rapidly suppressed by the culture methods used.

Up to this point the inclusion of temporal information has served to account for variation in expression values. We can also turn the question around and ask: Once temporal variation has been accounted for, what differences remain?

Figure 3.10 presents one way to approach the problem. We have selected samples from within ± 2 hours post invasion, and compared these against time-matched samples cultured *in vitro*. A t-test was then performed to test whether the means of the different groups are equal, and the top 50 differentially expressed genes are shown.

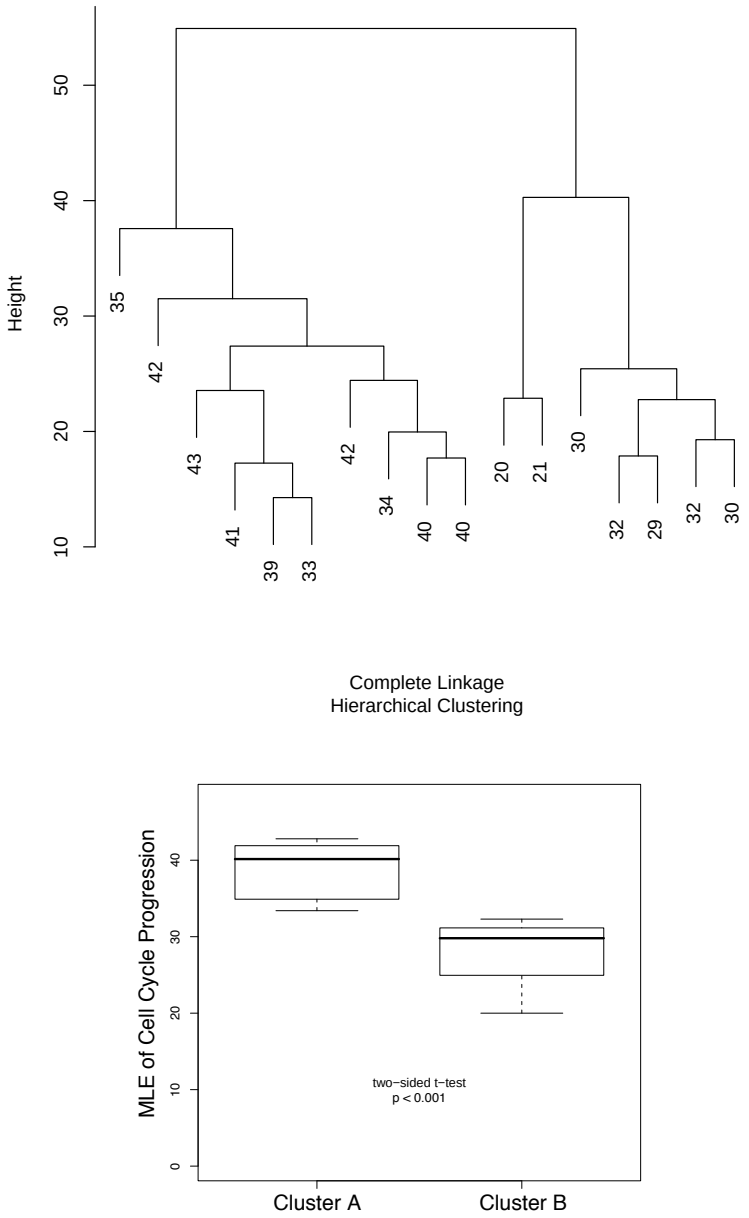

Figure 3.8: The same dendrogram as presented in Figure 3.7, this time with maximum likelihood estimate of sample age labelled for each sample (top). The average temporal development is significantly different between the two groups, suggesting that differences in the expression profiles reflecting variable underlying temporal progression are driving the clustering among samples – there is a strong association between cluster membership and sample temporal progression (bottom).

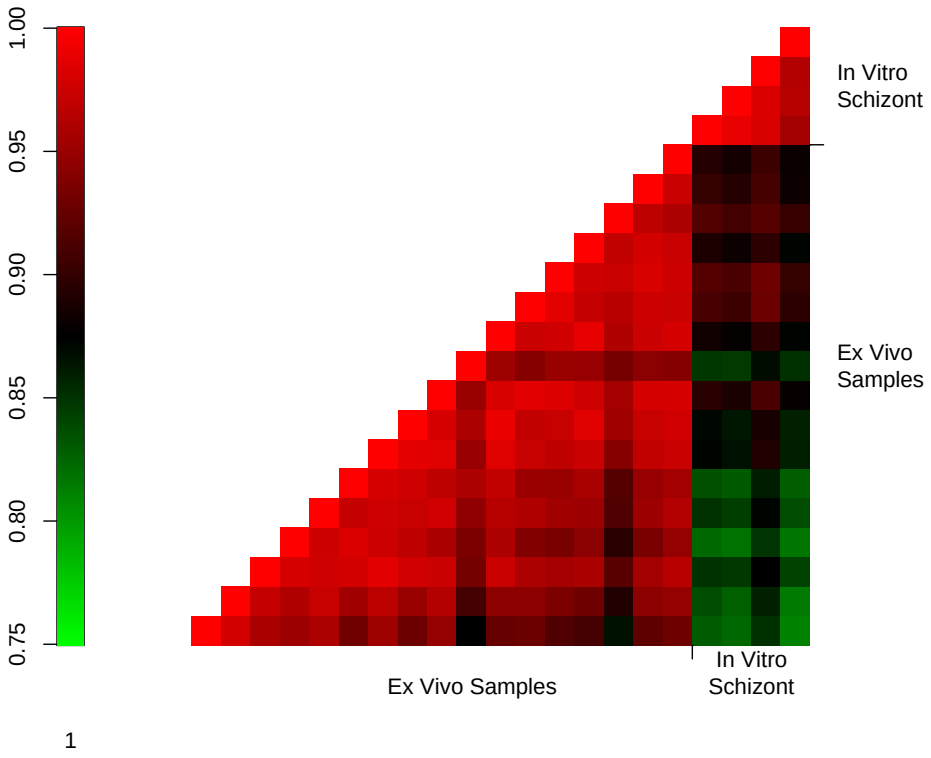

Figure 3.9: This figure shows pairwise Pearson correlation coefficients among *ex vivo* samples. Since the matrix of correlation coefficients is symmetric, only the lower triangular part is shown. The major discrepancies between *ex vivo* samples and *in vivo* can be attributed to different culture age. Pairwise correlations ordered by MLE are shown above. *Ex vivo* samples show increasing dissimilarity to *in vitro* schizont samples as their estimated temporal progression deviates from $\approx$ 42 HPI.

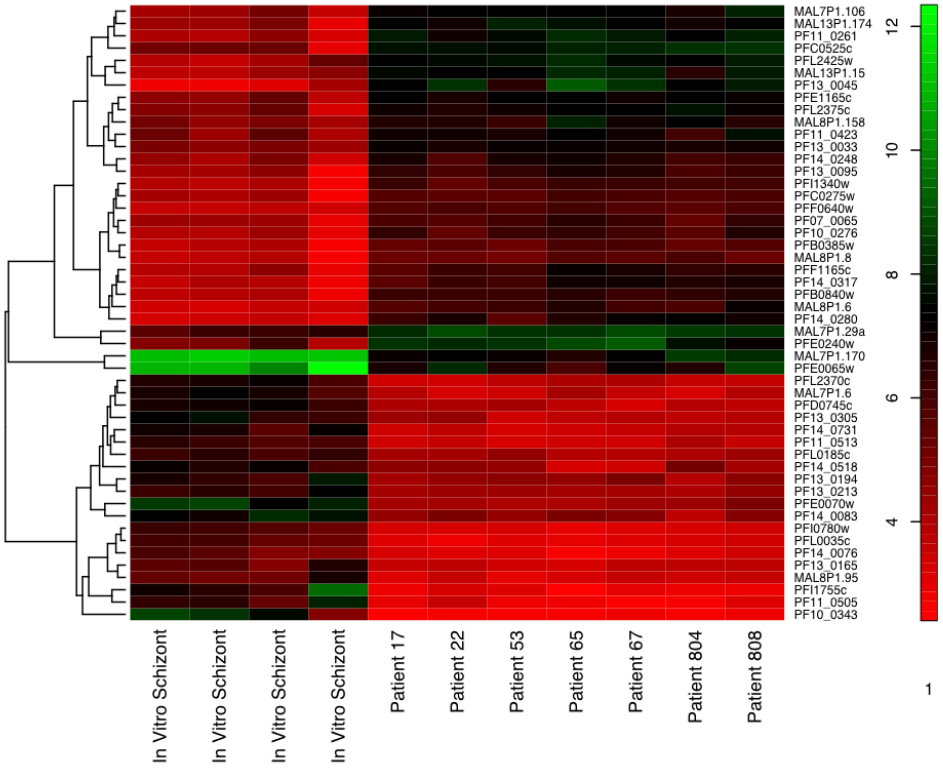

Figure 3.10: Heatmap showing differentially expressed genes between computationally time-matched samples. The RMA expression value of individual genes is given in the rows, which have been individually normalized. Time-matched samples were selected from the *ex vivo* and *in vitro* datasets, and a t-test for differential expression was performed. The top 50 hits are shown above.

3.4 Lineage Commitment

The result above, that the large-scale patterns of expression are similar *in vitro* to those after short-term *ex vivo* culture, differs from the result observed by Daily et al. in reference [27]. While these contrasting results may be explained by the differences in study design, in particular the presence of a brief culture period in our study which was absent in theirs, it was important to examine the differences in detail. Since the similarity between *ex vivo* samples and *in vitro* samples was much more apparent after accounting for stage, we considered whether the same could be true of the samples in reference [27]. To test this, we established approximate ages for the samples, and both maximum likelihood estimates and log-likelihood curves are shown in Figure 3.11.

The samples included in the study from Daily et al. [27] all lie within the region $2 - 16$ HPI, indicating the absence of a systematic association between individual clusters and early or late time points. Nevertheless, the width of the confidence intervals for the clusters does appear to be different, with clusters 1 and 2 of [27] containing broader confidence intervals and decreased curvature of the likelihood curves, suggesting a progressive shift away from an asexual signal. We noticed also that this shift away from an asexual signal corresponded to an increase at $\approx$ 30 HPI, reminiscent of the apparent asexual age for gametocyte samples previously noted (Figure 3.6).

We investigated this further by studying the projection of the samples in the coordinate system defined by the left singular vectors of the gametocyte development space, using the samples from the study by Young and colleagues [152]. In Figure 3.12, we projected the 43 *in vivo* ring-stage samples from reference [27], along with asexual samples obtained using the same scrmalaria microarray [122]. The samples span a continuum in this space, moving from close resemblance to the early (predominantly asexual) time points toward the time points on Day 4

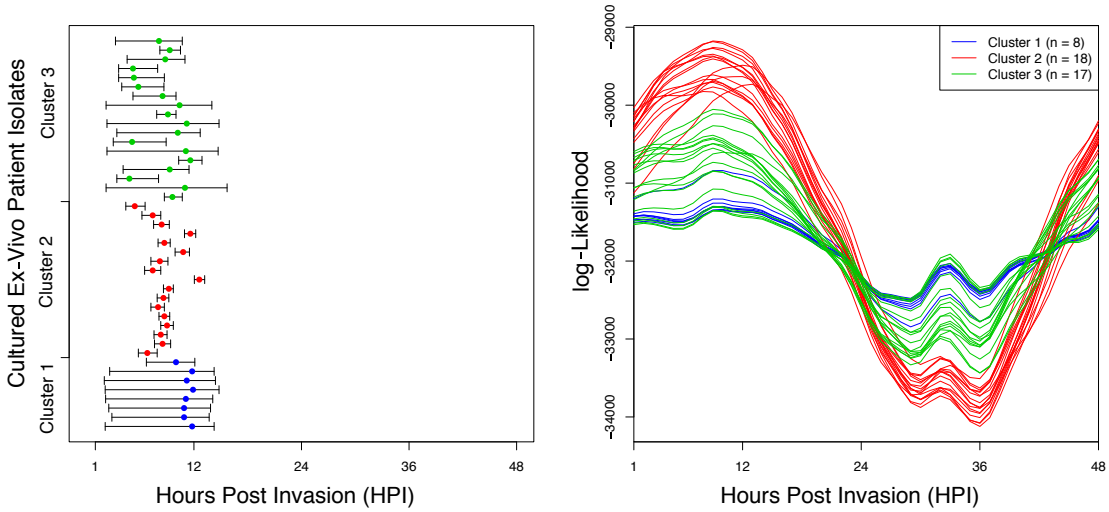

Figure 3.11: Maximum likelihood estimates (left panel) and log-likelihood curves (right panel) are shown for the 43 *in vivo* samples from reference [27] and colored by assigned cluster from that study. The samples are distributed over the ring stage interval, with equal temporal progression between clusters, suggesting that differential temporal progression does not account for cluster membership. A peak can be seen at $\approx$ 30HPI, reminiscent of the apparent asexual age for gametocyte samples, suggestive of gametocyte gene expression in some of the samples.

and 5, which are composed almost completely of immature gametocytes. The samples do not subdivide into discrete clusters, instead falling continously along a line which experimentally corresponds to increasing gametocytemia and gametocyte maturation. This analysis provides further evidence for a smooth progression toward gametocyte-like expression profiles in the samples from reference [27], a trend that had been suggestive from Figure 3.11.

We hypothesized that this result was due to mixtures of both sexual and asexual forms in the culture. We set out to investigate this more directly by extending the stage estimation methods developed in Section C.3.1 to provide estimates both of progression into the asexual lineage (stage estimation) and well as commitment to the sexual lineage. We consider each expression profile as a mixture of asexual and sexual expression patterns, and sought to calculate the correct mixing coefficient.

Recall that in the maximum likelihood approach to estimating temporal progression, we modeled the gene expression vector, $y \in \mathbf{R}^p$, as

$$y = \mu(t) + \epsilon,$$

where

$$\epsilon = \mathcal{N}(0, \mathrm{diag}(\sigma^2)).$$

We can extend that to a two-parameter mixture,

$$y = (1 - \alpha)\mu(t) + \alpha\gamma(s) + \epsilon,$$

which models each expression profile as a weighted sum of asexual (the $\mu(t)$ term) and sexual (the $\gamma(s)$ term) components plus multivariate Gaussian noise with diagonal covariance. The relative amounts of asexual and sexual terms are given by $(1 - \alpha)$ and α respectively. The parameter α is to be estimated from the data;

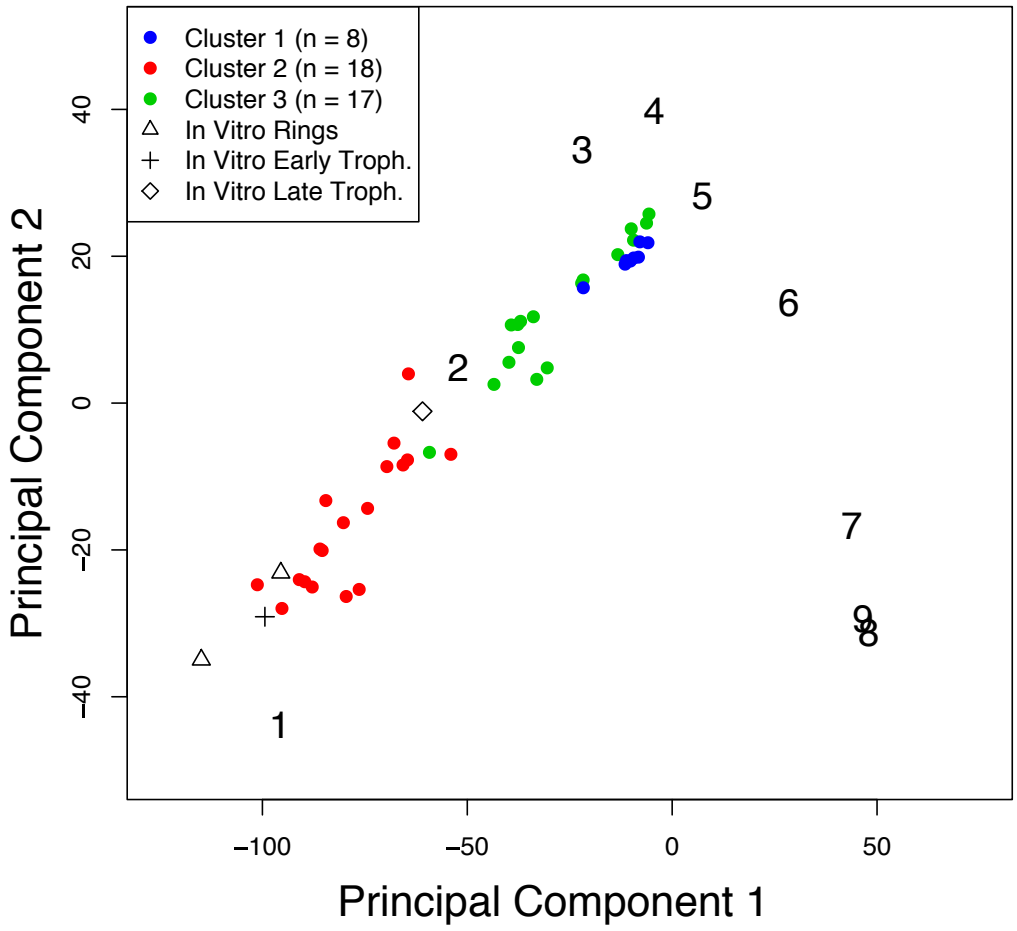

Figure 3.12: Projection of samples from [27] into space defined by first two components of gametocyte development from [152]. The *in vivo* samples form a continuum along the trajectory defined by the early days of gametocytogenesis, suggesting that there is a continuous shift toward sexual-stage gene expression in the sample set. Array-matched controls from ring and trophozoite timepoints (from reference [122]) project between gametocyte days 1 and 2, which are both heavily contaminated by asexual stages.

we refer to it as the gametocyte mRNA fraction. We make a further simplification of the model so that the gametocyte development transcriptome, $\gamma(s)$, is modeled as a single gametocyte-specific profile $\gamma \approx \gamma(s)$. We make this simplification because in fact it is true during middle and late timepoints that $\gamma(s) \approx \gamma$ (c.f. Section 2.6), and also because we do not know what the early gametocyte profiles look like; we only have access to expression profiles of mature gametocytes and so are constrained to use those.

The likelihood function for this model,

$$l(t, \alpha) = \prod_i P(y_i|t, \alpha),$$

is a two dimensional function which can be evaluated on a grid of points. The maximum likelihood estimate for the parameters t and α is the point in the (t, α) plane which maximizes the function. An example for a single sample is given in Figure 3.13. Maximization is done by exhaustively sampling the grid of points. It is worth noting that this is a non-convex function of t and α which is concave for single values of t, i.e. the conditional likelihood $l(\alpha) = P(y_i|t = \tilde{t}, \alpha)$ is concave.

The approximation $\gamma(s) \approx \gamma$ is a weakness of the model, and brings to light an issue which is sometimes called the "Anna Karenina Problem".[1] Although we know that during mature stages, gametocyte gene expression patterns change very little, the assumption that early gametocytes resemble late gametocytes is approximately true for the stages currently measured but remains an important and unvalidated assumption in our model.

We validated the 2-parameter estimates of gametocyte development by testing the ability of the method to detect gametocytemia in samples with known percent-

[1]Tolstoy began *Anna Karenina* with the quotation, "All happy families are alike; each unhappy family is unhappy in its own way." In other words if we have seen one happy family we know what others look like, but the same is not true for unhappy families. In terms of malaria gene expression profiles, we know what all stages of the asexual temporal development cycle look like and we can approximate $\mu(t)$ at all values of t. On the other hand we do not know what gene expression looks like at early stages of the gametocyte development transcriptome.

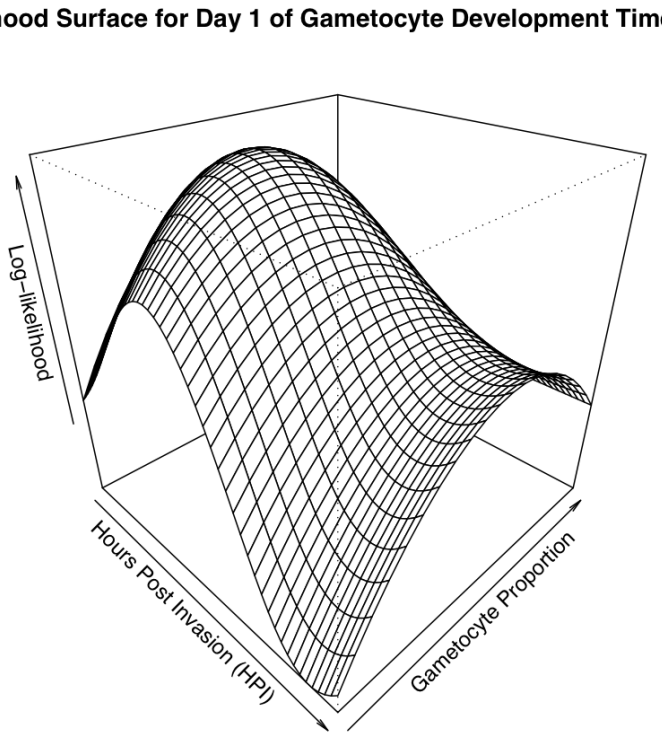

Figure 3.13: An example log-likelihood surface for is shown. The maximum and curvature of this plot are used to obtain the two-parameter maximum likelihood estimate and confidence ellipse. The sample above contains a mixture of ring stage parasites (reflected in the log-likelihood curve which peaks at approximately 12 HPI on the HPI axis, along with approximately 50% gametocyte RNA, reflected in the log-likelihood curve which peaks at $\approx \alpha 0.5$ along the Gametocyte Proportion axis.

ages of gametocytes. The plots in Figure 3.14 show our estimates $\hat{\alpha}$ for samples from [152]. Gametocytemia in those samples was measured by Giemsa staining as well as immunofluorescence against the gametocyte marker Pfs16. Comparing the estimated gametocyte mRNA fraction to the measured gametocytemia shows a strong correlation. Where the estimates diverge we suspect it is due to the fact that α estimates mRNA fraction and not true gametocyte fraction, since different stages are known to produce different quantities of mRNA [102]. In general, however, the estimation of α is an accurate way to detect the presence of gametocytes in an expression profile.

The plots in Figure 3.14 appear to consistently overestimate the number of gametocytes present in early days of the gametocyte development time course. It is interesting to speculate that, in some settings, α may provide a more accurate estimate of the gametocytemia in the culture, and that immature ring forms which are committed gametocytes but morphologically indistinguishable from asexual parasites are not being counted by the staining methods. The relationship between gametocytemia and α is in general complicated, however, because different stages produce different relative amounts of mRNA [102]. Only in situations in which asexual stages that make equivalent total quantities of mRNA per cell to gametocytes, such as mature stages, will our estimates of α closely approximate gametocytemia.

After validation of the 2-parameter estimates, we evaluated temporal progression as well as gametocyte mRNA in the samples from our study as well as the study of Daily et al. [27]. Consistent with the conclusions in Section 3.3, the major variation in the study from our samples derives from temporal variation. However, there is additional variation due to the presence of a small percentage of RNA that appears to come from gametocytemia, a nuance not not seen with the previous analysis. On the other hand, the samples from Daily et al. [27] show substantial variation in the fraction of the expression profile which is attributable

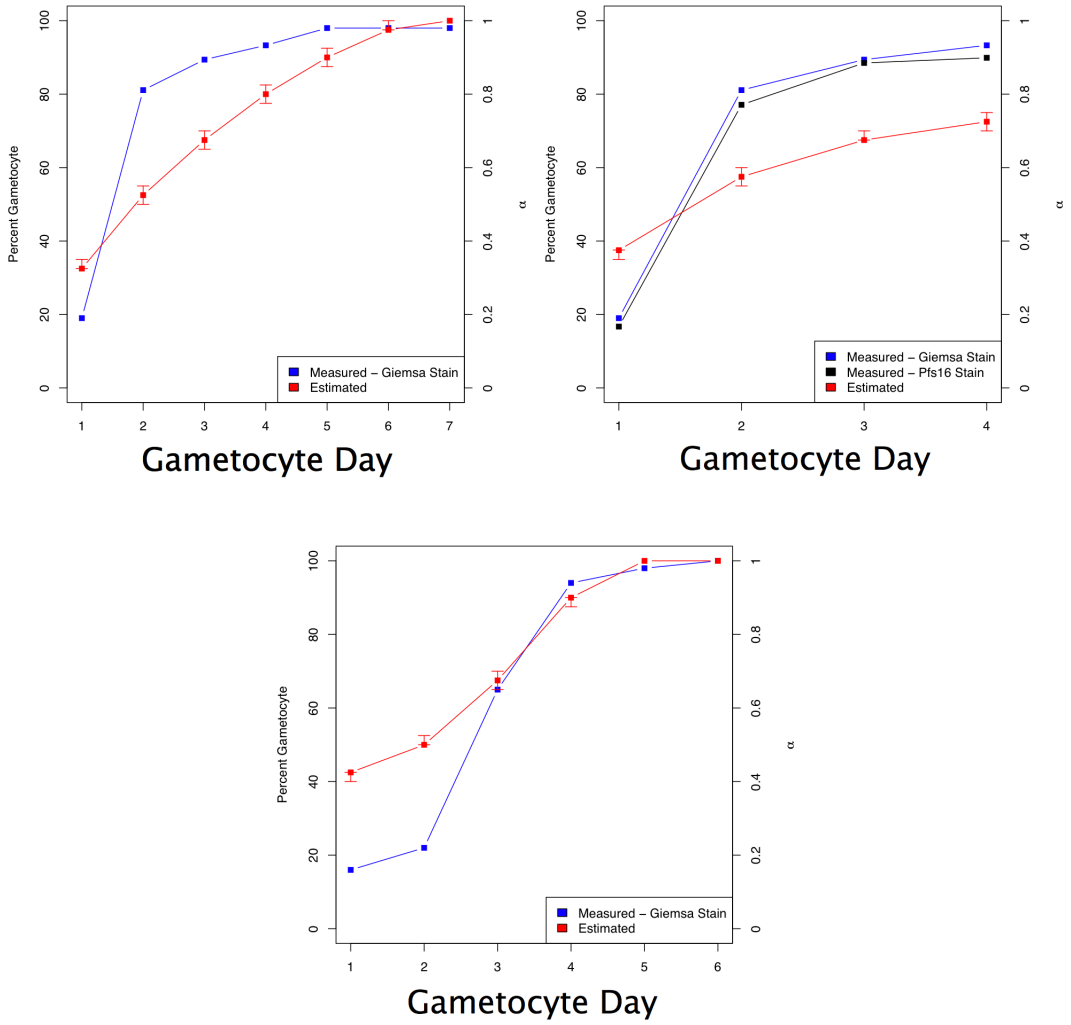

Figure 3.14: Validation of α using samples of known mixtures from reference [152]. Values of α, the gametocyte mRNA fraction, were estimated from samples for which both microarray expression profiles and accurate gametocyte counts (as measured by Giemsa staining and Pfs16 immunofluorescence) were available. Estimated gametocyte mRNA fraction (α) tracks closely with measured gametocytemia. Notably, a consistent overestimate of gametocytemia can be seen at early timepoints, which may indicate detection of committed gametocytes which morphologically resemble asexual ring forms.

to gametocyte RNA but limited temporal variation. This analysis is in accord with the results from PCA. In both studies of patient isolates, the amount of gametocytemia is continuous and roughly uniformly distributed across a range of values.

The two-parameter estimates highlight the commonalities of expression signatures derived from patient isolates. Once varying temporal progression and lineage commitment have been accounted for, the expression patterns observed in culture remain similar to those observed *in vitro*. Rather than a continuous adaptation of the transcriptional program of individual parasites to each patient, it is parasite populations themselves that seem to restructure. All patient infections appear to lie along a continuous gradient of gametocyte commitment, though it is far greater in the *in vivo* parasites than it is in the *ex vivo* isolates. There are (at least) two possible explanations for this. The first is that culture conditions suppress gametocyte growth, and that while the samples which begin *ex vivo* culture may be higher in gametocytemia, after 24 - 48 hours in culture most of the gametocytes have died or stopped producing mRNA. An alternate hypothesis states that since trophozoites and schizonts make more RNA per cell than immature ring forms, it is possible that the ratio of cells may be similar between the studies even if ratio of RNA is not. While the second hypothesis is almost certainly true to some degree, it is not clear whether it accounts for the substantial observed difference between the studies.

Overall, we conclude that every patient infection contains transcriptional evidence of gametocytes, and that the amount of gametocyte-specific transcription is almost uniformly distributed across an interval. We attribute this to a relatively large population of ring-stage parasites which will ultimately become morphologically identifiable gametocytes, but have already begun expressing the gametocyte transcriptional program.

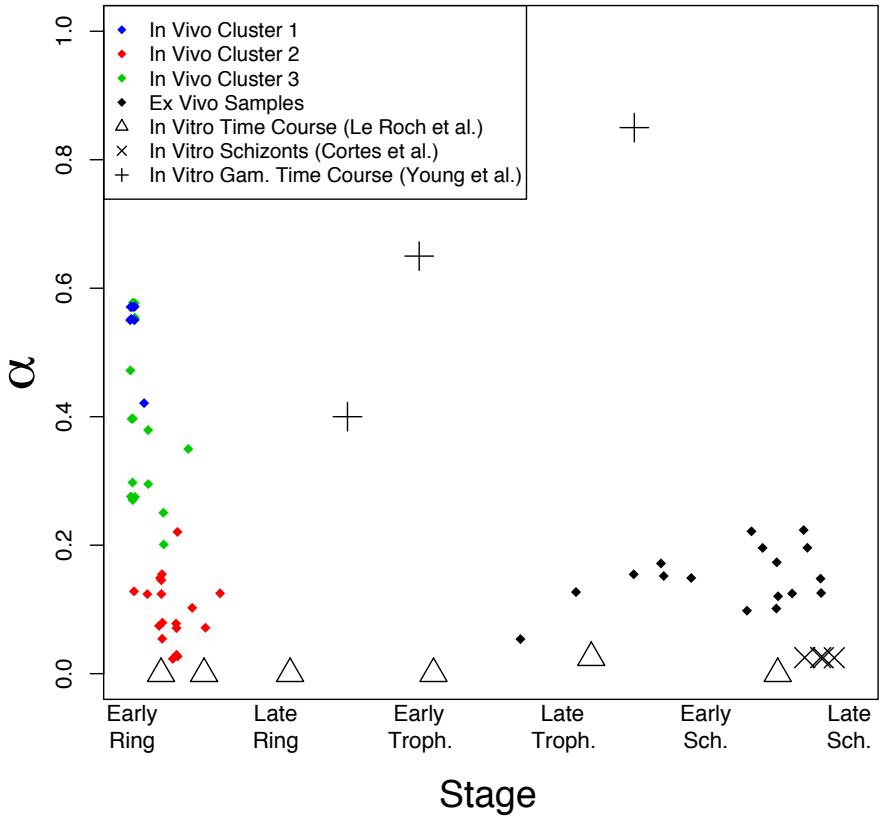

Figure 3.15: Estimates of α HPI for samples from Lemieux et al. [85], Daily et al. [27], Le Roch et al. [122], Young et al. [152], and Cortes et al. [21]. Each 2-parameter estimate reflects the coordinates at which the 2-dimensional likelihood function obtains its maximum (c.f. Figure 3.13). All samples from patients have a distribution of α values as well as a distribution of estimated asexual ages. The samples from [27] show substantial variation in the gametocyte fraction and limited temporal variation, while the samples from [85] contain more variation in temporal progression but less in gametocyte mRNA. *In vitro* controls of various stages are also plotted for comparison.

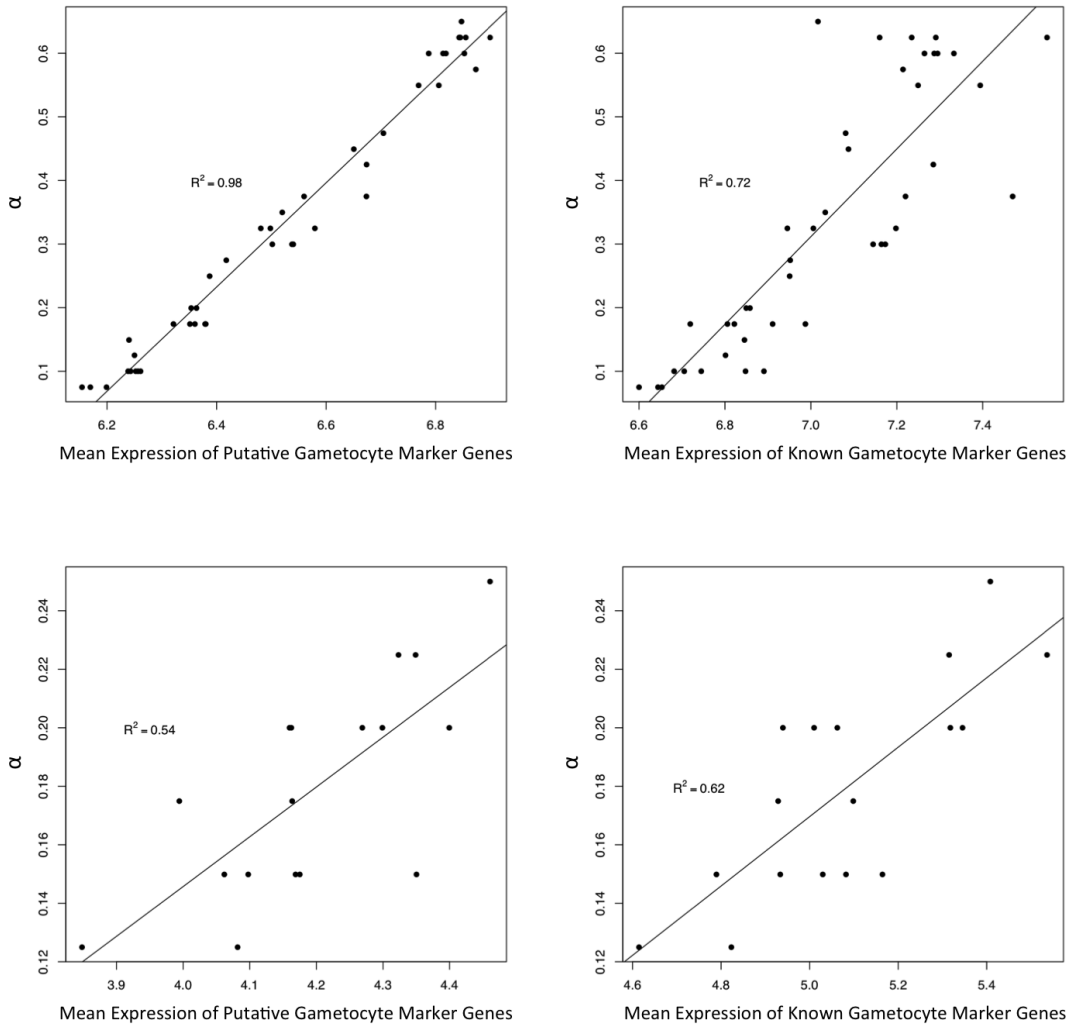

Figure 3.16: Correlation of α with gametocyte marker genes using the two sets of gametocyte marker genes from reference [152]. The genes in the right column plots are 15 gametocyte marker genes which are known from the experimental literature. The left column uses the 246 genes identified algorithmically by Young and colleagues [152] as having similar expression profiles to known gametocyte marker genes. The top row shows relationship between α and the mean expression of gametocyte marker genes for the *in vitro* samples from Daily et al. [27], whereas the bottom row shows the relationship for *ex vivo* samples introduced in Chapter 2. A strong, linear relationship can be seen for all sets of samples, providing further evidence that the samples from Daily et al. [27] are contaminated with committed sexual stages, and also documenting that α, obtained using the gene expression pattern of all genes in a sample, accurately reflects the patterns of gamtecytemia that would have been inferred using gametocyte-specific marker genes.

3.5 Batch Effects

The methods described above for estimating the temporal progression of samples, and the results obtained by applying them to several datasets, were published in reference [85]. When the study was published, Wirth and colleagues wrote a reply [148] questioning some of the points of the study, mainly those in which the data from reference [27] was reanalyzed. The main area of disagreement was that in our analysis we concluded that each patient infection could be placed along a continuous spectrum of mixtures of sexual and asexual stages, whereas their analysis had identified three discrete clusters in the dataset.

We had addressed this question in [85] by showing that their analysis, which included samples hybridized in batches, showed substantial effects of hybridization date on the data. Arrays hybridized in 2004 showed a distribution of fluorescence characterized by a larger variance, whereas those hybridized in 2005 showed a distribution with smaller variance (Figure 3.17). These discrete groupings of hybridization intensity were non-randomly associated with cluster assignment. In our view their clusters result from the discrete variation due to batch superimposed onto the continuous variation in gametocyte commitment in the samples.

We performed a more sophisticated analysis in reference [84] in response to the comments from Wirth et al. [148]. We obtained the date of hybridization from the header lines of the .cel files which contain the fluorescence intensities for each probe. The hybridization dates revealed that the 43 arrays were processed in four batches. Figure 3.18 shows that batch membership was strongly predictive of cluster membership ($p \ll 0.001$). Treating each batch as a single experiment, within each batch, our estimates of α correlated strongly with the first principal component, suggesting that the major source of variation within batch was gametocyte mRNA commitment (Figure 3.18).

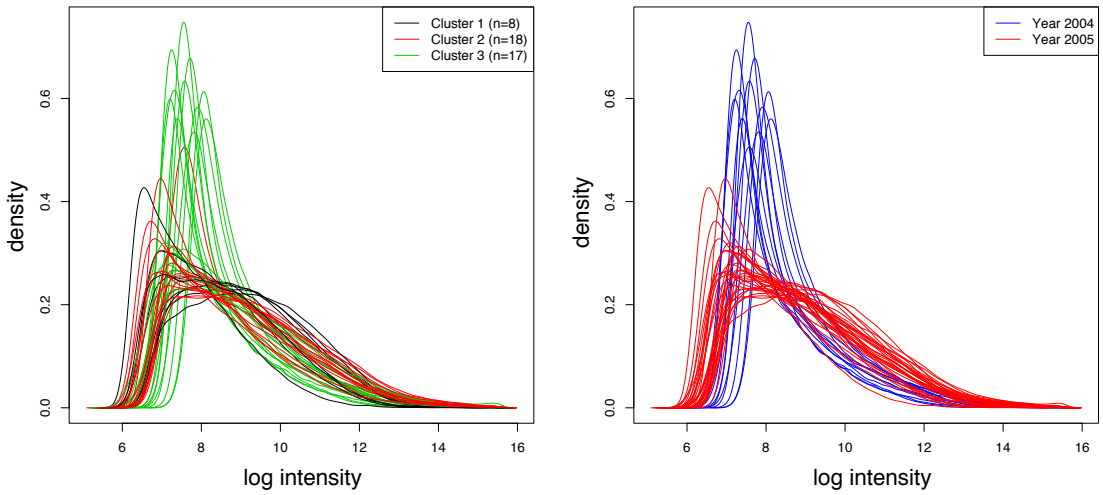

Figure 3.17: The distribution intensity values for each array is shown for the samples from reference [27]. The intensity distribution should be roughly equivalent for each hybridization, and while some differences can be corrected in normalization procedures, gross differences will impact the global expression profile and influence the study conclusions. The same curves are displayed in both the left and right panel; however, in the left panel they have been colored by cluster whereas in the right panel they have been colored by the year of processing. Samples processed in year 2004 have a more peaked intensity distribution of lower variances while the samples processed in 2005 have a broader distribution that is less peaked, reflective of systematic differences in sample processing between these times. The differences in sample processing correlate with cluster assignment. For example, the peaked profile observed in year 2004 is disproportionately associated with membership in cluster 3.

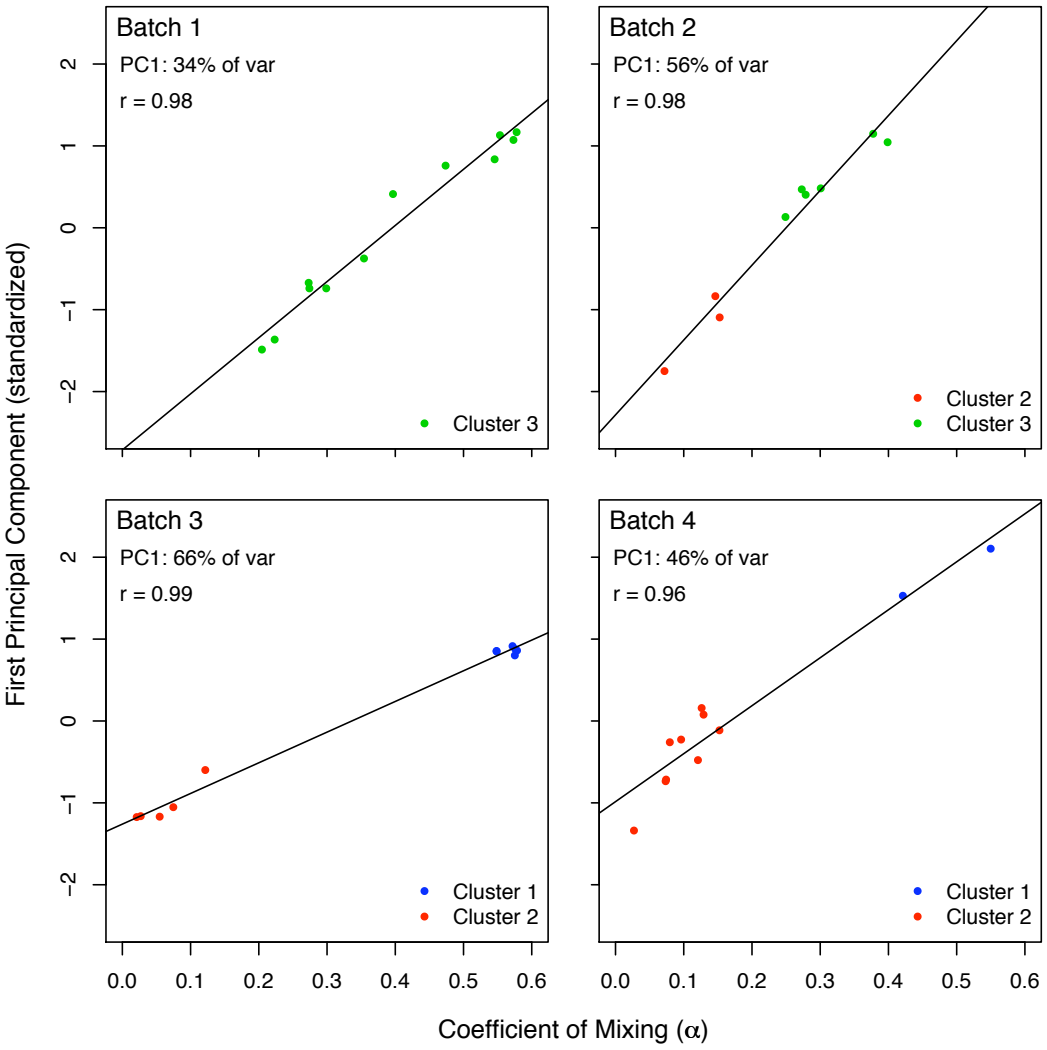

Figure 3.18: Affymetrix arrays record their date of hybridization directly on the array in time stamp which can be extracted during analysis. When the date of hybridization for each sample is measured, processing in four batches was apparent, and the batches were strongly predictive of cluster membership. For example, all samples processed in the first batch were classified into cluster three, suggesting that the clustering approach of Daily et al. [27] is detecting the technical similarity due to batch processing more strongly that biological variability in the sample. In this figure, for each date of processing, α estimates for each sample from Daily et al. [27] are plotted on the x-axis, while the first Principal Component of the dataset. There is a strong, linear relationship between α and the first Principal Component, suggesting that the dominant source of variability in this dataset *is* α. Yet the slope and intercept of this relationship differ between samples, likely as a result of experimental handling. This suggests that α, or gametocyte mRNA fraction, is the correct biological interpretation of the variability observed in the study by Daily and colleagues [27], but that this relationship can be obscured if the confounding effect of batch is not removed.

3.6 Conclusion

Each malaria infection contains a heterogenous population of cells which are genetically similar but phenotypically distinct. Individual cells vary in terms of their developmental lineage (asexual vs. sexual) and their temporal progression within this lineage. This heterogeneity at a population level makes interpretation of expression data difficult. This is a challenge which has affected studies in our own lab as well as studies published in the primary literature. One way to fix this problem is to infer temporal progression and lineage commitment from observed expression data. In this chapter and the accompanying appendix, we presented several methods which perform that function. We also compare these methods to other approaches used in the literature, highlighting similarities and differences between our approaches and those used, for example, by PlasmoDB.

After considering multiple approaches to stage and lineage estimation, we select one method (the maximum likelihood method), demonstrate its robustness and accuracy, and use it to improve the analysis of the severe vs. mild study presented in Chapter 2 as well as the study by Daily and colleagues [27]. Our results identify genes which are upregulated in patient samples, highlight reservoirs of early, committed or submicroscopic gametocytes present in nearly all patient infections, and suggest a model in which the circulating parasite burden of every patient infection contains mixtures of asexual and committed sexual stage parasites. In comparing the evidence in favor of our model to that of Daily and coworkers, we identified the presence of batch-processing artifacts in their dataset which appear to have contributed to the presence of a putative third state.

Chapter 4

Chromatin and Gene Regulation

4.1 Introduction

In the previous two chapters, we investigated the global patterns of expression among *P. falciparum* genes. By examining the levels of parasite transcripts in patients with severe and mild malaria, as well as by reanalyzing publicly available datasets, we emerged with a picture of a highly structured, but nevertheless predetermined, expression program. During asexual development, the majority of genes are expressed periodically; however, when gametocytogenesis begins, gametocyte-specific genes are induced and maintained at a constant level. These contrasting patterns, and the apparently seamless transition between them, points to extensive regulatory mechanisms that must exist in the parasite. Nevertheless, we know surprisingly little about these regulatory mechanisms – what they are, and what their relative importance is in relation to one another.

Our goal in the remainder of the thesis is to investigate mechanisms that control the elegant, structured patterns that were observed in the preceding chapters. This is too big a task to be solved completely, but by studying two types of regulatory mechanisms—chromatin state and chromosome folding—we can begin to learn about their relative importance in determining parasite gene expression. And with the advent of second-generation sequencing technologies, it is possible to do this in a genome-wide capacity, enabling inference which may be valid on a genomic level, not only at the level of individual genes.

In the first part of this chapter, we review what is currently known about gene regulation in *Plasmodium*. Then, in experimental models, we begin to study the epigenetic control of gene expression with a focus on the *var* gene family. Taking advantage of recent technical advances in DNA sequencing, we map the global distribution of histone marks in the IT isolate using chromatin immunoprecipitation followed by sequencing, a technique known as ChIPseq. Shared patterns of chromatin modifications reinforce the idea that the genome is partitioned

into euchromatin, enriched in activating marks, and heterochromatin, enriched in silencing marks. Interestingly, these areas of distinct chromatin state are not continuous along the length of the chromosome, suggesting that an understanding of epigenetic processes in the parasite requires a knowledge of higher-order chromatin structures, which are the subject of the next chapter. The focus of this chapter is on defining the basic epigenetic landscape upon which higher chromatin structures are built.

4.2 Background

4.2.1 Established Mechanisms of Control

Transcription of *P. falciparum* genes is mono-cistronic and controlled at the level of RNA polymerase [80, 81]. Consistent with the presence of expected transcriptional machinery, molecular studies of *Plasmodium* transcription have demonstrated that the components appear to function as they do in other eukaryotic organisms. Horrocks and colleagues, summarizing the results of functional promoter assays on 22 genes described in the literature [66], document the presence of directional promoters upstream of transcribed regions which possess both activation and repression sequences, similar to other eukaryotes. Nevertheless, the molecular factors which bind to these upstream are not entirely known. Following completion of the genome project, sequence-based searches of homology failed to identify specific transcriptional regulatory factors, a surprising finding given the known complexity of gene expression patterns [9]. Later studies suggest that some specific transcription factors are present but were initially missed because of limited primary sequence homology. Balaji and colleagues used sensitive domain-based searches to identify a set of 26 candidate transcription factors similar to the plant AP2 transcription factor family [3]. These transcription factors appear to

have undergone a radiation in the *Apicomplexa*, and have recently been shown to bind to promoter elements in *P. falciparum* [16,134], though their role in the regulation of specific genes, and in orchestrating the genomic program of the parasite as a whole, remains to be experimentally established.

The parasite contains almost all the components of the basal transcriptional machinery, including RNA polymerase II [22]. In the face of observed transcriptional complexity, and the uncertainty around the function of AP2 proteins, it has been suggested that the parasite uses other means to regulate its gene expression patterns, such as epigenetic control of gene expression [22]. The case in favor of epigenetic control has also been bolstered by several lines of evidence. The drug apicidin, a potent inhibitor of parasite histone deacetylase activtity, was shown to have anti-parasitic effects [29], demonstrating the importance of proper histone modifications for survival in the parasite. Furthermore, the continuous switching among *var* genes along with reversible silencing highlighted a possible role for epigenetics in *var* gene regulation [126]. Later, with the discovery of PfSIR2-dependent silencing of *var* genes (discussed in the Introduction), the importance of epigenetic factors was confirmed.

The *var* genes are some of the best studied genes in the malaria genome, and the effort to understand the factors controlling *var* gene expression and switching has led to much of our current understanding about mechanisms of gene regulation in malaria. Our current understanding of these mechanisms was outlined in Section 1.3. While much is understood, many open questions remain, and we take up several of them here. Because of their importance in so many critical aspects of malaria biology, *var* genes remain one of the main focuses of our research. On the other hand, we also seek to generalize results from particular gene families to the genomic level.

Based on earlier results indicating a role for epigenetic regulation of particular gene families [17, 36, 50, 126], we hypothesized that epigenetic factors play a

major role in regulating gene expression in *P. falciparum*, and that this occurs in part through higher order chromatin structures and the spatial properties of the nucleus. We therefore set out to investigate those mechanisms, first by examining the epigenetic changes at a single locus and then extending these studies genome-wide. By conducting genomic experiments with cloned lines of *P. falciparum* in which we can select for the expression of a particular *var* genes, it is possible to investigate gene regulation in general as well as for particular *var* genes under study.

4.3 Results

4.3.1 Epigenetic mechanisms are active at the A4 var locus

Using cloned lines of the IT strain of *P. falciparum*, we selected parasites for the expression of the A4var by repeated selection with the BC6 monoclonal antibody [137]. This is a monoclonal antibody that recognizes a portion of the PfEMP1 encoded by a sub-telomeric *var* gene, termed A4var, on the beginning of chromosome 13. After three rounds of selection, we assayed levels of two chromatin marks, H3K9me1, and H3K9me3, across the locus of an active and inactive *var* gene. The results indicate lower levels of H3K9me3 and H3K9me1 along the length of the locus in the active *var* gene (Figure 4.1), including modifications which stretch through to the adjacent *var* gene, R29. The H3K9me3 is consistent with other results in the literature for transgenic parasites [18] and the 3D7 strain [89]. We also noticed a depletion of H3K9me1 at the active *var*, which is different from the published result [18]. It is possible that this is due to cross-reactivity with the H3K9me3 antibody, or it may be that the modification patterns can vary between parasite strains.

We next assayed the time-dependence of modification patterns. Based on our

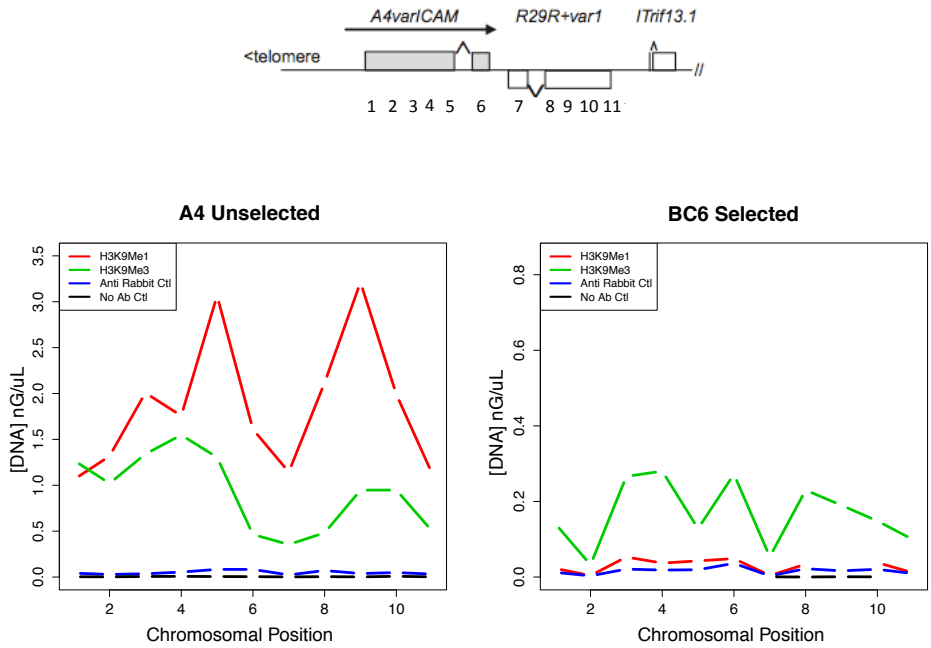

Figure 4.1: Native Chromatin Immunoprecipitation (nChIP) of cloned lines of the IT strain selected for different var gene expression patterns. The distribution along the A4/R29 locus (schematic of A4/R29 locus used with kind permission of S. Kyes, from reference [76]). The qRT-PCR primers are tiled across the locus and correspond to the numbers in the locus schematic. H3K9me1 is several-fold enriched in A4 unselected parasites relative to BC6 parasites which predominantly express the A4 var gene. The scale is an absolute measure of immunoprecipitated chromatin; note that the scale of the y-axis differs between graphs.

previous study which emphasized the rapidly occurring changes in transcriptional state, we sought to profile the occupancy of modifications at high temporal resolution. In order to do this, we sampled the 3D7 reference strain every six hours for a total of eight timepoints. The results, summarized in Figure 4.2, reveal that chromatin marks are highly dynamic during the asexual lifecycle. At the locus of the gene PFA0005w, levels of trimethylation at lysine 9 of histone H3 varied more than 20,000 fold (figure 4.2, *right panel*), mono-methylation and acetylation at that same residue by more than 150 fold, and H3 and RNA pol II varying 50 and 30-fold, respectively. By initial inspection, there seem to be three classes of modification patterns present: elevated in late rings (H3K9me3), elevated in late rings and schizonts (H3K9meAc, RNAPolII), and elevated in schizonts (H3 and H3K9Me1). These classes of patterns seem to correspond to the periods of transcriptional activity (Figure 4.2, *lower panel*). It is not clear from these data whether marks are being added to existing nucleosomes or nucleosomes are being repositioned in the site under study.

The reference strain, 3D7, was selected for these assays because we wished to expand our study of chromatin modifications genome-wide. In doing so, we wanted to take advantage of recent, dramatic increases in sequencing capacity. Second-generation sequencing technologies provide amazing capacity to study the patterns of epigenetic modifications at single-base resolution, and in collaboration with the Pathogen Sequencing Unit (PSU) at the Sanger Institute we used Illumina sequencing to begin to study the epigenetic landscape in malaria. The ChIPseq data from the 3D7 timecourse is not yet available for analysis, and we plan to analyze the results of this experiment in future work. In the next section, we describe the results of ChIPseq studies in the A4 clone.

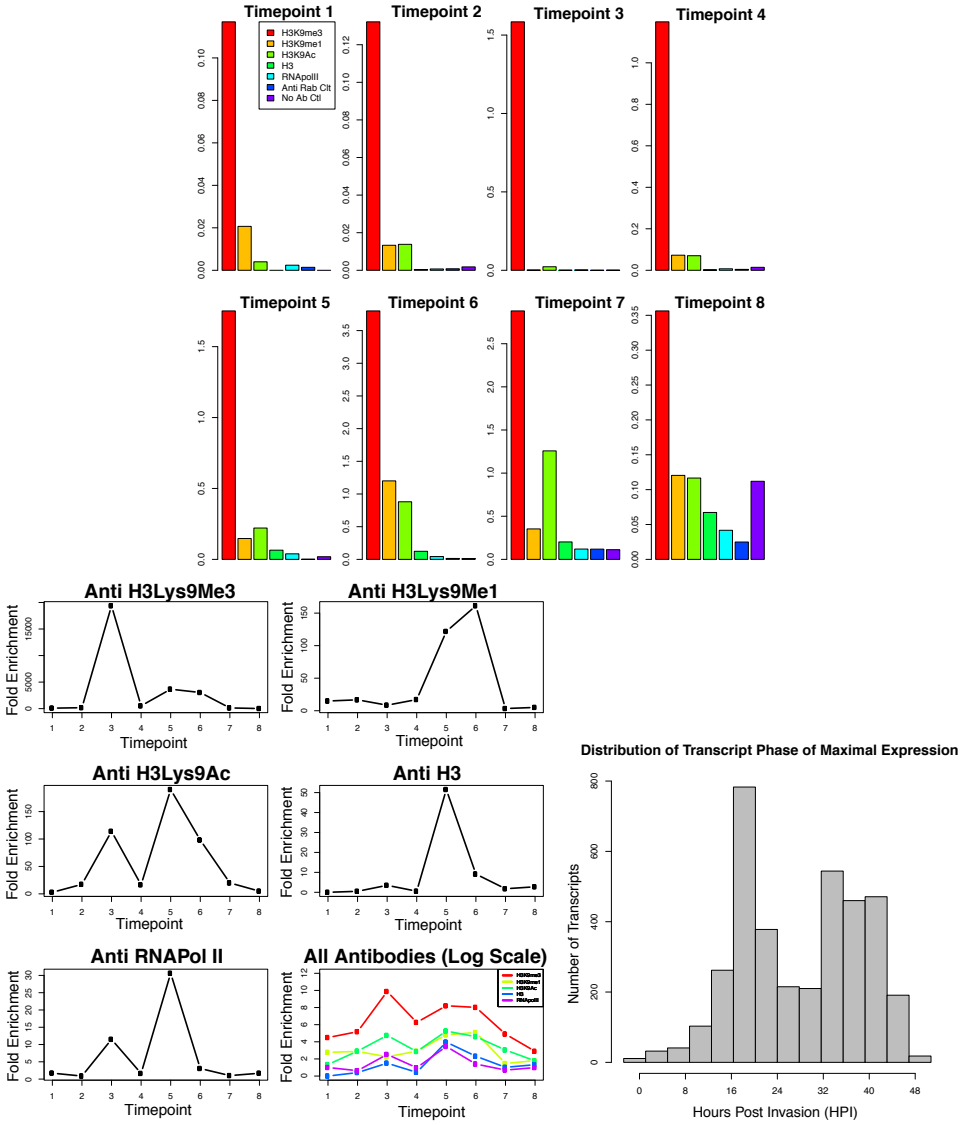

Figure 4.2: Chromatin immunoprecipitation (ChIP) consists of purifying protein complexes and their associated DNA from solution using antibodies, and then quantifying the bound DNA. The absolute quantitity of immunoprecipitated DNA (log scale) for 6 antibodies and beads-only control at the 5' region of exon1 for PFA0005w in 3D7 at each timepoint *(upper panel)*. H3K9me3 is the dominant modification at the *var* exon. The ratio of specific antibody to a non-specific anti-rabbit IgG control is shown for the eight timepoints *(lower left panel)*. The level of modifications is dynamic during the lifecycle and is unique for each modification studied. There appear to be two peaks of modification activity during the lifecycle. These may correspond to the peak transcriptional activity *(lower right panel)* (data from reference [9]).

4.3.2 Genome-wide Assays for Chromatin State

In order to assay the distribution of chromatin modifications in a genome-wide manner, we paired chromatin immunoprecipitation with second-generation sequencing on the Illumina platform. These experiments, done in collaboration with the Pathogen Sequencing Group at the Sanger Institute, allow us to profile the relative quantity of histone modifications at single base-pair resolution. Using the IT isolate, we studied two marks, H3K9me3, which is a silencing mark, and H3K4me3, an activating mark. Table 4.1 shows the numbers of reads sequenced for each of the samples.

Generating libraries for sequencing given the small amounts of immunoprecipitated DNA presented technical challenges, because the amplification procedures that are typically used for ChIPseq rely on PCR-based amplification prior to sequencing. The PCR conditions used in these protocols amplifies DNA in a biased fashion, resulting in an over-representation of GC-rich sequences in the resulting sequence data. This effect can be mitigated by modifying the PCR conditions to include a reduced-temperature extension step [91], but the GC bias is not eliminated entirely. After an initial sequencing attempt generated data contaminated with extensive GC bias, the Pathogen Sequencing Group optimized their library generation protocols. In particular, use of the Kapa HiFi polymerase reduces GC bias sufficiently such that high quality ChIP sequencing libraries can now be generated in *P. falciparum*. The ChIPseq libraries described below were sequenced using the improved protocols.

The genome-wide distribution of H3K9me3 and H3K4me3 is shown in Figures 4.3 and 4.4, respectively. The occupancy of H3K9me3-modified nucleosomes shows enrichment at telomeres and internal antigen clusters. This is consistent with the results of [90] and [125], which showed a nearly identical pattern of modification for H3K9me3 in the strain 3D7. The H3K9me3 is a known silencing

mark and it was shown in [90] that this modification demarcates heterochromatin and colocalizes with silenced foci at the nuclear periphery.

In contrast, the pattern of H3K4me3, shown in Figure 4.4, is nearly opposite to that of H3K9me3. This mark, an activating mark [125], localizes to internal portions of chromosomes but is reduced at telomeres and internal antigenic gene clusters. The opposing pattern of these two marks suggests a genome which is largely partitioned into two regions. One region contains the transcriptionally silenced heterochromatin which spans the telomeres and antigenic gene clusters; the second region encompasses the remaining portions of the chromosomes and contains genes in euchromatin which are either transcriptionally active or available for transcription. Furthermore, the identification of similar patterns of histone marks in the 3D7 strain [90, 125] suggests that this a general feature of the *P. falciparum* which is conserved among isolates.

In addition to providing a global view of the residency of epigenetic marks, the ChIPseq data also offer impressive spatial resolution which extends down to single base pairs. Figure 4.5 shows H3K4me3 coverage (thick blue line) as well as the micrococcal nuclease-digested input control (thin green line) for a portion of chromosome 2. The fine spatial resolution of the data allows identification of individual nucleosomes in both the input control as well as the H3K4me3 sample. The nucleosome peaks are consistent between samples, and nucleosomes enriched in H3K4me3 are clearly visible in the data, particularly at the 5' end of the coding sequence.

The percentage GC content of the DNA is shown in a thin black line above the plot in Figure 4.5. Because of the small quantities of immunoprecipitated DNA generally available in ChIP studies, amplification must be performed. Preferential representation of GC-rich regions as an artifact of the exponential PCR amplification process is a concern. Several of our initial sequencing libraries were affected by this problem, which was subsequently corrected by altering the library prepa-

ration pipeline at the Sanger Center. The libraries used in this analysis, however, demonstrate that the problem has been effectively corrected by the new library preparation procedure. Sequence is available for relatively GC-poor regions of the genome, and GC-rich regions are not overrepresented, suggesting that the Illumina sequencing protocol effectively captures the genome irrespective of base composition. While there is some evidence of a weak GC bias in the data, it is not a large effect and may also reflect the sequence preferences of nucleosomes. Overall, the data generated by ChIPseq offers greatly improved genomic and spatial resolution relative to ChIP-qPCR or ChIP-chip assays and yields important insights into the global and fine-scale structure of the epigenetic landscape.

The data were collected for the IT clone and here have been mapped to an IT reference assembly developed by Thomas Otto and colleagues in the Pathogen Sequencing Unit at the Sanger Center. Initial mapping of sequence to the 3D7 reference was sub-optimal because of the genome changes that have occurred in the IT strain. After much concerted effort by the Pathogen Sequencing Unit, we were able to generate an improved IT reference sequence suitable for mapping short read data. This greatly facilitates the genomic study of gene expression using this parasite, as we discuss in the next section.

Studying the associations between ChIPseq libraries is also biologically and technically informative. Figure 4.6 gives scatterplots of the coverage values for each base on chromosome 1 between the various libraries sequenced. The strong correlations between input libraries from different samples provide evidence of the reproducibility of the technique, while the weaker correlations but still significant correlations between other samples quantify the similarity between marks and baseline nucleosome occupancy. For example, the weakest associations are observed between H3K9me3 and the corresponding input control, suggesting that most nucleosomes are not marked by H3K9me3. On the other hand, there is a relatively stronger correlation between H3K4me3 and its input control sample,

Modification	Read Type	Total Reads	Mapped	Percent Mapped
H3K9me3	75bp PE	18377682	16901024	92.0
H3K9me3 Input	75bp PE	18699984	17588855	94.1
H3K4me3	75bp PE	17452786	16484248	94.5
H3 K4me3 Input	75bp PE	21670506	16989865	78.4

Table 4.1: The read count and mapping statistics for ChIPseq libraries.

suggesting that more nucleosomes contain H3K4me3 than H3K9me3.

In summary, the global pattern of chromatin modifications in the IT isolate describes a genome partitioned between heterochromatin, which predominates at sub-telomeres and internal *var* clusters, and euchromatin, which encompasses the rest of the genome. Many non-adjacent regions seem to share chromatin status. This finding, along with the known foci of heterochromatin at the nuclear periphery, suggest that the spatial relationships of the nucleus are of central importance in determining the epigenetic state of genes and in coordinating gene expression.

4.4 Conclusion

In the previous two chapters, we described some of the difficulties involved with studying gene expression in malaria which result from the confounding effect of temporal and lineage heterogeneity between cultures. Having devised methods which are successful in overcoming these issues, we turned our attention to studying the molecular mechanisms which control gene expression in the malaria parasite. In this chapter we reviewed known mechanisms of gene expression in the malaria parasite, with a particular focus on the *var* gene family because of its central importance in immune evasion and pathogenesis, as well as its unusual expression pattern of mutually exclusive expression and continuous switching.

We then investigated the epigenetic changes which occur during *var* gene switching. We show that selection of the IT parasite line with an antibody against

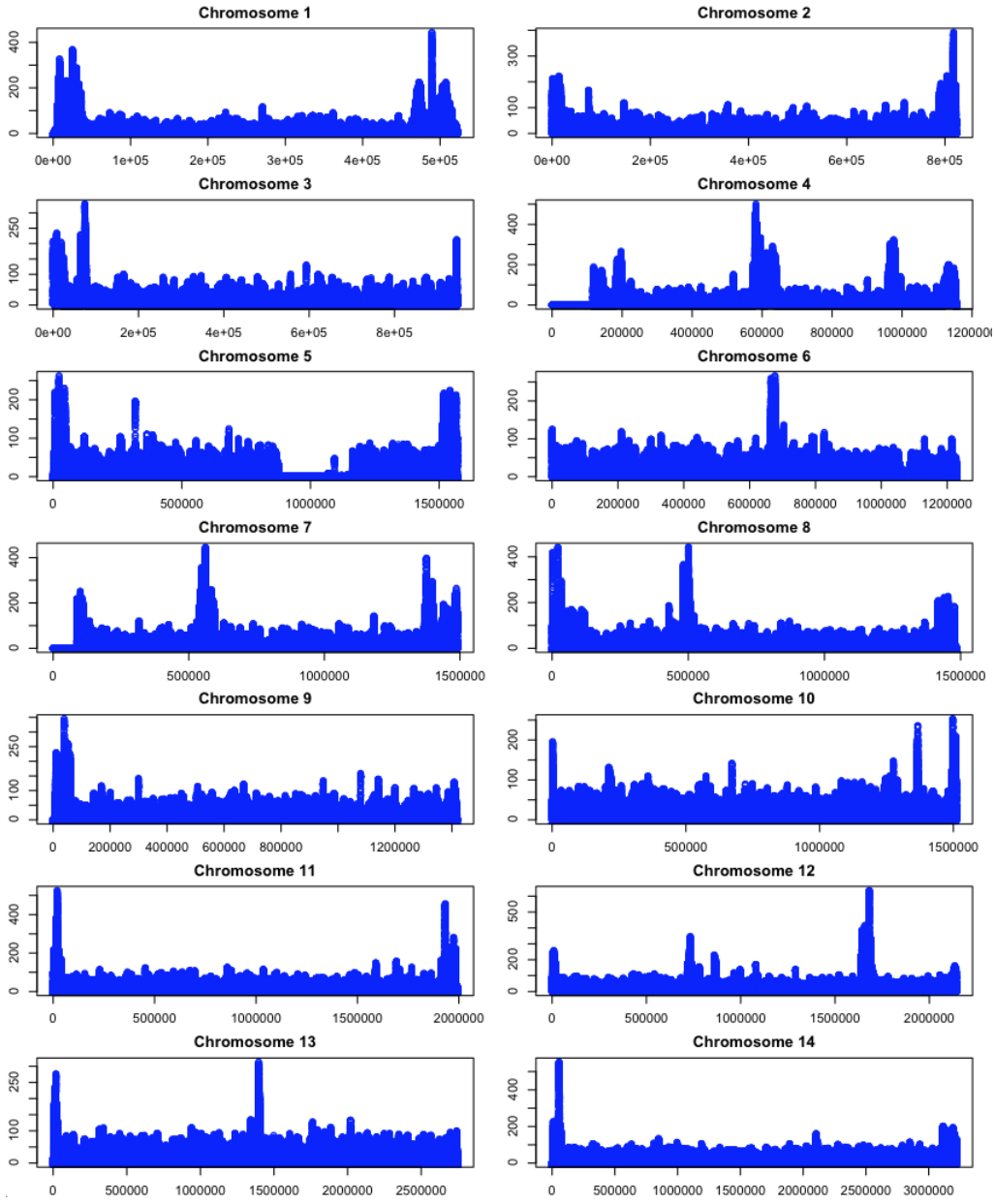

Figure 4.3: In conventional ChIP (as in Figures 4.1 and 4.2), DNA is quantified by hybridization methods of quantitative PCR. A related method involves sequencing the immunoprecipitated DNA and using the summed read counts, a metric also known as coverage, as a measure of the genomic occupancy of a particular modification. This approach offers a significant advantage in that the entire genome can be surveyed at base-pair resolution. This figure shows the distribution of H3K9me3 in the IT parasite clone. The data are expressed as reads in the H3K9me3 library with the input control signal subtracted. Greater numbers of H3K9me3-enriched nucleosomes are seen at telomeres and subtelomeric regions along with internal clusters of *var* genes. The pattern observed for this mark in the IT isolate is nearly identical to that observed by the authors in references [90, 125].

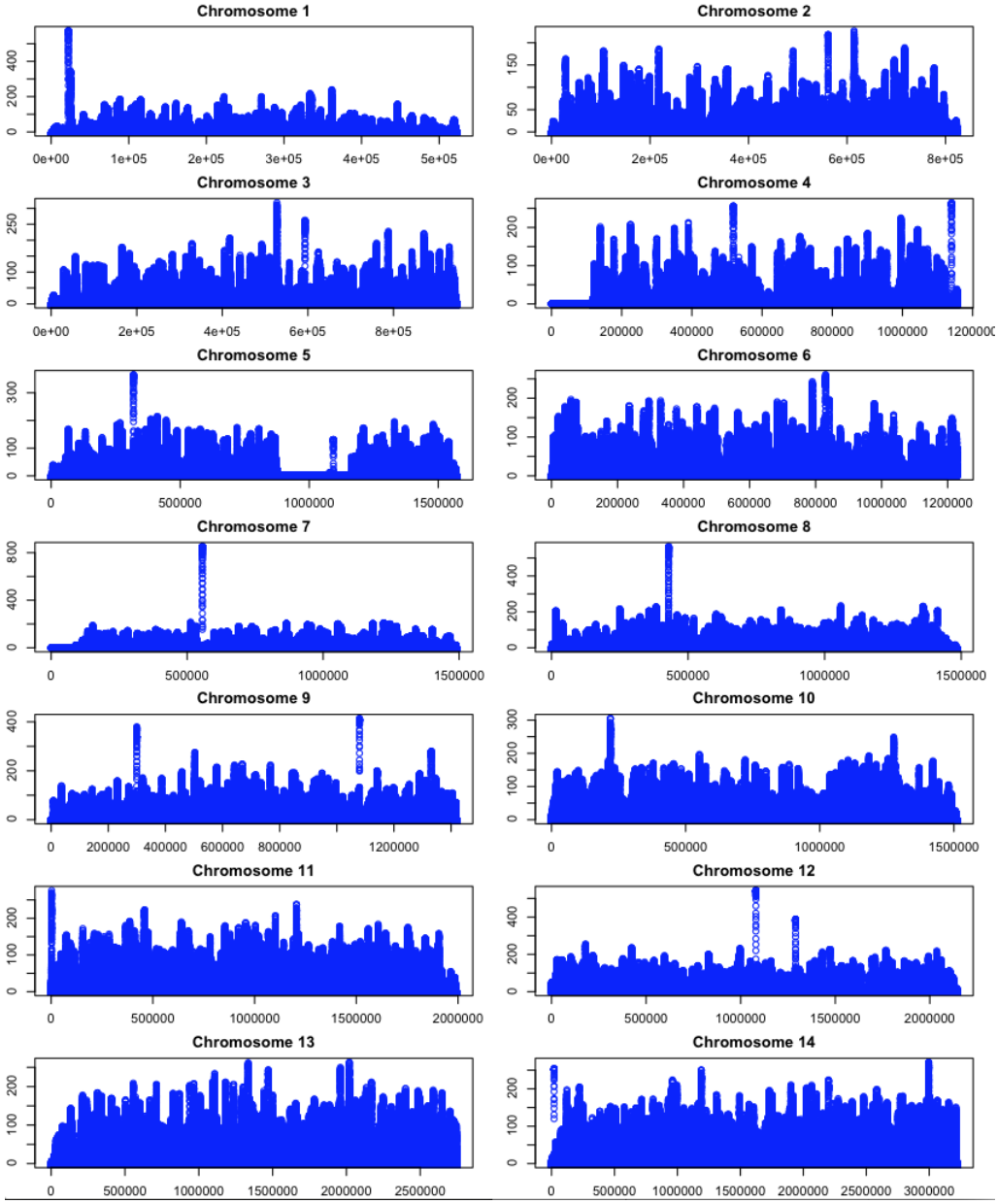

Figure 4.4: Distribution of H3K4me3 in the IT parasite clone. As above, input control reads have been subtracted from the signal track. The H3K4me3 is an activating modification whose patterns are strongly correlated with other activating marks such as H3K9Ac [125]. H3K4me3 was enriched in central areas of the chromosomes and outside of internal *var* clusters, opposite what was observed for H3K9me3 above.

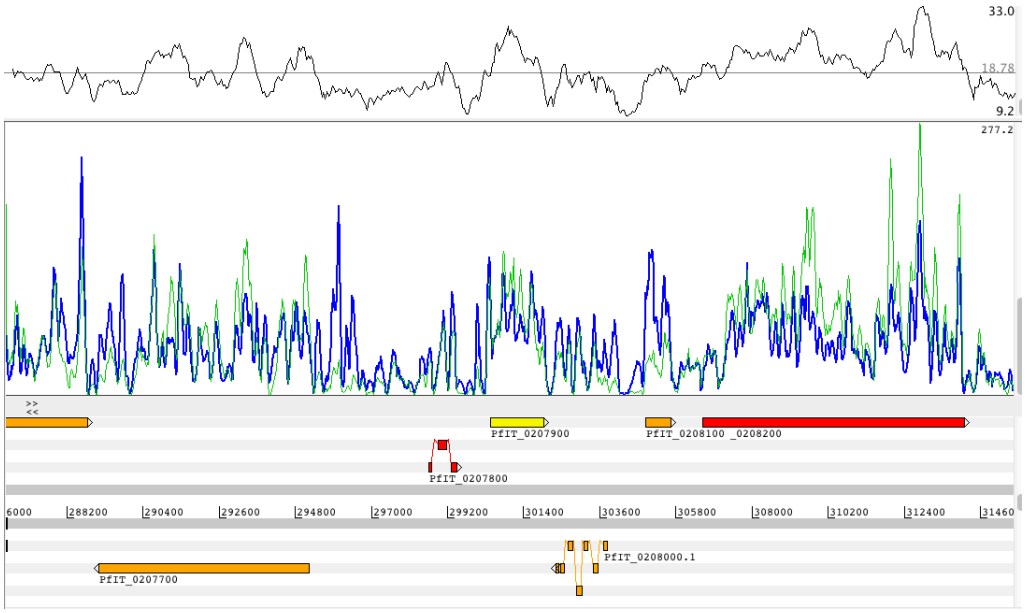

Figure 4.5: Screenshot of coverage for H3K4me3 (thick blue line) and input DNA (thin green line) showing base-pair spatial resolution and the visualization of individual nucleosomes. H3K4me3 is an activating mark which appears to be especially enriched in intergenic regions and at the 5' boundaries of genes. The GC content of the underlying DNA is shown above the plot as a black line, demonstrating detection of nucleosomes both in relatively GC-poor and GC-rich areas of the genome and confirming that weak or no GC-bias is introduced in the amplification of ChIP DNA in the current protocols for library preparation.

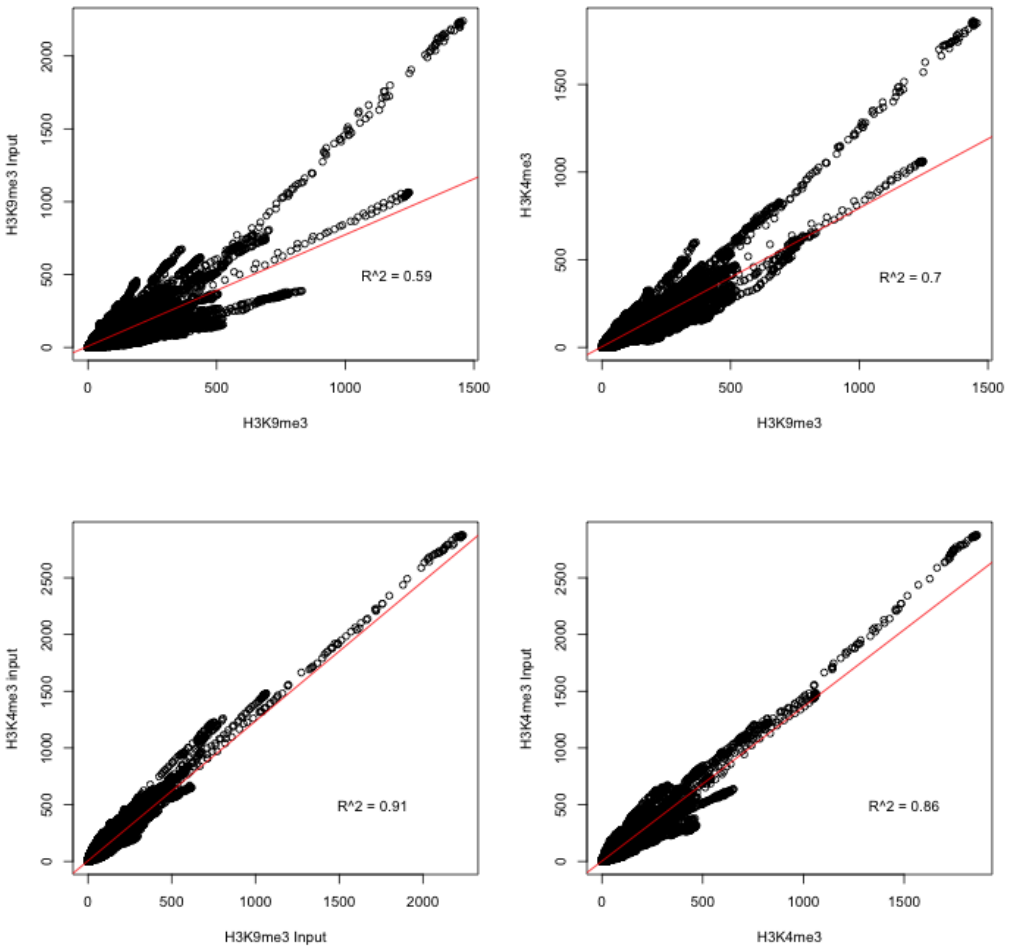

Figure 4.6: The pairwise associations between coverage along chromosome 1 is shown for selected ChIPseq libraries. Each point represents a base on chromosome 1, and the coordinates give its coverage in the libraries labeled on the plot axes. The strong linear relationship for input control libraries ($r = 0.95$) demonstrates the reproducibility of the ChIPseq libraries across biological replicates. The relationships between the different marks and input control provide information about the relative similarity between the underlying nucleosomal pattern and the superimposed epigenetic modifications.

one of the specific *var* genes, BC6, results in depletion of silencing histone marks at that locus. We also demonstrate that the residency of modified nucleosomes is dynamic throughout the lifecycle, at least for the single *var* locus in the 3D7 parasite which was studied at fine temporal resolution. This is a finding which we hope to investigate further when the whole-genome data are available from those samples.

After studying epigenetic changes at a single locus, we go on to study the genome-wide distribution of an activating mark and a silencing mark in the IT clone. We show that the silencing mark, H3K9me3, clusters at telomeres and internal *var* clusters, while activating marks preferentially localize to the internal portions of chromosomes. The spatial resolution of the ChIPseq data lets us identify the position and modification status of individual nucleosomes, offering an unprecedented view into the chromatin and epigenetic landscape of *P. falciparum* chromosomes. The pattern of epigenetic modifications we observe is consistent with that reported by other groups, indicating that the epigenetic portioning of the genome into heterochromatin and euchromatin is strain-transcendent. Nevertheless, much of the meaning of the fine detail of the epigenetic modifications, and the relevance of the patterning of individual nucleosomes, remains to be investigated.

ChIPseq datasets are rich information, and there is much that can still be done with the ChIPseq data presented in this section. In particular, there is an opportunity to integrate ChIPseq datasets, from our experiments as well as those in the literature, with gene expression data. We plan to revisit such questions in future work, but at this point, we chose to follow up on a different aspect of the ChIPseq data: The correlations between epigenetic state across non-adjacent sections of the genome suggest that higher-order spatial relationships are superimposed upon the chromatin, and we study these in depth in the next chapter.

Chapter 5

Spatial Properties of the Folded Genome

5.1 Introduction

The active *var* locus possesses a distinct epigenetic status involving looser chromatin, fewer silencing marks, and a privileged sub-nuclear location. These factors act together to facilitate expression from a chosen locus and to maintain the rest of the *var* alleles in a silenced state. In addition to modulating the density of the chromatin fiber and facilitating or occluding access by polymerase, the epigenetic marks at the active and silent loci confer on the cell a "memory" of the *var* gene expressed in the previous generation. This model accounts for many aspects of silencing, but it does not explain all features of *var* gene expression dynamics. In particular, how is the strict "counting" mechanism – in which a cell only activates a single *var* – enforced? Furthermore, the transition rates between *var* genes are not uniform – what accounts for the differences in "on" and "off" rates between genes? Finally, how do the active and inactive *var* genes arrive at their respective and proper subcellular locations?

We hypothesized that the spatial structure of the genome is highly choreographed and involved in the control of gene expression. In order to test this hypothesis, we set out to study the large-scale folding properties of chromosomes by combining biochemical approaches with second generation DNA sequencing methods. Using chromosome conformation capture techniques [33] and massively-parallel DNA sequencing, we generate high resolution maps (on the order of $5-10$ kilobase resolution) of interaction preferences in the nucleus. These maps overcome the limitations of FISH for mapping sub-nuclear events and provide information about the folding behavior of malaria chromosomes. After generating an initial set of maps, we then go on to study how spatial interactions change when *var* gene expression is altered.

Chromosomes are polymers, and in places we draw on the polymer literature to make sense of the data produced by our experiments. An appendix (Appendix D)

is included which discusses the necessary background.

5.2 Development of a 3C Assay for *P. falciparum*

We adapted the chromosome conformation capture assay of Dekker and colleagues [33] to malaria parasites. The final assay protocol is given in section A.7; we briefly discuss its development here.

In a 3C protocol, chromatin is cross-linked, digested with a restriction enzyme, and then ligated together. These steps are shown in Figure 5.1. In the cross-linking step, a chemical cross-linker (typically formaldehyde) is added with the intention of covalently linking protein-protein and protein-DNA interactions in the cell. The DNA is then digested with restriction enzymes under biochemical conditions that allow enzymatic access to the chromatin (0.3% SDS). The digested DNA is diluted and ligation buffer is added in order to generate intramolecular ligation products.

The relative concentration of ligation products in the resulting library is a measure of the pairwise affinity between two segments of DNA. In the original description by Dekker [33], the relative concentration of the ligation sites was measured by semi-quantitative PCR. This approach is time-consuming and only suitable for questions about particular sets of interactions which are suspected *a priori* to be of interest. A newer approach, sometimes known as Genome Conformation Capture (GCC) [123], involves direct sequencing of the 3C library. A further modification was made by Lieberman-Aiden et al. [86], who inserted a chemically modified nucleotide, containing biotin, in the ligation site. Prior to second-generation sequencing, the authors then enriched the library for genomic fragments that contain ligation sites.

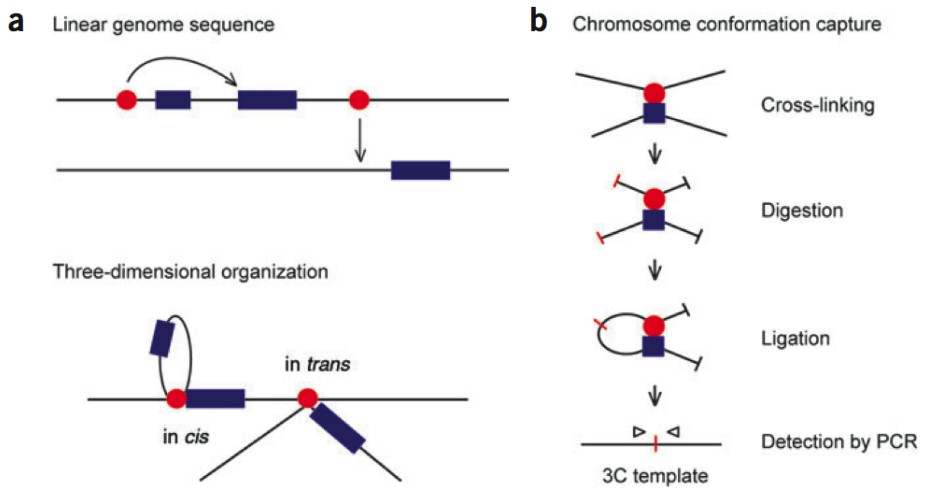

Figure 5.1: This image, taken from reference [32], shows the 3C protocol as described in reference [33]. The left panel shows an enhancer (red dot), separated by physical distance or on separate chromosomes, looping in three-dimensional space to regulate the expression of target genes. The right panel describes the steps of the 3C assay, which include formaldehyde-based cross-linking, restriction endonoclease digestion, and ligation. Fragment interactions which occur during the cell cycle in a living cell are preserved via the chemical cross-linking, and then encoded into a linear DNA molecule through digestion and ligation steps. The resulting library is then interrogated by PCR, microarray hybridization, or sequencing.

Chromosome conformation capture assays are not straightforward to perform [32, 58]. At the level of cross-linking, choices must be made about which cross-linker, how much to use, how long to cross-link, and whether to cross-link cells or nuclei. In the digestion step, a suitable restriction enzyme must be chosen and its concentration optimized. The length and concentrations of the DNA ligation reaction must also be optimized, as well as downstream purification and sequencing steps.

In our assays, we repeatedly found unreliable yields of DNA when cross-linking was done on isolated nuclei (c.f. Section A.7). We therefore chose to perform the cross-linking directly on in tact cells. The cross-linking conditions of 10 minutes at 37 °C in 0.5 - 2% formaldehyde followed by quenching with 125 mM Glycine are widely used in the literature. In our protocol, nuclei are isolated after the cross-linking step. We tested several concentrations of formaldehyde and found that 1% was the most reliable concentration. On a test locus using *Eco*RI for digestion, this concentration of formaldehyde was found to give the expected decay in contact probability as a function of genomic distance. A representative pilot experiment is shown in Figure 5.2. Both 0.5% formaldehyde and 1% formaldehyde yielded the an appropriate decrease in relative interaction frequency as genomic distance increased, but the curve with 0.5% was noiser; the amount of interaction between sites 1 and 3 was apparently larger than that between 1 and 2 in the reactions performed using 0.5% formaldehyde. Based on consistently more reliable performance with the 1% formaldehyde conditions for cross-linking, we used this concentration in all of our subsequent experiments.

For enzymes, we used ones which have all been validated in the literature for use in chromosome conformation capture assays. These enzymes must cut DNA efficiently under harsh conditions (0.3% SDS and 1.6% Triton x-100), be heat-inactivatable, and available affordably for large quantities. The enzymes we used were *Eco*RI [33], *Hind*III [86], and *Mbo*I [58]. We also occassionally used *Dpn*II, an isoschizomer of *Mbo*I, which was recommended to us by Dr. Jim Hughes and

validated for chromosome conformation capture in the group of Professor Doug Higgs. In conditions of 1% formaldehyde, 500 units *Hind*III was found to give reliable digestion > 90% on DNA from 10 million ring-stage parasites. *Mbo*I, a restriction enzyme whose restriction is 4 base pairs, was found to give digestion ≈ 70% of sites. While we would have liked the *Mbo*I digestion efficiency to be higher, this was considered acceptable performance for an enzyme which cuts very frequently.

Following digestion, DNA from each reaction, containing approximately 10 million parasites, was resuspended in a volume of 8mL. Ligation was performed using 10 units of T4 DNA ligase from Invitrogen for a 3C reaction; for HiC reactions using blunt-ended restriction sites, at 40 - 50 Units of ligase was used. After ligation, DNA was purified by phenol-chloroform extraction and precipitated using ethanol.

Further detail for the entire protocol is given in section A.7.

5.3 GCC vs. HiC

The chromosome conformation capture assay fundamentally involves encoding three-dimensional spatial information into one-dimensional linear DNA sequence. The resulting 3C library[1] posses enormous complexity, because for a genome consisting of n restriction sites, there are $\binom{n}{2}$ pairwise ligation products. Over-representation of particular pairs of ligation products in the library signifies a spatial interaction. In order to calculate over- or under-representation of a ligation product, the frequency of each of the $\binom{n}{2}$ ligation products must be estimated, a formidable task, because for *Mbo*I this number is approximately 2 billion. In the original 3C experiments of Dekker and colleagues [33], only a small number

[1]Note that 3C and GCC refer to the same procedure in terms of library construction, but I try to use GCC to refer specifically to a 3C library that has analyzed by second-generation sequencing

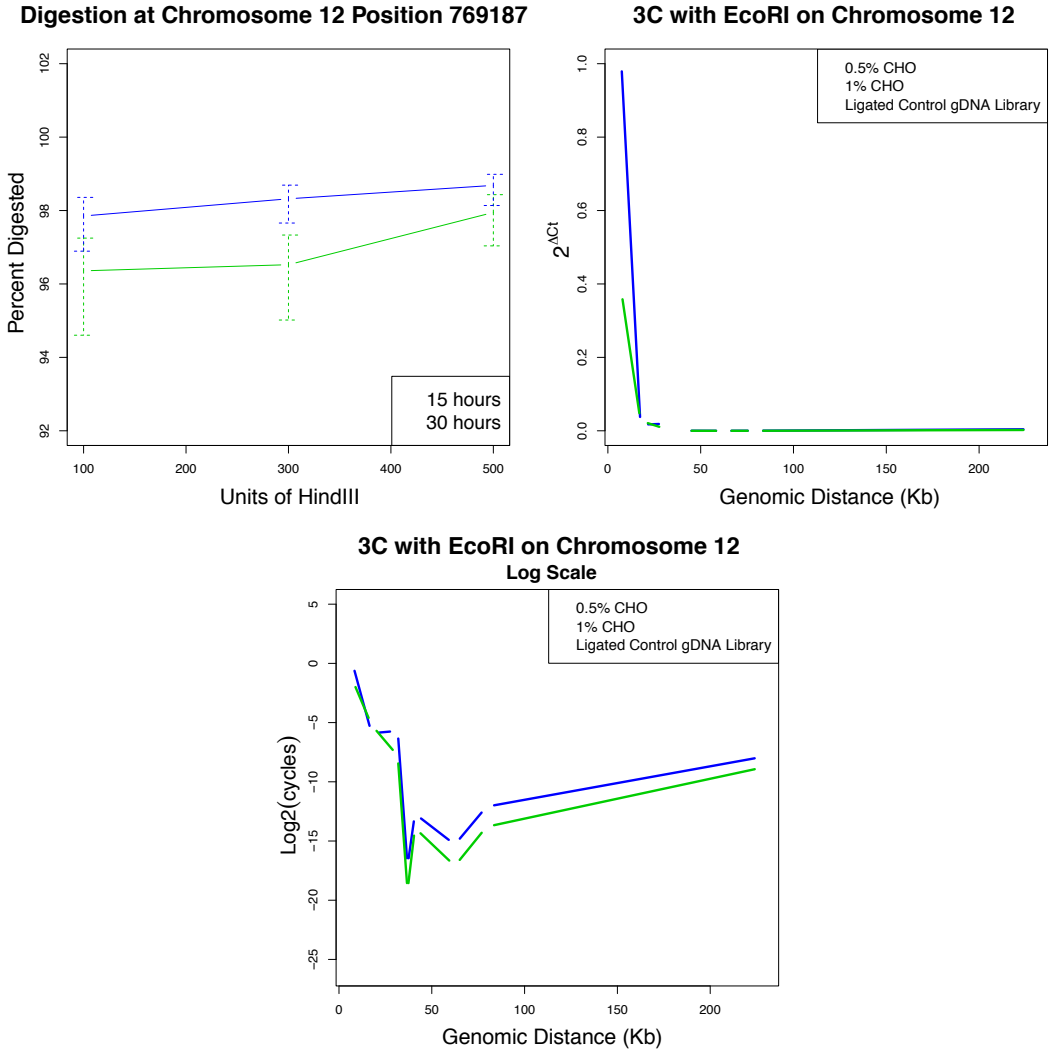

Figure 5.2: **Evaluation of digestion efficiency and testing of the 3C assay.** In the upper left panel, parasites were cross-linked in 1% formaldehyde, nuclei were isolated, and DNA was digested with varying quantities of restriction enzyme for either 15 hours or 30 hours. Digestion with *Hind*III under these conditions is nearly complete though some improvements can be seen with increasing *Hind*III concentration and digestion time. In the middle and right panels, the results of a sample 3C assay using *Eco*RI on linear and log scales, respectively. This is a segment on chromosome 12 for which we were unaware of any looping or spatial interactions. The decline in concentration of the ligated product as a function of increasing genomic distance can clearly be seen. The results for 1% formaldehyde behaved consistently over several assays and we therefore chose this concentration in our assay. Quantitation in the middle plot is under the assumption of a perfectly efficient qPCR reaction and this assay is probably better interpreted as a semi-quantititive result which shows the qualitative features of a successful 3C assay. In the logarithmic plot, cycle number is shown. A ligated, digested control DNA library was also assayed in the reaction and, when detectable, is shown in red.

of sites was chosen; the relative frequency of the ligation products was analyzed by semiquantitative PCR.

Second-generation sequencing offers the power to sequence the 3C library at sufficient depth to estimate the relative abundance of all or almost all pairwise ligation products in the library (depending on the frequency of cuts), which is the approach we have used in this chapter. Nevertheless, only read pairs which map on opposite sites of a ligation site furnish information about the abundance of a ligation product, and therefore direct sequencing of the 3C library has the drawback that only a fraction of the total sequence generated provides useful information. There are several ways to improve the yield of useful information per sequenced read. One strategy, known as "HiC" [86], involves filling in the sticky ends of the restriction fragments before ligation, and including a biotinylated nucleotide in the blunting reaction. After removing the biotin from unligated free ends using the exonuclease activity of T4 polymerase, the library can be enriched using streptavidin affinity purification before sequencing, thereby increasing the percentage of read pairs that span a ligation site.

The small size of the malaria genome means that we can use both GCC and HiC and perform a direct comparison between the two methods. The location of mapped reads along chromosome 1 is shown in Figure 5.3 for both GCC (top panel) and HiC (lower panel). In both techniques, reads cluster around restriction sites (marked with red lines in the figure), which provides evidence that the technique works as it is supposed to. And in general, while the profiles look similar, suggesting good agreement between the techniques, there are more reads spanning restriction sites in the HiC library. Also, the number of reads not mapping to a restriction site is reduced. Therefore it seems that, in accord with what would have been predicted, the effect of HiC is to improve the quality of the interaction libraries. This is consistent with interaction maps of improved resolution in the HiC samples.

The ligation step for 3C and GCC libraries involves a sticky-end ligations, whereas HiC requires a less efficient blunt-end ligation. To overcome this, additional ligase is added; however, the libraries generated using sticky-end ligations typically have reduced electrophoretic mobility in agarose (Figure A.5 and A.6). It is interesting to note that this does not seem to affect the downstream results. Whether the altered gel mobility results from over-ligation and concatemerization of the sticky-end libraries, or simply inefficient ligation of the blunted sites, it appears that the effect is either offset or undetectable in the short read sequences.

5.4 Scaling of Contact Probability

The relationship between linear distance and contact probability is diagnostic of important spatial characteristics of a polymer. The underlying theory is summarized in section D.6.1. From the results assembled there, the most important for this section is that contact probability for a freely diffusing chain in equilibrium can be written

$$y = Cx^{-\frac{3}{2}},$$

where C is a constant, and x and y are scalar variables denoting linear distance along the polymer and contact probability, respectively. This relationship is known as a 'power law' and is often characteristic of scale-free objects such as polymers. The value to which x is raised is known as at the 'scaling exponent' and provides key information about the nature of the power law relationship between x and y. The relationship between polymers and power laws is explained further in Appendix D.

We examined the relationship between contact probability and genomic distance in the HiC libraries. The contact probability data contain a prominent scaling law which holds up until 1 megabase. The observed scaling exponent for

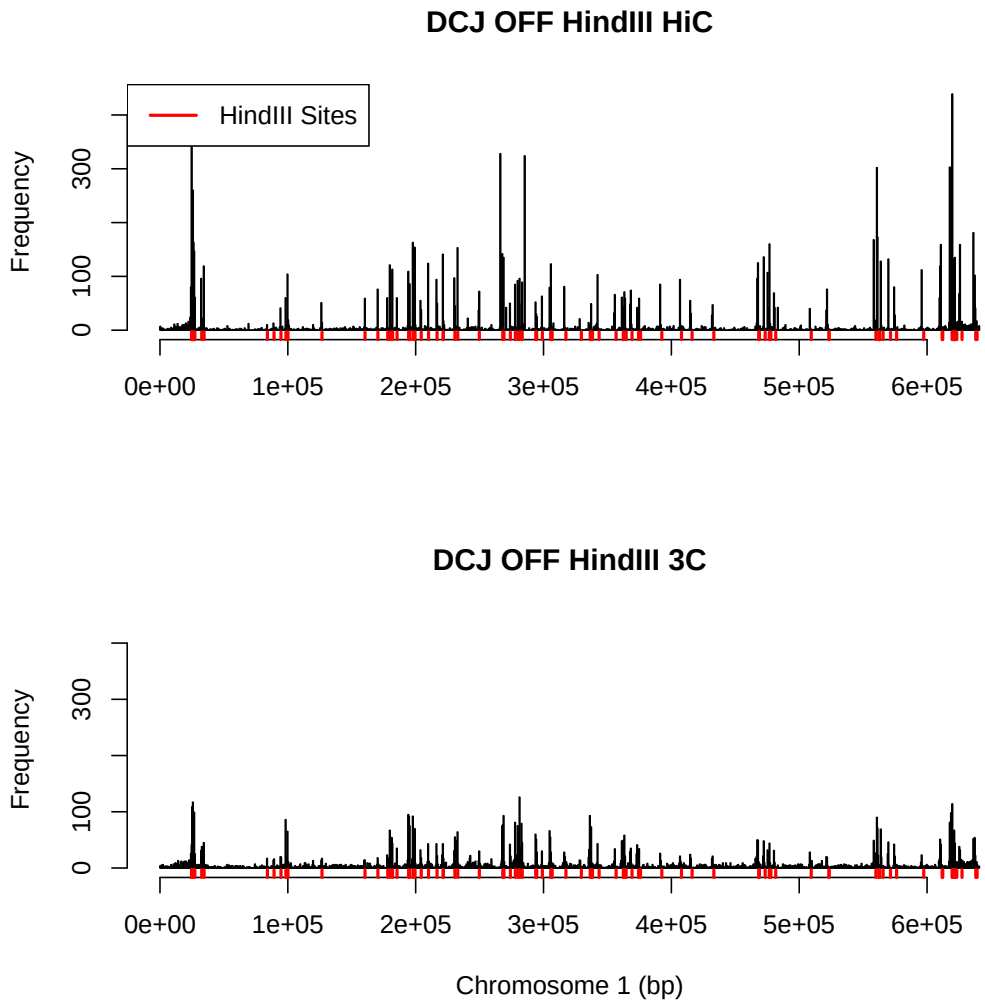

Figure 5.3: Histogram of read mapping sites on chromosome 1 in *Hind*III HiC experiment (top) vs. 3C experiment (bottom) using a sub clone of 3D7 (DCJ off, described in detail later). Restriction sites for *Hind*III are marked with red lines underneath the plot. The fact that the vast majority of split reads map close to a restriction suggests that the technique is capturing digested and then ligated DNA segments. The two assays yield largely equivalent results, demonstrating that the incorporation of biotin into the ligation site, and the use of blunt-end ligation, does not introduce bias into the technique. The HiC experiment has higher coverage adjacent to restriction sites and slightly reduced coverage between restriction sites, suggesting that the data in the HiC experiment is of higher quality with an improved signal to noise ratio.

all chromosomes was ≈ -1; the distribution of values for this exponent is shown in figure 5.6. The $-3/2$ exponent for a random walk polymer in equilibrium is not observed, suggesting that the folding principles of the malaria genome are inconsistent with such a model of polymer folding (termed the"condensed globule" or "equilibrium globule"). The measured value of of -1 is instead more consistent with the "crumpled globule" or "fractal globule" proposed by Grosberg [57] and invoked by Lieberman-Aiden [86] to explain their observed scaling exponents.

It is worth noting that we do not understand the sources of noise in a HiC experiment, and it is therefore possible that the observed scaling exponent is biased up or down due to a technical feature of the experiment. We considered the possibility that the scaling relationship was an artifact of the sequencing process; however, the scaling law was not observed in control libraries. Incomplete digestion of DNA could also generate a scaling relationship between genomic distance and ligation frequency (i.e. contact probability) through circular DNA products derived from linear molecules with n skipped cut sites. Assuming ligation is 100% efficient, the scaling relationship generated by such a process would be of the form

$$y \approx p(1-p)^n,$$

where p is the digestion efficiency. Nevertheless the data do not appear consistent with this possibility for two reasons. The first is that the decay in ligation efficiency is exponential for the above scaling relationship and would not lead to a straight line on a log-log plot in any portion of the curve. The second is that departure from power-law scaling occurs at a different point on each chromosome, roughly one half the chromosome length. Inefficient digestion depends only on local properties and would not be sensitive to the overall length of the chromosome.

The departure of the contact probability data from the power law scaling at

roughly half the distance along each chromosome (the red line in Figures 5.4 and 5.5) is puzzling. At large distances, the decrease in interaction probability is substantially faster than a power law, suggesting that some force is actively working to prevent the contact of distant sections of chromosomes. We speculate that the clusters of heterochromatin first described in [49], and detectable in our inter chromosomal interaction data, also have a biophysical effect on intra-chromosomal interactions by suppressing the likelihood of long-range intra-chromosomal contacts. This effect is reminiscent of constrained diffusive motion and the Brownian bridge (Section D.7).

We conclude that the observed scaling relationships are consistent with the polymeric nature of folded DNA. The scaling exponent of -1 is unexpected and inconsistent with most equilibrium polymer models used for DNA. This exponent was also observed by Lieberman-Aiden [86] and led the authors to proclaim their support for a fractal globule model of genome organization for mammalian genomes on the megabase scale. Our data suggest that the statistical relationships of the fractal globule may extend well below the megabase scale and may be conserved across long stretches of evolutionary time.

5.4.1 Intrachromosomal Contacts

We next considered specific patterns of intra-chromosomal interaction. We divided the results into bins of 10-50kb and placed the split reads in a matrix, A. Figure 5.7 shows the interaction matrix for chromosome 12. The number of reads found between bin i and bin j on chromosome 12 is inserted into position a_{ij} in the matrix. We refer to matrices of this type as pairwise interaction matrices; they are by nature symmetric since we cannot biologically distinguish between segment interactions of the form $i \to j$ and $j \to i$.

The type of plots shown in Figures 5.8 can be recovered from the interaction

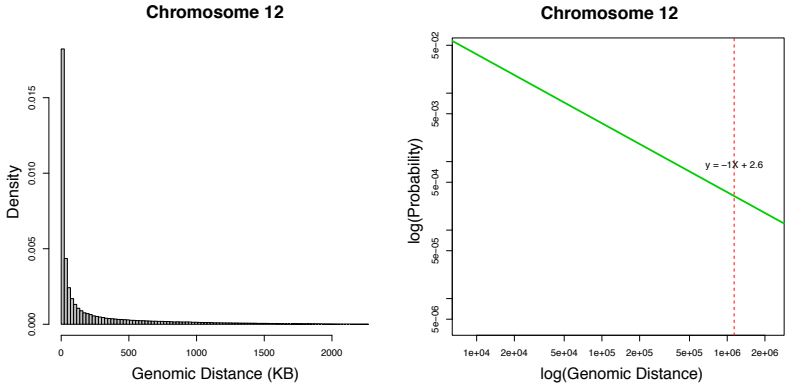

Figure 5.4: Contact Probability with linear (left) and logarithmic (right) axes, chromosome 12. The probability of observing a contact at a given genomic distance can be studied to learn about general folding principles of malaria chromosomes. Polymer theory (see Appendix D) suggests that for a freely diffusing polymer, contact probability should scale via a power law, which is identifiable as a straight line on a doubly logarithmic plot. Examination of the scaling of contact probability for malaria chromosomes reveals a prominent power law whose form holds until approximately half the chromosome distance (marked by the red dashed line). At greater distances, a sharper decline in contact probability is observed, which may indicate that anchoring forces such as heterochromatin forces reduce the likelihood of long-distance contacts on the chromatin fibre.

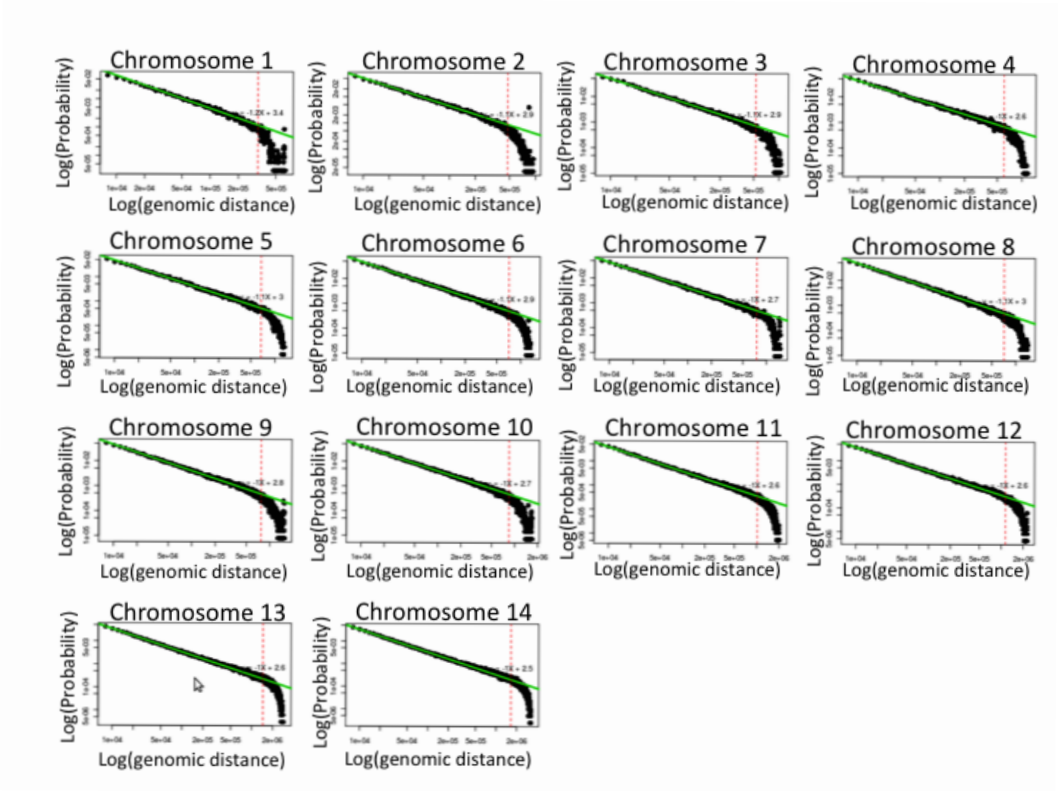

Figure 5.5: Doubly logarithmic plots of contact probability for all 14 *P. falciparum* chromosomes. For each chromosome, the logarithm of genomic distance is shown on the x-axis and the logarithm of the probability of making contact at a given genomic distance is shown on the y-axis. The prominent power law up to approximately half the chromosome length holds for all chromosomes, consistent with polymer theory (see Chapter D).

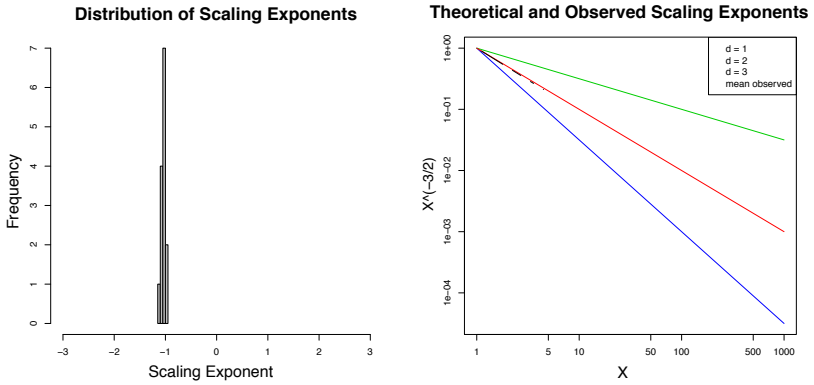

Figure 5.6: The distribution of scaling exponents is presented in the left panel. The slope of the lines in the above plot (Figure 5.5 has been quantified and the distribution of slope values is plotted. The slope value corresponds to the scaling exponent in the power law. The distribution is tightly centered around -1, suggesting that this may be a universal parameter characterizing the folding of malaria chromosomes. The plot at right shows log-log plots that would be obtained under theoretical models of chromosome folding. The plot shows the exact probability of making a contact with itself in a d-dimensional space under a random-walk model. The model is explained and derived in Section D.6.1. The exponent $-3/2$ is expected for a random-walk polymer in a 3-dimensional space, while a constrained folding space will generate larger (i.e. less negative) values. While the random-walk model may not be a good model for chromosomes, this is a general result, which is that the folding of chromosomes occurs in a way which increases the probability of contact between loci.

matrix by simply selecting a row or a column. Figure 5.8 shows interaction data for two loci, one on chromosome 12 suggesting no detectable spatial interactions, and another on chromosome 5 suggestive of a loop.

Three dominant patterns are visible from the intra-chromosomal interactions of chromosome 12. The first is that the vast majority of contacts occur on or close to the diagonal. This is a result of the polymeric nature of chromosome folding, in particular the constrained set of confirmations which are available to a molecule with a connected backbone.

The intra-chromosomal interactions also seem to contain block-like portions of locally interacting domains. This suggests that 100 - 200 kb blocks may form globules which are locally interacting compartments in which the probability of within-compartment interaction is higher than the probability of between-compartment interaction.

The third observable pattern in the intra-chromosomal interaction matrices is the presence of occasional, isolated interactions which are specific to a single bin. Such interactions appear to be rare in the malaria genome but nevertheless do appear to occur. Without technical replicates it is difficult to say statistically which of these is a true biological feature of the dataset and which is generated by technical noise. Such interactions are candidates to be tested in FISH.

The intra-chromosomal contacts for chromosomes 1-8 are shown in figure 5.9. The intra-chromosomal contact matrices for all chromosomes show features broadly consistent with the pattern observed on chromosome 12. This includes the dominance of the main diagonal, the presence of locally interacting domains, and specific interactions between individual loci.

There are several features of these maps which are poorly understood and present opportunities for future research. The structure of chromosome "domains" is not the same across chromosomes. For example, chromosome 12 shows prominent block patterns, discussed above. These blocks occur both as rectangular

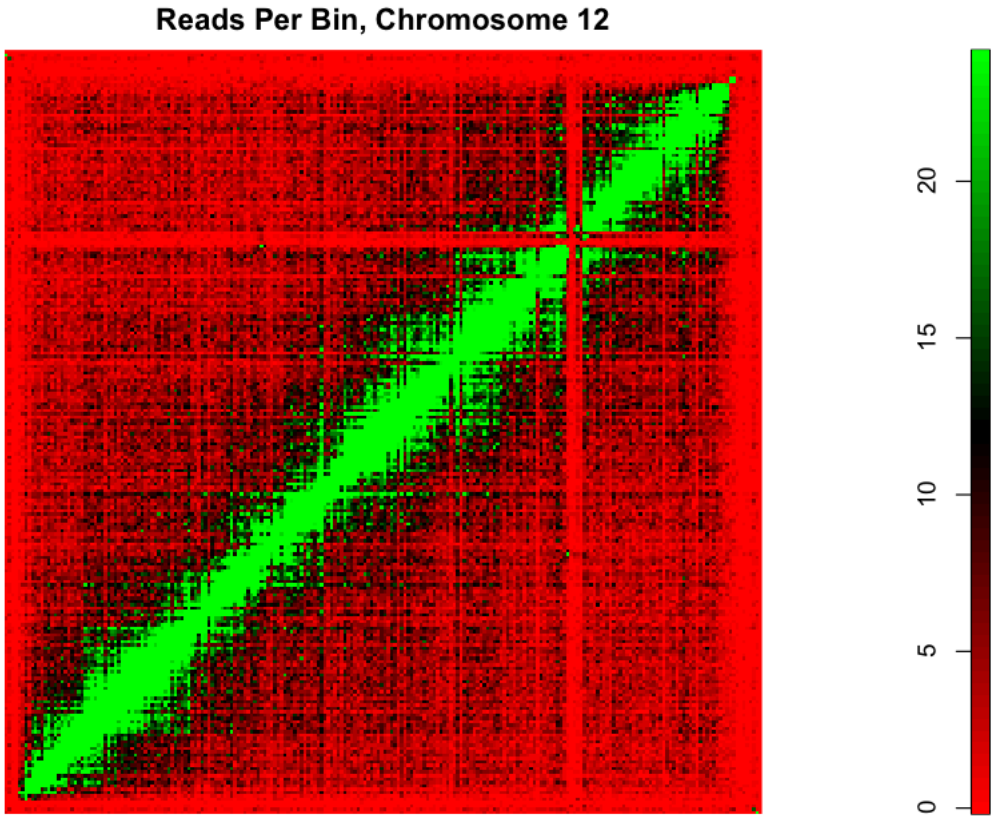

Figure 5.7: Heatmap of intra-chromosomal spatial contacts, chromosome 12. In this figure, the number of contacts observed between locus i and locus j, at a resolution of 10KB, is plotted as a color values (according to the scale given in the legend at right) in the matrix element a_{ij}. A threshold of 24 reads on the intensity value was set in order to maximize contrast in the relevant portions of the image. The largely green diagonal indicates that loci which are close to each other in linear genomic distance are more likely to contact each other in physical space, a result which is consistent with polymer theory. Broad interacting domains as well as isolated, locus-specific interactions can be seen in this analysis. These maps were generated using the IT clone, and the broad red band of limited contacts represents a gap in the current genome assembly.

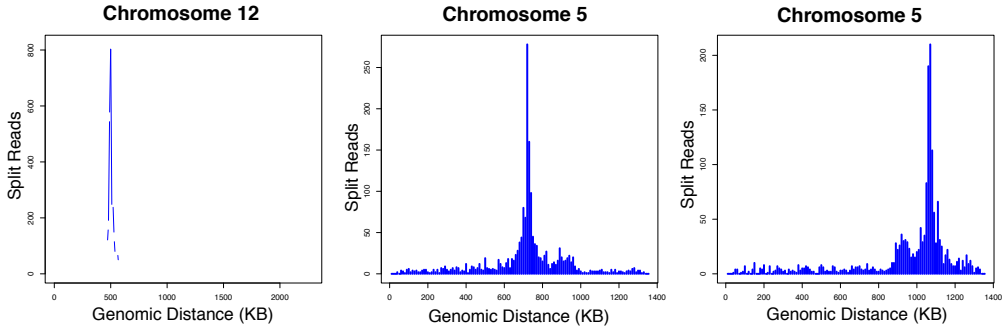

Figure 5.8: The interaction data from a single locus can also be displayed by placing linear distance on the x-axis and split read count on the y-axis. In the left panel, HiC data from a single locus approximately 500 KB into chromosome 12 are displayed. The data from multiple restriction sites have been pooled to generate the 10KB resolution, which reduces noise by smoothing the data. The prominent decrease in contact as a function of genomic distance can be seen. At this locus, no major spatial interactions appear to be occurring above background. In the middle and right panels, the interaction data from two loci on chromosome 5 are displayed. A second peak can be seen in the split read data, suggestive of a loop. This corresponds to the MDR1 locus; a second peak in the data could arise from a genuine loop or a copy number variation in the MDR1 region. We investigated both possibilities further by generating interactions maps in other parasite clones and imaging studies, and the results suggest that this increase in interaction frequency results from an clone-specific expansion of the IT locus.

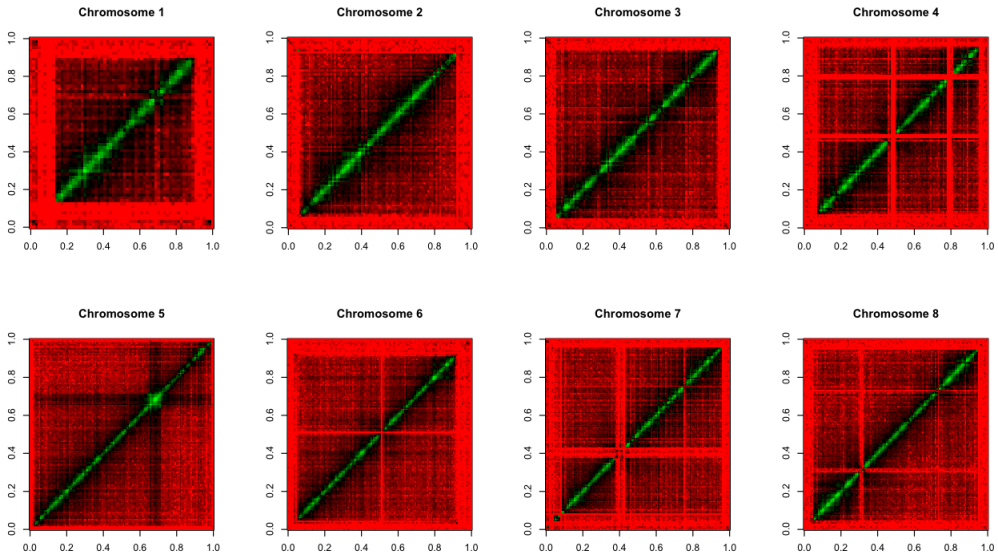

Figure 5.9: Intra-chromosomal spatial contacts, chromosomes 1 - 8. Similar to Figure 5.7, this figure shows the heatmaps of chromosomal interactions for the first 8 chromosomes. As above, locally interacting blocks, locus-specific interactions, as well as areas in which interactions are difficult to call at present due to the evolving IT assembly can be seen.

blocks which suggests that all vs. all interactions are favored within a block; however, there are also "arrow" type interactions indicative of a one vs. all interaction pattern, or a repeated loop structure, within a block. It is not yet clear what these correspond to. Example of an "X" shaped pattern, suggestive of a region of a chromosome apt to form specific loops, can also be found on chromosomes 2 and chromosome 12.

Chromosome 5 shows a prominent, block pattern residing on top of the MDR1 locus. This region caught our attention because of the known association, identified by Gonzales and colleagues [56], of this region with inheritance of a large number of differentially expressed transcripts or "expression QTLs". The role of this region in intra-chromosomal interactions remains unclear, but several pieces of evidence suggest that it may be an artifact.

We imaged this putative interaction using 2-color FISH. Representative images are shown in Figure 5.10. The majority of the time, the red and the green probes do not co-localize, suggesting that there is no interaction most of the time. Occasionally, the locus is seen to co-localize, but this occurs $\approx 20\%$ of the time and is attributed to background collision frequency.

A second piece of evidence comes from considering the inter-chromosomal interaction data (Section 5.4.2). The MDR1 locus contains a transporter linked to drug resistance and is known to readily amplify, with some parasite clones carrying multiple copies [23,113,114]. The genome-wide interaction maps in figure 5.12 show a prominent band of increased interaction with the MDR1 locus. The interaction intensity is uniformly distributed across nearly the entire genome. It seems physically unlikely that the locus interacts with the entire genome at a single timepoint during development, suggesting that the observed interaction may be an artifact generated by increased copy number.

Further evidence for this hypothesis is given by looking at the the total number of inter-chromosomal interactions, binned across the 14 chromosomes, with

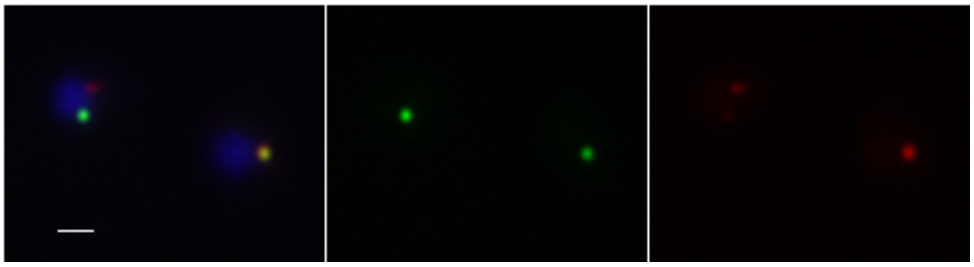

Figure 5.10: Fluorescence In-Situ Hybridization was performed against the MDR1 locus and a gene approximately 100 kilobases upstream, PFE_1050c. As described in Methods (Section A.8), a 2KB probe was generated by random synthesis using the method of Feinberg and Vogelstein [41]. This probe was then hybridized to the denatured chromosomes of individual cells fixed onto a microscope slide. The MDR1 locus was detected with a fluorescein-labeled probe and appears as green, while the PFE_1050c locus was hybridized with a biotin-labeled probe, detected with an AlexaFluor 568-conjugated streptavidin, and appears as red. DNA is stained with DAPI and appears as blue. Two ring-stage parasites are shown. Scale bar is 1 μM. The left-most panel shows the green, red, and blue channels merged, whereas the middle panel and right panel show the green and red channels, respectively, in isolation. The locus is occasionally seen to co-localize (as shown in the nucleus at right), but not at a frequency greater than background, and the vast majority of signals did not overlay (as in the nucleus at left).

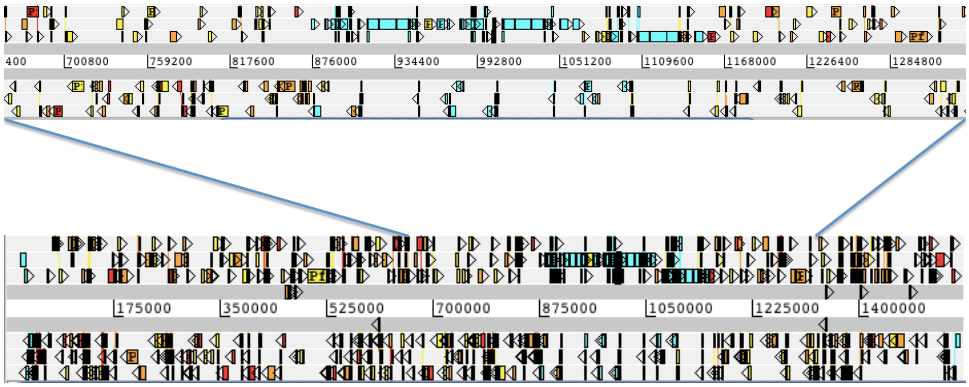

Figure 5.11: **The IT parasite possesses three copies of the MDR1 locus.** The reannotated chromosome 5 of the IT clone, showing whole chromosome view (bottom) and zoomed view (top). An amplification event has occurred, resulting in three copies of the locus in the IT clone. This generated an artificially elevated interaction signal from the locus when the short reads were mapped to a reference which included only a single copy of the locus. Interestingly, the intra-chromosomal interaction signal was elevated greater than 3X, suggesting that 3C followed by sequencing may be a sensitive way to detect amplification events.

noise filter (i.e. minimum number of reads mapping). Background signal, or noise, should feature prominently in these types of plots, and a locus that is over-represented in the genome should be overrepresented in the noise. These data, shown in Figure 5.16, shows that the MDR1 has roughly three times the number of inter-chromosomal interactions as the rest of the genome when no minimum threshold is applied. When a minimum of 4 split reads is applied (Figure 5.15), this increased signal is removed, suggesting that it is due to noise.

Based on the results from FISH along with the inter-chromosomal interaction data, we re-examined the assembly of the IT clone. Optical mapping on chromosomes from the IT clone was performed at the Sanger, and this revealed that the parasite possessed three copies of the locus. This demonstrated that chromosome conformation capture data can be used to detect structural rearrangements. The improved annotation for chromosome 5 is shown in Figure 5.11. It is interesting to speculate that simply incorporating a digestion and ligation step prior to sequencing may be useful for improving genome annotation or detecting structural genomic variations, for example in cancer genomes.

A final prominent feature of the intra-chromosomal interactions is that there are large gaps in several areas of the genome. These gaps are present at the telomeres and sub-telomeres, as well as in the internal clusters of gene families including *var* genes and *rifins*. The coverage within these regions is non-uniform, and only small numbers of reads map overall. The apparent paucity of interactions in these regions is due to the quality of the reference sequence in these areas. Second-generation sequencing methods, such as the Illumina/Solexa platform we used, achieve their throughput by generating large numbers of short reads in extremely large numbers of sequencing reactions run in parallel. These short oligomeric sequences must be aligned to the reference genome. In these initial studies we used the parasite clone A4, which we chose for its surface expression of PfEMP1 proteins and the ability to select it for expression of the A4var gene using

the BC6 monoclonal antibody [137]. Nevertheless, because the reference sequence is imperfect for this parasite clone, some information cannot be obtained from these libraries until the reference is improved. We have been working hard with the Pathogen Sequencing Unit to refine the IT assembly, and we expect to be able to fill in many of these gaps in the near future.

5.4.2 Inter-chromosomal Interactions

We next considered contacts that occur between chromosomes. Inter-chromosomal interactions are potentially both more interesting and more difficult to handle. Unlike intra-chromosomal interactions, in which the physical constraints enforced by the polymer backbone specify a quantitative structure for the expected distribution of interactions (discussed at length in Appendix D), we know much less *a priori* about the expected stucture of inter-chromosomal interactions.

Figure 5.12 shows the entire set of interactions, at a resolution of 50kb, for all fourteen chromosomes comprising the genome as a whole. The ordering of chromosomes is by chromosome number, i.e. chromosome 1 is first and chromosome 14 is last. This ordering is roughly by size but is arbitrary in terms of the interactions between chromosomes. The whole genome interaction map is broadly consistent with features observed so far in the data: Inter-chromosomal interactions are prominent, with elevated numbers of interactions occurring within chromosomes, and these interactions decreasing as a function of increasing genomic distance. The scaling

$$p(r) \approx Cr^{-1},$$

discussed in section 5.4, within chromosomes, is the defining feature of the maps when in the genome is viewed large.

Contacts away from the diagonal are also visible, although they are difficult to resolve at this resolution. Many contacts are visible between telomeres and

internal *var* clusters. Such contacts have been widely reported in the literature using FISH [49, 50, 90]. Most of the interactions cannot be understood from figure 5.12, for two reasons. The first is that the resolution of the matrix, even at 50KB, exceeds what the digital image is capable of showing. Many pixels are over-plotted (i.e. dots of different color are repeatedly plotted on top of the same pixel), making it difficult to separate interactions from background. The second difficulty is that a matrix representation of the contact data is not intuitive; in particular, it does not communicate well the network structure of the interactions (e.g. of type $i \to j$, $j \to k$, $i \to k$ and higher order clusters of interactions are not readily identified by inspection of the matrix).

One way to address the issue of overplotting is to filter the data by removing bins with fewer than k interactions and to represent the number of interactions by some function of point size (in this case point size = $\log(k)$). This is shown in Figure 5.13 for $k = 6$. Like Figure 5.12, this plot emphasizes the importance of diagonal or polymeric interactions. One thing that Figure 5.13 emphasizes, which Figure 5.12 does not, is the importance of interactions at telomeres and internal *var* clusters. These appear to be the major structural features of the genome at large.

Condensing the interactions across a chromosome into a single bin per chromosome yields the representation shown in Figure 5.14, which shows pairwise interactions between chromosomes. This symmetric matrix shows inter-chromosomal preferences or affinities. The most prominent interactions are between chromosomes 10 and 2, and chromosomes 9 and 3. Some chromosomes appear to interact more than others, for example chromosomes 1, 7, 9, 10, and 12 have large numbers of interactions with multiple chromosomes, whereas chromosomes 4, 11,13, and 14 appear to interact less.

Moving away from the matrix representations, we can consider where along the chromosomes the inter-chromosomal interactions tend to occur and begin to

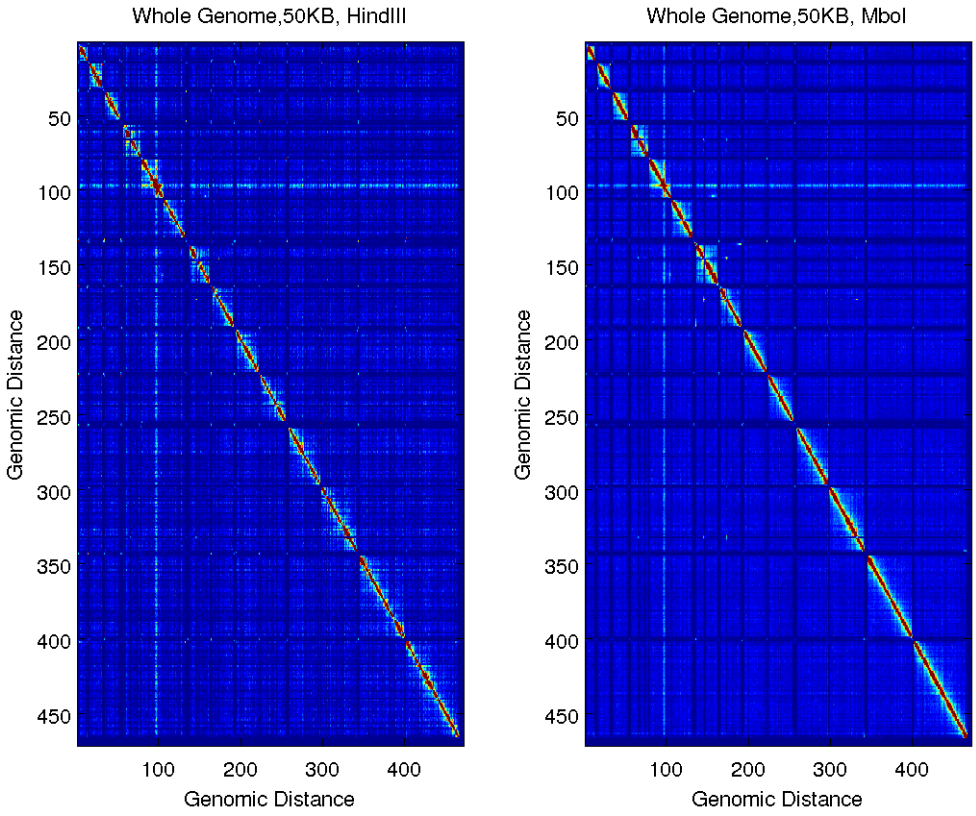

Figure 5.12: **Genomic view of interaction libraries produced with different enzymes.** Genome-wide interaction matrix at 50 kb resolution, generated using both *Hind*III (left panel) and *Mbo*I (right panel) HiC. Interactions between all pairwise sites on all 14 chromosomes are shown. As in Figure 5.7, the color matrix element a_{ij} represents the number of reads between genomic segment i and genomic segment j, except in this plot all fourteen chromosomes have been concatenated and the bin size has been increased to 50 kilobases to facilitate the comparison between the two enzymes (*Hind*III produces maps of coarser resolution). Comparison of the two plots shows that maps produced using different enzymes are reproducible. Interactions between telomeres and among rDNA genes can be seen, although detail is obscured by the high density of data displayed in this plot.

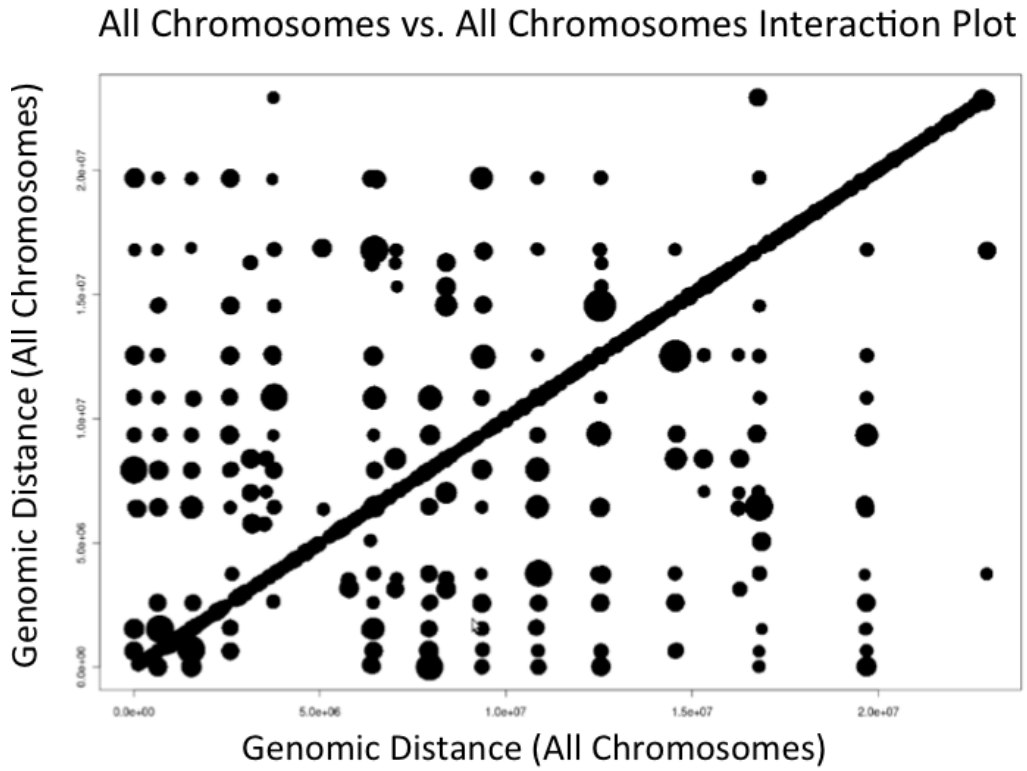

Figure 5.13: Another representation of the interaction matrix between of all 14 chromosomes. In this presentation, the dataset has been filtered into bins of 300 bases, and the number of reads in each bin is shown as the logarithm of point size. As in Figure 5.12, the chromosomes have been concatenated and both the $x-$ and $y-$ axes run from 0 to 23 megabases.Once again, the prominent diagonal band down the plot demonstrates the likelihood of loci nearby in linear, genomic space to be proximal in real space. One advantage of this display is that overplotting is reduced relative to Figure 5.12, and the interaction of individual loci can be seen. Clear preferences among chromosomes and telomeres can be detected.

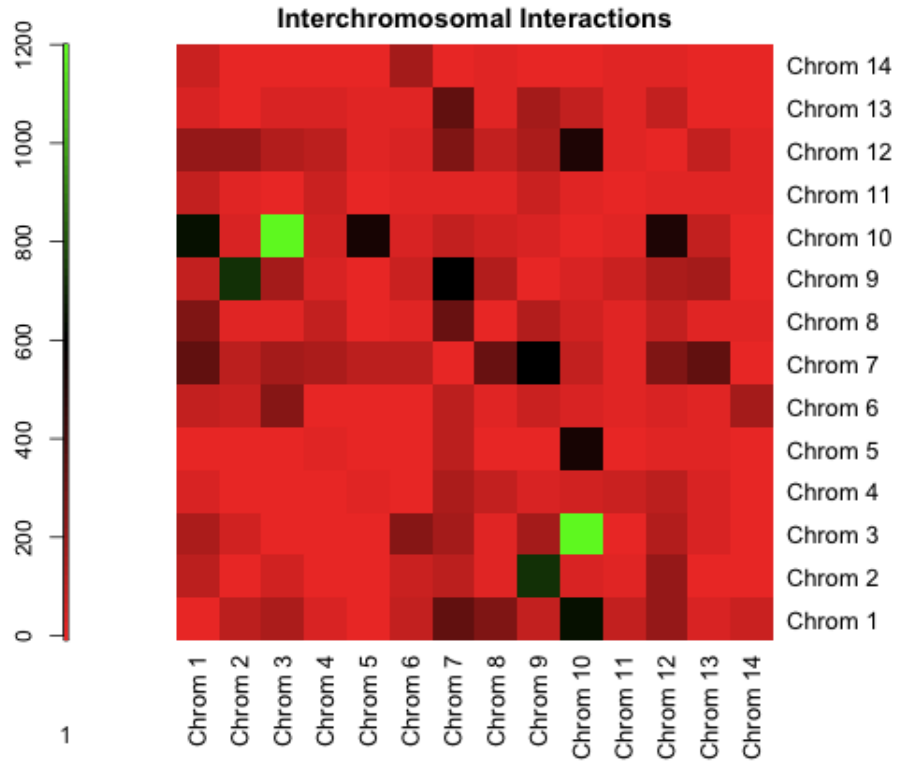

Figure 5.14: **Inter-chromosomal interaction preferences.** Chromosomes have non-random affinities. Raw numbers of pairwise interactions between chromosomes, showing that chromosomes preferentially interact with one another. The number of split reads indicating a pairwise association between chromosomes is plotted in the matrix as a color, with the legend shown at left. This shows the relative pairwise affinity between chromosomes. For example, chromosome 10 interacts strongly with chromosome 3, whereas chromosomes 14 and 11 appear relatively isolated.

compare the results with known structural features of chromosomes. Figure 5.15 shows that the most prominent inter-chromosomal interactions are between telomeres and internal *var* clusters. As noted above, this agrees with FISH data showing that telomeres cluster together at the nuclear periphery [49, 90].

Spatially, the inter-chromosomal interactions overlap almost perfectly with known sites of heterochromatin. We have marked with red lines in Figure 5.15 the results of a genome-wide chromatin immunoprecipitation ("ChIP-on-chip") study using an antibody that targets tri-methylated Lysine 9 on the core histone protein H3. This is known to be a silencing mark in *P. falciparum* [18, 89] and to colocalize with heterochromatin protein 1 (HP1), a known organizer of heterochromatin [116]. This suggests that areas of heterochromatin dominate the total set of interactions; in conjunction with the data obtained using FISH [49, 50], these results indicate that the heterochromatin hubs are the major structural organizing force of the *P. falciparum* genome.

We further investigated the relationship between histone modifications and chromosomal interactions by assembling publicly available chromatin immunoprecipitation datasets. Integrating 2-dimensional data such as ChIP (datasets of the form (x, y) where x is a position and y is a quantity of immunoprecipitated DNA) with 3-dimensional data such as HiC data (datasets of the form (x, y, z) where (x, y) is a pairwise interaction between chromosomes and z is the intensity of the interaction) is a significant informatics challenge. Circos [71] is a software tool that facilitates representation datasets with different dimensionality. The 1-dimensional coordinates are arrayed around a circle, and a second dimension of data can be plotted by height above or below the coordinate system. Interactions are then plotted as lines drawn between points on the circle, which in this case are drawn as smooth bezier curves for aesthetic purposes.

Figure 5.17 shows inter-chromosomal action data from HiC data combined with chromatin immunoprecipitation followed by sequencing ("ChIP-seq") data from

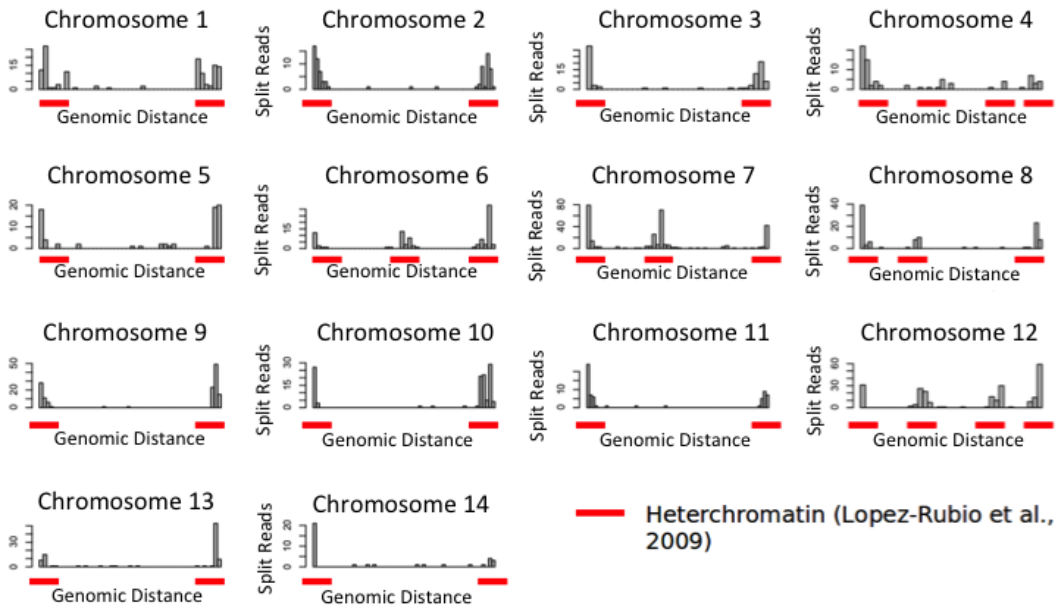

Figure 5.15: **Total interactions from individual chromosomal loci.** Sites of inter-chromosomal interaction overlap almost completely with known heterochromatin foci. In this plot, the total number of spatial contacts within a bin are displayed, i.e. for each bin, b_i, the height of the bin, h_i is plotted according to the column sums of the interaction matrix A: $h_i = \sum_j a_{ij}$. To reduce noise, sub-bins of 300 bases pairs were first generated and only sub-bins with greater than 3 reads have been considered in this analysis. This provides a measure of the total inter-chromosomal interactions stemming from each region of each chromosome. Since these are the marginal tallies of the interaction matrix, we sometimes refer to the h_i as the marginal interactions at locus b_i. The chromosomal regions which interact most are the sub-telomeric regions and the internal *var* clusters. These are the same regions identified by Lopez-Rubio and colleagues as heterochromatin containing H3K9me3 marks [90] (identified by red bars in the figure).

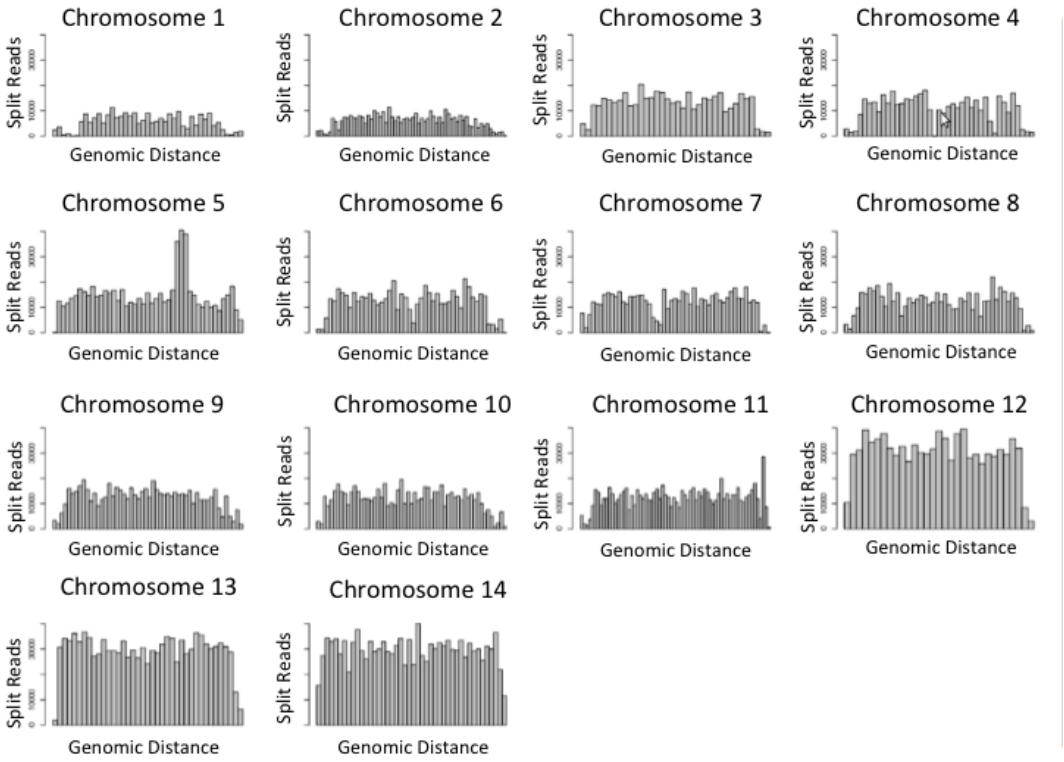

Figure 5.16: **Total interactions at a given chromosomal locus, with a less stringent noise filter.** This plot shows the same marginal tallies as in Figure 5.15 but with a threshold of 1 split read. This has the effect of including many 'random' interactions which amount to statistical noise. With the noise intentionally included, the increased interaction frequency of the MDR1 locus on chromosome 5 can be seen, indicative of an amplification event.

Bartfai and colleagues [15], who measured H3K4me3 (shown in grey), H3K9Ac (shown in blue), and H2A.Z (shown in green). ChIP input DNA from their study is also plotted in red. The interactions we observed in the HiC data at a treshold of k reads, where $k = 6$, are plotted as curved blue lines across the interior of the circle.

The inter-chromosomal interaction data for chromosomes 4, 7, and 8 are shown in Figure 5.17. Data are available for all fourteen chromosomes, but these were chosen because they are representative, and because they contain both subtelomeric *var* clusters and internal *var* clusters. The genome-wide inter-chromosomal maps for a given chromosome are shown in the left panel; a expanded view of the interaction network for chromosomes 4,5,7 and 8 is shown in the right panel.

The data again show that the subtelomeres and internal *var* clusters dominate the inter-chromosomal interaction datasets. The hubs of interaction are nearly all emanating from either internal clusters or subtelomeres. A surprising feature of the datasets is that subtelomeric clusters tend to interact with subtelomeric clusters, while internal *var* clusters tend to interact with other internal *var* clusters. This is contrary to the interactions observed in [90], and we are not certain why the HiC data in this case appear to disagree with the FISH data.

In all cases, there is a strong correlation between the presence of chromatin immunoprecipitation marks and inter-chromosomal interactions. Sites of high levels of spatial interaction are low in acetylated lysine 9 of histone H3 and low in histone H2A variant Z (H2A.Z), while they are high in tri-methyl lysine 9 of histone H3 (see also Figure 4.3 and 5.15). Overall, marks indicative of highly methylated histones and tightly coiled DNA present in the heterochromatin compartment suggest that substantial interaction is present, while acetylated marks characteristic of loose chromatin and the euchromatin compartment are associated with reduced spatial interaction.

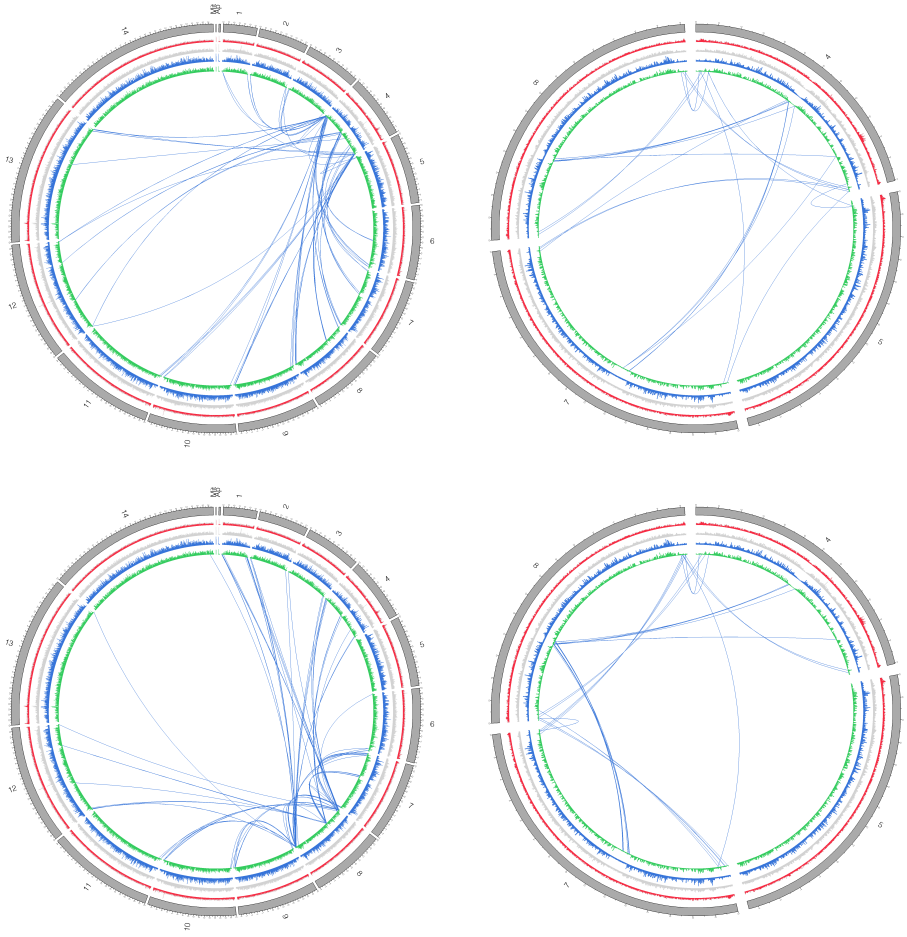

Figure 5.17: The inter-chromosomal interactions on chromosome 4 (top) and chromosome 8 (bottom). All chromosomes are shown on the left panel; only chromosomes 4 - 8 are shown on the right. The analysis presented in Figure 5.15 demonstrates that areas of increased inter-chromosomal interaction tend to overlap with known sites of heterochromatin, but the specific nature of these interactions was not displayed on that plot. Circos diagrams [71] provide an elegant visualization tool with which interaction data as well as chromatin IP data can be simultaneously displayed. The chromosomes are plotted in gray, wrapped end-to-end around the circumference of a circle. Interactions between chromosomes are displayed as smooth curves running through the center of the circle to the corresponding points on the periphery. Interaction data have been filtered by sub-bin as in Figure 5.15. Chromatin IP data, obtained from reference [15], are displayed as colored plots just interior to the chromosomes. ChIP input DNA is shown in red; H3K4me3 is in light gray; H3K9Ac is in blue, and H2A.Z is in green. Inter-chromosomal interactions occur predominantly between regions of the chromosome that are low in activating marks such as H3K9Ac and H2A.Z (and high in silencing marks such as H3K9me3 – not shown in this plot). The internal *var* clusters tend to interact preferentially with one another. This finding is to some degree inconsistent with the FISH data of Lopez-Rubio and coworkers [90], although this may be an artifact of poor resolution due to assembly issues in the IT clone.

5.5 Reconfigurations During *var* Switching

The original impetus for developing 3C and HiC methods was to use these methods as a tool for study gene expression, in particular the spatial relationships of higher-order chromatin structures and *var* gene expression. Are some *var* genes closer to each other in the nucleus than others? Linear genomic positioning is known to affect the switch rate and switch preferences (sometimes called switch bias) among *var* genes [47, 119], but the mechanisms driving this effect are not understood. Does spatial proximity correlate in any way with *var* gene switch rate or switch bias?

These are not straightforward questions to study because of the highly polymorphic nature of *var* genes and other gene families in their associated chromosomal regions. This level of polymorphism makes it difficult to assemble these regions in parasite genome projects. Therefore, while the complete nucleotide sequence of the majority of the IT strain is known, including many of the differences between this parasite and the 3D7 reference clone, the subtelomeres and internal *var* genes are not assembled for the IT parasite genome. Furthermore, the high rate of recombination between these regions, and the conservation and constant exchange of block sequences within *var* genes [14,43,48] makes any interactions in a non-base perfect reference genome suspect. On the other hand, *var* expression is weak in 3D7, and the parasite is difficult to select for expression of individual *var* genes. Overall, studying switching in spatial interactions that accompanies switching in *var* gene expression is a formidable challenge that requires highly specialized experiments. In general, there is a tradeoff between using lines that are selectable for *var* expression, and lines that are isogenic to the reference.

In order to address this issue, we took two approaches. The first is to study spatial reconfiguration in the A4 clone with the BC6 antibody and to try to overcome the mapping problem by refining the reference sequence. The second is

to use transgenic parasites, generated from the reference clone, in which *var* gene expression can be modulated. The data from the BC6 selected vs. unselected experiment has not yet cleared the sequencing pipeline; however, we do have data from the initial A4 clone as well as the switching experiment in transgenic parasites from the 3D7 background, which we present below.

The whole-genome interaction maps derive from an A4 culture which expresses $\approx 60\%$ A4 var gene (a subtelomeric *var* gene located on the left arm of chromosome 13). A schematic of the locus is shown in Figure 5.18. We can use the HiC data to search for interactions between this locus and the remainder of the genome. Figure 5.19 shows a close-up view of the locus itself for the *Hind*III HiC data. Red lines indicate *Hind*III interaction sites, and the number of split reads mapped within 100 base bins are shown on the y-axis of the top panel. The interactions of these loci with the entire 23 megabase genome are shown in the bottom panel.

The decay in interactions as a function of linear distance within chromosome 13 is immediately visible as the series of adjacent bins of increased interaction frequency. This is again the hallmark of polymer folding and the increased probability of contact that is dictated by the constrained conformational space of a polymer. We note that this behavior can be used to assist with genome assembly. In general, it can be difficult to localize individual *var* genes on partially assembled chromosomes or supercontigs. With chromosome conformation capture data, the localization is immediately clear from the polymeric signature of the spatial interaction data.

There are a small number of elevated sites of inter-chromosomal interaction with the A4 *var* locus. These include the right arm of chromosome 9 and the left arm of chromosome 8. Nevertheless, it is unclear if these are the actual locations of these blocks of interacting sequence within the IT strain as the telomeres remain unassembled and contain mosaic sequence from different parts of the genome. We have therefore not yet attempted to validate these loci by FISH. What is clear,

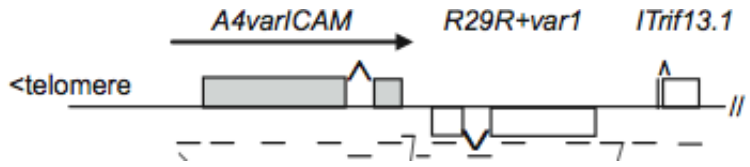

Figure 5.18: Schematic of A4/R29 locus from reference [76]). The IT clone can be selected to express a particular *var* gene, termed the A4var, by selection with a monoclonal antibody that recognizes the protein product of this gene. The A4 gene lines on the Watson strand on the beginning of chromosome 13, proximal to the telomere. Another *var* gene, R29, and a *rifin* lie adjacent to it. The sequence of this locus is from reference [76].

however, is that the locus does possess defined interaction preferences. It will be of interest, when the data become available, to identify if changes in these long-range contacts are associated with switches in *var* gene expression.

It is also possible to investigate the spatial relationships between *var* genes. While these again suffer from the problems discussed above, this type of analysis will be important in analyzing datasets coming shortly which address some of the shortcomings of the present datasets. The network of interactions between *var* genes is shown in Figure 5.20. It is clear that *var* genes have defined spatial preferences for one another. While again this network is only as good as the fidelity of the underlying sequence assemblies, we expect this type of analysis to be informative once interaction data from the 3D7 reference strain is available.

5.5.1 DCJ lines

In general, population-level studies of gene expression in multi-gene families are difficult. These obstacles are further confounded for *var* genes which are expressed in a mutually exclusive fashion in a single cell. For example, in a population of parasites, a change in the level of expression of a particular *var* is much more likely to signify that a greater fraction of the cells have switched to expression of that gene than it is to mean that the gene is "up-regulated" in the traditional

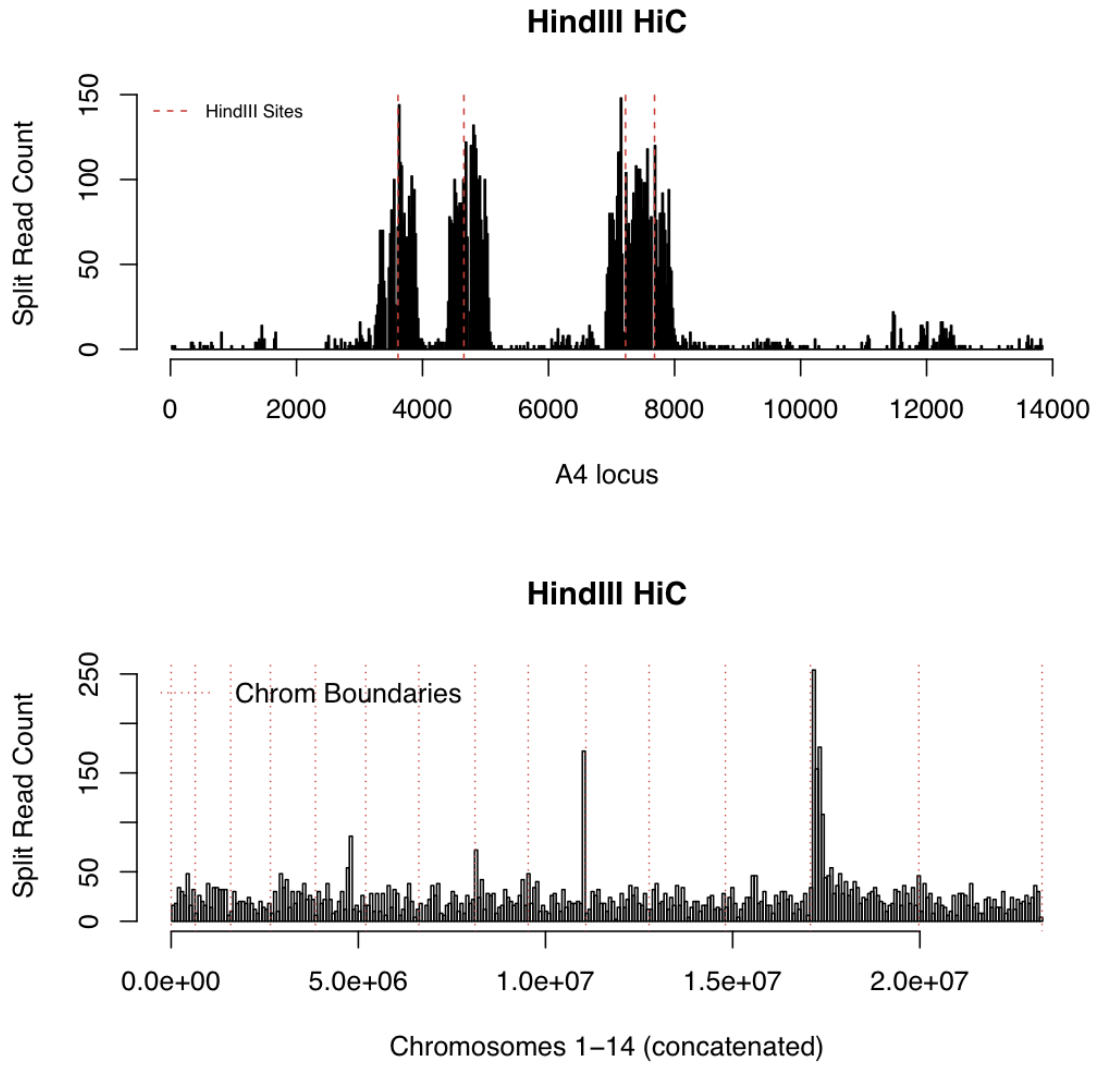

Figure 5.19: The split reads mapping to *Hind*III digestion sites along the A4 locus (top panel), and the contacts of this locus with the rest of the genome (bottom panel). The top panel is similar to Figure 5.3 but at a much higher spatial resolution. Split reads cluster around restriction sites confirming that the data provide high quality interaction information on a site-by-site basis. In the bottom panel, the decay in interaction probability as a function of linear distance can be seen on chromosome 13, which marks the physical location of the *var* gene. Specific interactions are apparent with other areas of the genome, including chromosome 7 and chromosome 9.

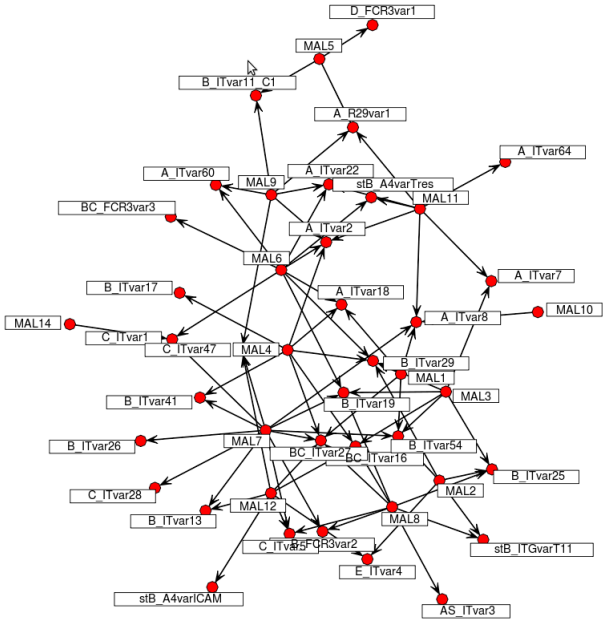

Figure 5.20: A draft of the *var* interaction network with available sequence from the IT clone. In this diagram, the nodes represent the genomic sequences of the *var* genes in the IT parasite, and edges represent interactions supported by more than a threshold of reads (in this case, 2 split read pairs). This picture remains a draft because the underlying genomic regions containing these *var* genes are not assembled in the IT parasite. Nevertheless, this type of analysis highlights the power of HiC to provide detailed information about the interactions between an arbitrary number of loci, a significant advantage of specialized imaging techniques such as FISH. In this case, interaction datasets have the resolution to study the interaction between individual *var* genes.

sense of gene expression changes. In general, studies of *var* genes require reagents which can be used to generate a culture which is homogeneous for the expression of single *var* genes. The monoclonal antibody BC6 (see Appendix A) recognizes the A4 *var* gene present in the IT parasite strain and be used to select parasites of that strain away from or toward homogenous expression of that gene; however, assembly of the IT genome is not perfect, and the areas of greatest interest to us in the case of *var* gene regulation are also those areas—sub-telomeres and internal *var* clusters—that are most challenging to assemble from short reads.

We therefore sought to use lines, generated in the laboratory of Professor Kirk Deitsch and kindly shared with us, which contain a drug-selectable marker that activates a *var* gene promoter. These lines, termed "DCJ", contain a BSD-inducible promoter, introduced at the PFB1055c locus by a double-crossover recombination event. The transgenic parasites were sub-cloned by limiting dilution and in the absence of Blasticidin-S, preferentially express PFD1015c, an internal *var* gene on chromosome 4. When grown under Blasticidin pressure, the entire population of parasites can be selected to activate the PFB1055c locus, a sub-telomeric gene approximately 915KB into chromosome 2.

We performed GCC and HiC in this system, using *Hind*III and *Mbo*I, to detect changes that occur on activation of a different *var* gene. DCJ lines were evaluated for expression of *var* genes just prior to chromosome conformation capture, confirming that PFD1015c was expressed in the lines off blasticidin-S and varBSD was expressed in lines after addition of the drug.

Preliminary results (data analysis is still ongoing) for the chromosome conformation capture are shown in Figure 5.22, which displays a heatmap of the intra-chromosomal interactions for chromosome 4. The "DCJ off" lines are grown without drug and express PFD1015c. The "DCJ on" lines are grown in the presence of drug for 4 weeks. Several changes can be seen in the interaction pattern of chromosome 4, although the effect is visually subtle because background polymer

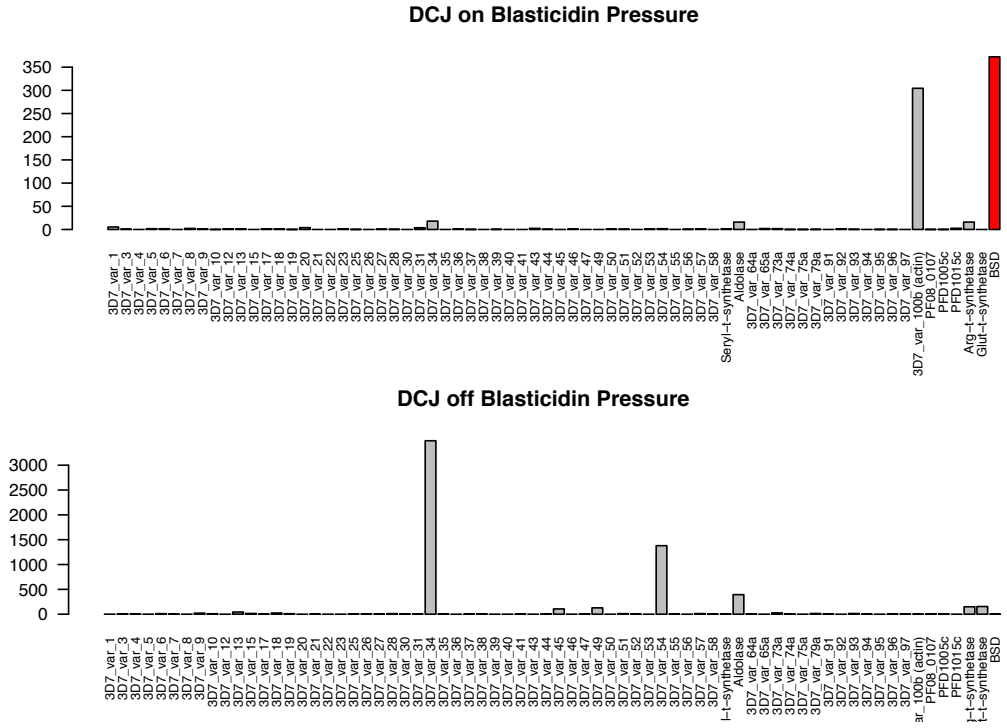

Figure 5.21: Expression of *var* genes in the DCJ lines was monitored prior to performing the chromosome conformation capture experiments. Primers from reference [124] with modifications from [37] were used to study the expression of 50 var genes. Relative expression is on the y-axis for the individual *var* genes marked on the x-axis. The names refer to the primer sets from [124]. The DCJ line off blasticidin expresses the primer set var_34, which corresponds to PFD1015c, while after blasticidin treatment the parasites express varBSD, a blasticidin S deaminase gene driven by a chromosomal integrated *var* promoter. The primer set var_54 is known to cross-react with PFD1015c [124], accounting for its increased expression in the bottom panel. Levels of varBSD are colored red in both plots. Interestingly, the levels of actin are also induced after BSD treatment. Expression is measured relative to seryl-t-synthease controls and calculated by the $2^{\Delta Ct}$ method.

contacts dominate the interaction picture. The most prominent difference occurs at the internal *var* cluster that contains PFD1015c, approximately four-fifths of the way into the chromosome. A heatmap of difference is shown in Figure 5.23. Prominent red and green areas (representing 2KB bins that represent increased and decreased interactions, respectively) can be seen in the location of the active *var*. The "marginal interaction" in Figure 5.25 represents the row (or column) sums of the interaction matrix and is a measure of total interactions sequenced for a locus. In the BSD-selected lines, in which the locus has been turned off, we find substantially more interactions, suggesting that in order for the locus to become active it has moved to an area of the nucleus which contains fewer total spatial interactions or less densely packaged chromatin. This observation is consistent with the results of Freitas-Junior, who observed a transition from a silenced *var* residing in a heterochromatin to a euchromatin area of the nucleus marked by acetylated H4 [50].

It is interesting to examine the fine-resolution changes that do occur. The right hand-panel of Figure 5.23 shows a zoomed version of the changes that occur at the PFD1015c locus. At the 3' end of the gene, one can note prominent increased contacts with the 3' region of adjacent *var* genes. The region just 5' of PDF1015c seems relatively unchanged when it is silenced; however, the upstream non-coding RNA, creates a new contact with the intron PFD1005c when PFD1015c becomes silenced. This non-coding RNA, RNAzID:3217, also appears to lose a slightly weaker interaction with the region 5' of PFD1015c. Interestingly, some of the most active spatial reconfiguration appears to occur at the intervening *rif*, PFD1010w, which both loses and creates strong contacts when the locus is silenced. These effects are largely reproducible in another independent clone, B15C2, which possesses the BSD cassette integrated via a single-crossover on chromosome 12 (the integration event is described in detail in [37]), grown with blasticidin S. In this comparison, similar to the one described above, changes

occur most substantially in the second exon of PFD1015c and at the adjacent *rif*, PFD1010w. The consistency of these changes, and the localization of the signal to the site of altered expression in both lines, suggests that the HiC assay can reliably detect genomic spatial changes in an unbiased fashion.

Overall, we see two major effects on spatial interactions when the active *var* is changed. The first is a decrease in total interactions when the locus is activated and the silencing mechanisms removed, consistent with the known movement of the locus away from areas of heterochromatin and into euchromatin [50, 117]. Second, a reconfiguration of the local chromatin environment occurs which includes decreased interaction of the 5' region of the gene with the 5' region of downstream *var* genes, and an increased interaction of the 5' region with the adjacent locus containing a non-coding RNA.

5.5.2 A Role for Entropy?

The reconfigurations that occur after the inactivation of PFD1015c are notable for an apparent absence of specific long-range contacts that are generated or lost. This was in some ways an unexpected finding, since in the olfactory receptors, a mammalian system of gene family silencing and mutually exclusive expression, Axel and colleagues found that spatial interaction a single long-range enhancer, the "H-element", conferred expression activity on a single olfactory receptor gene [88]. The data we have collected so far suggest that malaria may not possess an H-element-like mechanism to enforce the mutually exclusive expression of *var* genes. In the absence of a deterministic mechanism to enforce the preferential affinity of two loci (i.e. an enhancer and an active locus), it is instructive to consider what the possibilities are for generating proper sub-nuclear localization of the active *var* genes.

One possibility is that there is a set of non-DNA factors that preferentially

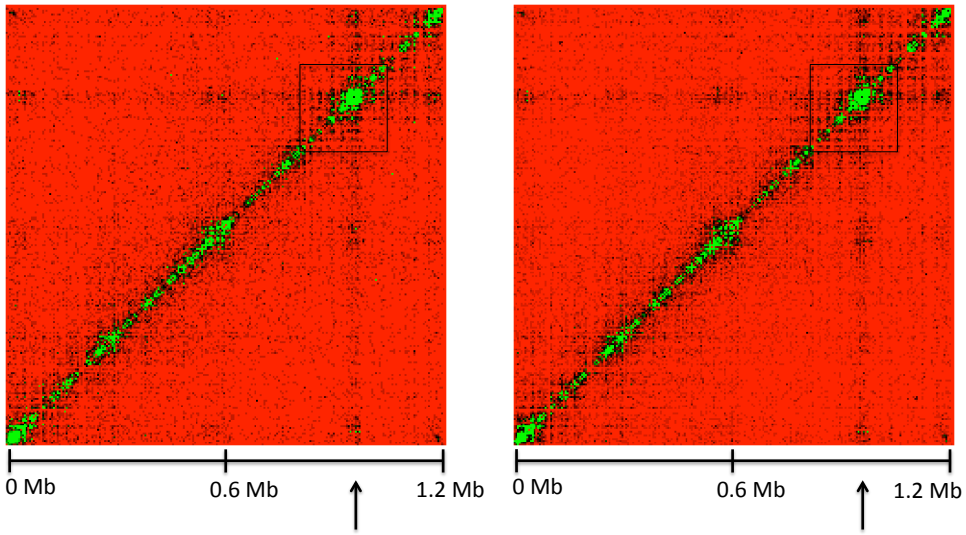

Figure 5.22: The intra-chromosomal interaction heatmaps for chromosome 4 are shown for DCJ lines grown in the absence (left panel) and presence (right panel) of Blasticidin S. In the left panel, the parasite expressed PFD1015c, an internal *var* gene in the second cluster on chromosome 4, whereas in the right panel, the *varBSD* locus on the left end of chromosome 2 has been activated. Rearranged contacts can be seen at several sites along the chromosome. In particular, the area adjacent to the PFD1015c locus shows altered spatial contacts, and the locus makes fewer total contacts, suggesting that it has both changed conformation and is present in a less constrained area of the nucleus when active. The increased sites of interaction within the locus cluster around the promoter of PFD1015c, suggesting that a local chromatin loop is formed locally when the *var* gene is active. While portions of the chromosome directly in the promoter of the active gene appear to increase in interaction frequency, those regions just upstream and downstream of the promoter and 3' end of the gene seem to decrease in interaction frequency, shown by the green shading box-like structure.

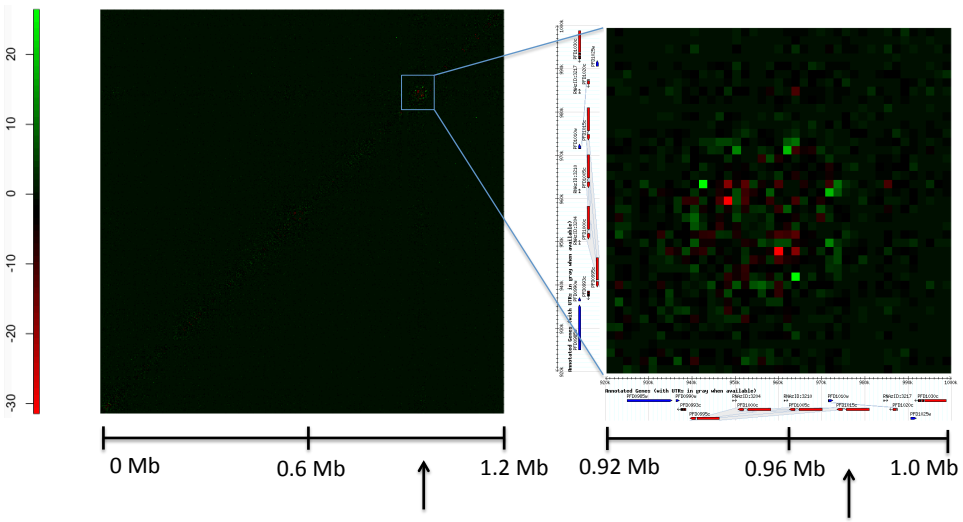

Figure 5.23: Heatmaps showing the difference in the number of split reads between parasites in which the DCJ line was grown with blasticidin S pressure and without. Unlike the heatmaps for the interaction matrices, A, in which each element corresponds to the total number of interactions between chromosomal locus i and locus j, these heatmaps correspond to the difference of two interaction matrices, i.e. the heatmap matrix $C = A - B$ where A is the interaction matrix for one line and B is the interaction matrix for another line. In this case, green dots correspond to sites of increased interaction when the PFD1015c locus is on (the DCJ clone without blasticidin S), whereas red dots correspond to increased interaction when the locus is off (DCJ clone with blasticidin pressure). The locus-specific rearrangement and modified interaction with adjacent *var* genes can be seen. When the locus is on, there appear to be interactions among the 3' regions of the adjacent *var* genes.

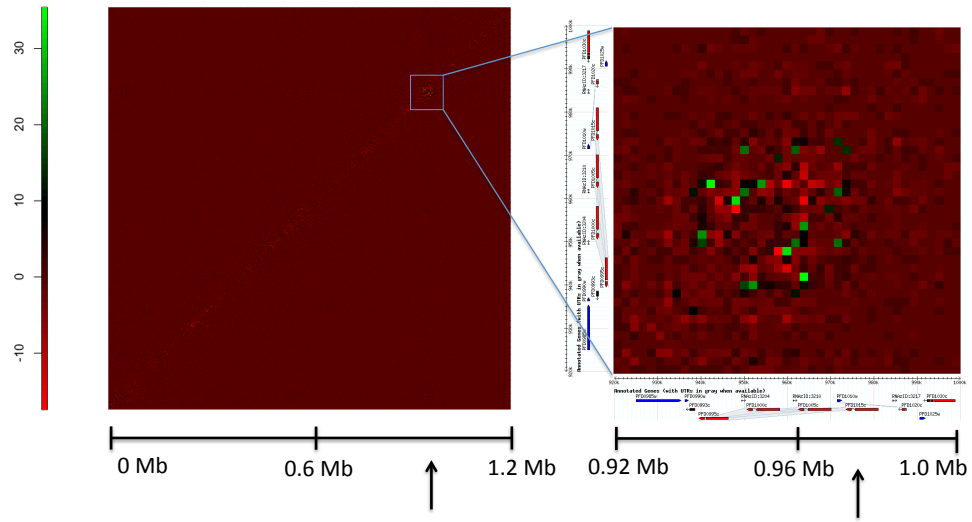

Figure 5.24: Chromatin rearrangements when a *var* gene is turned off are reproducible. Heatmaps showing the difference between the DCJ clone grown without blasticidin S and an independent clone, B15C2 [37], which possesses a copy of the BSD gene integrated into chromosome 12 via a single-crossover event. The plot on the right gives a zoomed view of the internal *var* cluster containing PFD1015c, showing major reconfiguration at that locus. Green dots correspond to sites of increased interaction when the PFD1015c locus is on, whereas red dots correspond to increased interaction when the locus is off. Similar to the effect in Figure 5.23, locus-specific rearrangement can be seen, with increased interaction between the 3' regions of the adjacent *var* genes. The red hue of the plots results from the fact that a smaller number of reads was generated in the B15C2 library. This effect can be removed by normalization, although for equivalence to Figure 5.23, we chose to present the raw data here.

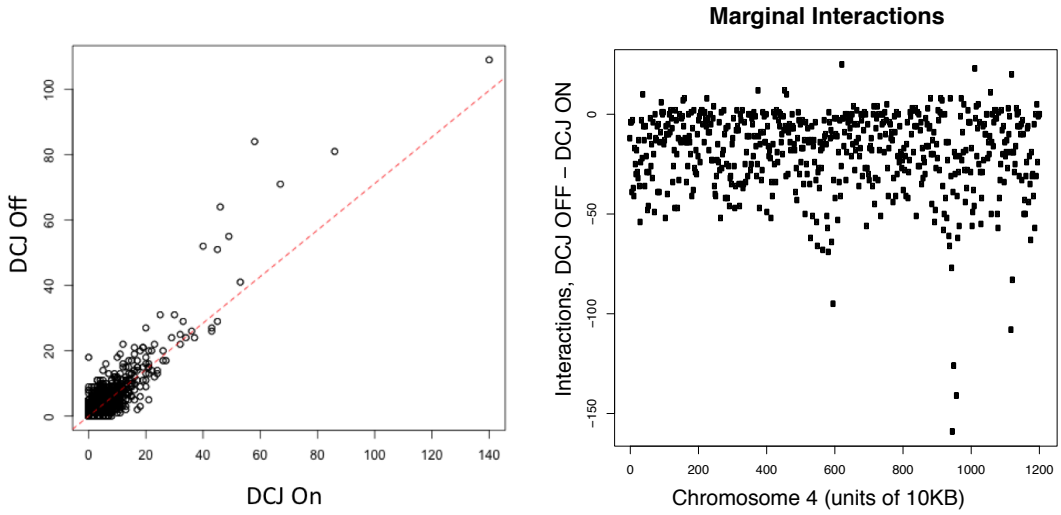

Figure 5.25: Scatterplot of interactions in DCJ Off vs. DCJ On lines (left), showing high reproducability of the technique and changes in only a small number of chromosomal locations. The row sums of the difference matrix, also known as marginal interactions, provide a measure of the total number of interactions that the locus is making. The difference in marginal interactions between DCJ off and DCJ on (right panel) demonstrates that the active *var* locus participates in fewer overall spatial interactions than the inactive locus, consistent with a model in which the silenced locus is condensed and moved away from actively transcribing foci.

associate with the active locus. These factors, which would not be detected in our assays, might be responsible for anchoring the active locus in a particular area of the nucleus. Chromatin-associated proteins or non-coding RNAs are an attractive mechanism for mediating such an effect; a molecule with affinity for the active locus could interact with the modified chromatin environment at the active *var* and actively transport it to this site.

A second possibility is that biophysical properties of the chromatin fibres themselves generate the unique spatial positioning. There have been a number of recent studies on the role of polymer entropy in defining spatial relationships in the nucleus. For example, Jun and Mulder showed that entropic forces can cause segregation of chromosomes in a rod-shaped bacterium [69]. Cook and Marenduzzo also showed, using simulations, that entropic mechanisms alone are sufficient to generate the preferential positioning of heterochromatin sequences at the nucleus [20]. For polymers, entropy is directly related to the number of conformations they can access. The nuclear periphery is a 2-dimensional surface which has additional, different spatial constrains relative to the 3-dimensional freedom of the nuclear interior. As Cook and Marenduzzo report, at low polymer concentrations, stiff polymers have larger radii of gyrations and therefore "feel" the impact of the nuclear membrane sooner in terms of their obtainable statistical configurations. As a result, at low concentrations, stiffer fibres are prone to reside in the nuclear center. Yet as the concentration of polymers in the nucleus increases, stiffer fibres lose conformations more quickly, pushing them to the exterior of the nucleus.

This finding may account for the finding that heterochromatin tends to be located at the nuclear periphery in eukaryotes including malaria [50, 139]. Nevertheless, demonstrating that entropic forces can generate the observed positioning of heterochromatin *in silico* is different from showing that entropic forces are actually responsible for such positioning *in vivo*. Still, it is interesting to speculate that perhaps it is the loss of the heterochromatin state itself that drives the active

gene away from condensed foci of heterochromatin at the nuclear periphery. Like the active transport by non-DNA factors, such a model would be consistent with the absence of specific new long-range interactions.

5.5.3 Ongoing and Future Work

There are a number of questions raised by the results in this chapter. For example, How reproducible are the changes in interaction frequency that we observe? How do these changes in chromatin structure correspond to changes in epigenetic state and transcript levels? Do similar changes occur when a sub-telomeric *var* gene is activated? What is the functional meaning of the individual changes in interactions we observe? It would, for example, be interesting to test the effect of deletion of upstream elements of the *var* gene on spatial interaction.

For some of these questions, we expect to have answers soon. For example, we are in the process of obtaining RNAseq, ChIPseq, and GCC/HiC libraries on BC6 selected and unselected lines. With the improved reference now available for the IT clone, we anticipate that the data from this experiment will go a long way toward answering several of the questions above, particularly those concerning correlated changes in mRNA levels and epigenetic marks. Since the A4var gene is a sub-telomeric gene, it will also be interesting to examine similarities and differences in the chromosomal reconfigurations that occur.

Another area of ongoing work is in generating 3-dimensional models of chromosomes which are most consistent with the observed contact probability measurements. If one is willing to assume a relationship between contact probability and average spatial distance, then the contact probability matrix can be converted to a distance matrix, and the structure most consistent with this distance matrix can in principle be found by optimization methods. For example, we show the best-fit structure for Chromosome 3 in Figure 5.26. The details of how this structure is

computed are given in Appendix E. We place this under ongoing and future work because a number of issues remain in generating these structures.

The main problem with this approach is that it is not clear what the relationship is between contact probability and distance. A second problem is that it does not really make sense to describe a dynamic entity such as chromatin by a single, static structure. A third problem is that the computations rapidly become intractable for all except the most coarse-grained models or small sections of chromosomes. In principle, a structure for the entire genome could be computed but it would be too coarse-grained to be meaningful. The same is true for large chromosomes. In general, while we believe generating these types of structures is important and useful, we are still investigating this area and looking for good solutions to some of the problems mentioned above.

5.6 Conclusion

We hypothesized that the spatial folding of chromosomes plays an important role in influencing gene expression. To test this hypothesis, we generated high resolution maps of chromosome folding using cross-linking, restriction digestion, proximity-based ligation, and massively parallel sequencing. Using two different restriction enzymes and two different parasite lines, and hundreds of millions of sequence reads, we generated a catalog of interactions of the malaria chromosomes at the late ring stage. Examination of intra-chromosomal interactions confirmed that the dominant spatial organizing force in the malaria parasite is the statistically constrained set of conformations which are available to connected molecules such as polymers. The intra-chromosomal interaction data also showed s^{-1} scaling reminisci]ent of the 'fractal globule' proposed by Lieberman-Aiden and coworkers [86], while at large intra-chromosomal distance the chromosomes appeared to be affected by an 'anchoring' such that distant ends interacted far less often than

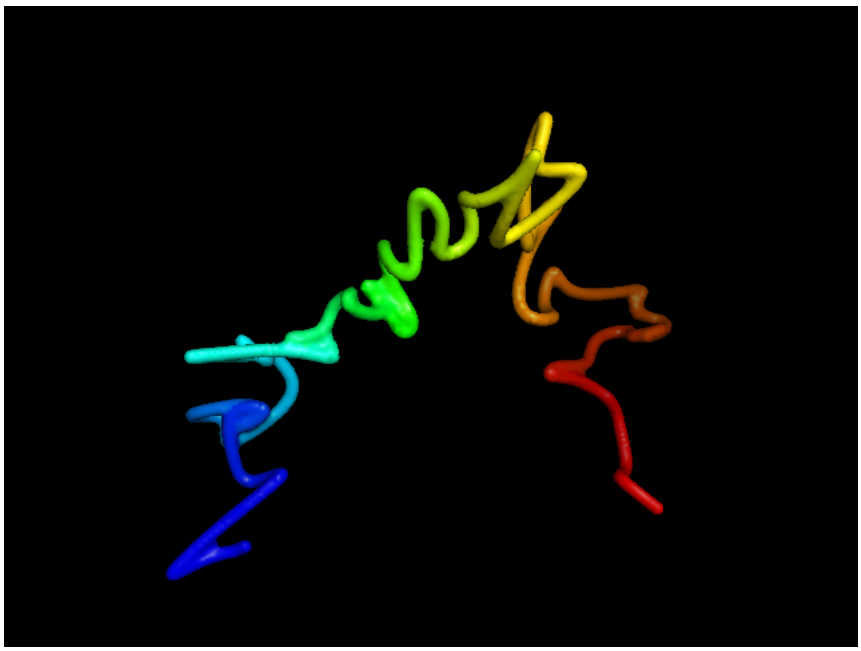

Figure 5.26: Best-fit structure for Chromosome 3. This structure is a model created under the assumption that contact probability can be considered an inverse function of distance. If this is the case, the interaction matrix, A, can be converted to a distance matrix, D, and the the configuration of points, $\{x_i\}$ can be found which minimizes the difference between the "measured" distance, d_{ij}, and the distance in the reconstructed space, $\|x_i - x_j\|$. Choosing the $\{x_i\}$ is an optimization problem. This approach is discussed in detail in Appendix E.

would be expected by chance.

Examination of inter-chromosomal interactions revealed strong connections between chromosome folding and epigenetic data. Regions of the chromatin fiber such as sub-telomeres and internal *var* clusters interact with each other at a much greater frequency than other portions of the chromosomes; these regions were also strongly correlated with the presence of heterochromatin repressive marks. Clusters of ribosomal DNA elements were also shown to interact with this assay. Both of these interactions were consistent with what has been previously reported in the literature using FISH, adding confidence to the sensitivity and specificity of the method to detect genuine interactions. Overall, the inter-chromosomal interaction data, when combined with epigenomic data, paint a picture of a *P. falciparum* nucleus whose principal structure is specific at the level of the epigenetic status of the chromatin fibers. The division between heterochromatin and euchromatin specifies which compartments are likely to interact; however, within a given compartment, free diffusion seems to play the major role in determining which loci contact one another.

Nuclear reconfiguration has been shown to occur upon activation of a silenced *var* locus, and we were able to study this process in depth using chromosome conformation capture-sequencing methods. By generating interaction maps in isogenic lines with a *var* promoter that can be activated in response to the drug blasticidin S, we found that an active *var* gene makes new contacts and loses old ones; however, it is predominantly local and not global reconfiguration which occurs. In contrast to the odorant receptors in which a specific "H" element appears to interact with the active gene, no element was found to 'bless' the activated *var* gene. Instead, a major change which was apparent was simply a reduced number of intrachromosomal contacts at the activated locus, suggesting it had moved from a heterochromatin compartment to a euchromatin compartment. This once again underscores the strong linkages between epigenetic state and the spatial struc-

ture of chromatin in determining gene expression. It also raises the interesting possibility that the chromatin density is itself a localization signal. Based on theoretical and empirical results in which the preferred statistical configurations of a fibre change based on the variations in density along its length, we suggest that one a model in which the sub-cellular location of a locus is determined is simply by the by fiber density. In this model, euchromatic loci are driven by entropic forces toward the middle of the nucleus, whereas denser heterochromatin is similarly "pulled" to the periphery. This model makes the testable prediction that subnuclear location is specific by chromatin compaction alone and not specific epigenetic marks, and it would be interesting to test in future experiments.

Chapter 6

Conclusion

6.1 Summary of Results

The malaria parasite *P. falciparum* undergoes profound morphological, cellular, and biochemical changes during its lifecycle. These changes are mediated by alterations in patterns of gene expression, and gene expression patterns, broadly defined, are therefore a primary mediator of essential parasite processes such as virulence, drug resistance and immune evasion.

The goal throughout this thesis has been to understand the patterns of gene expression and the mechanisms behind those patterns. In beginning to investigate the patterns of transcription, we repeatedly found ourselves asking whether the observed changes reported in the literature were genuine or whether they could be accounted for by differences in the temporal progression or synchrony of cultures. We therefore developed ways to address this issue before proceeding with mechanistic studies on the molecular determinants of gene expression. Because of the apparently conflicting results in the literature [27, 28], and because of difficulties in analyzing datasets from our own lab, we felt it was essential to answer these questions before further study of gene expression took place.

6.1.1 Statistical Descriptions of Gene Expression

In the first part of this thesis, we developed and tested methods for analyzing malaria gene expression profiles. Together, the toolkit of approaches we develop helps to overcome many of the problems initially encountered in the analysis of malaria expression profiles. It paves the way for valid experimental measurement of relationships between variables without the confounding effect of temporal progression. At a cell biological level, these include chromatin state and chromosome structure, while at a population level these include examining associations between severe and mild disease.

In Chapter 2, we identify genes whose expression is up- or down-regulated in

cases of severe vs. mild malaria. We discuss the issue of temporal heterogeneity among samples and argue why an improved analysis is required, though we defer a full analysis including stage until Chapter 3. We show how to decompose profiles into their molecular patterns through the use of matrix factorization techniques, particularly the SVD. Using this approach, we identify individual genes which are important during individual time points and at certain moments during sexual development. We also demonstrate how the SVD can be used to understand the expression patterns occurring in individual samples, and use it as a heuristic to decompose samples into their component mixtures.

In Appendix C, we provide a suite of algorithms and methods aimed at solving, in a technical sense, the issues raised in the introduction and in Chapter 2. Maximum-likelihood estimation under independent Gaussian probability densities for each gene is the first algorithm discussed (and the one most often used). We then consider improvements to this method, include robust densities, the incorporation of microscopy data, lack of independence between genes, as well as methods capable of handling mixtures of time points. An important special case is the case of commitment to the sexual lineage from the asexual one. We compare the methods developed to existing methods in the literature, and give conditions under which the approach is related to that of PlasmoDB. Finally, we address the issue of gene subset selection, and suggest a set of genes which could be used for stage estimation using real time PCR.

Chapter 3 addresses several real-world problems which are important to malaria research using the techniques developed in the appendix. We use stage and lineage estimation to suggest that previously observed transcriptional variation, proposed to correspond to altered physiological states of malaria *in vivo*, are not simply parasite populations of different synchrony or temporal progression. On the other hand, estimation of lineage commitment allowed us to establish the likelihood that these apparent clusters were in fact a continuous mixture of sexual and asexual

parasites, at a stage when those two are indistinguishable under the microscope. Furthermore, exploration of the apparent discrepancy between continuous mixtures of states vs. discrete clusters of states revealed that artifacts introduced by processing arrays during different times and in subsets were influencing the data and likely generating the patterns of discretized variation. We demonstrate that our approach to analyzing malaria microarray data effectively addresses most of the confounding issues of stage and sample mixtures that plague real lab experiments.

By removing the confounding effect of asynchrony and differing temporal progression, the statistical description of malaria gene expression patterns greatly facilitates the investigation of mechanisms of gene expression. The molecular mechanisms governing gene expression in *Plasmodium falciparum* are incompletely understood, and because an understanding of these mechanisms could lead to improved vaccines, reduced drug resistance, and new candidate drugs, we moved from descriptive modeling of gene expression to experimental approaches focusing on causal mechanisms underlying changes in gene expression.

6.1.2 Molecular Determinants of Gene Expression

Initial sequencing of the *P. falciparum* genome revealed an ostensible absence of specific transcription factors. To account for this deficiency, several groups suggested that epigenetic mechanisms were likely to play a role in structuring the gene expression program that underlies the morphological and functional diversity that occurs throughout the parasite lifecycle. This postulated role for epigenetic control of gene expression was later elegantly confirmed by several groups, first by identifying the role for the histone deacetylase PfSIR2 [36, 50], and then second by the discovery that tri-methylation at Lysine 9 of histone H3 (H3K9me3) serves as a memory mark for active *var* gene transcription [17].

The initial discovery of the loss of H3K9me3 on activated *var* genes was made using transgenic parasites which do not express a functional PfEMP1. We first sought to confirm this finding using wild-type parasites of the IT genotype which express normal levels of PfEMP-1. Using a system in which wild-type parasites can be selected for expression of a particular PfEMP-1 variant using a monoclonal antibody previously generated in the Newbold lab, we confirmed this finding in our parasite lines and also showed the H3K9me1 appears to be at reduced levels in an actively transcribing *var* gene. We then went on to profile epigenetic changes over time as well as over space in the genome. In a timecourse performed with samples isolated every 8 hours, we demonstrated the the amounts of modified nucleosome change dramatically during the lifecycle. In a genome-wide ChIPseq assay, we also showed that the distribution of H3K9me3 marks localized with high specificity to antigenic clusters at sub-telomeres and internal repeat regions. This suggests that PfEMP-1 is under genome-wide epigenetic control and also that other gene families in these regions such as *rifins* and *stevors* may similarly be controlled by epigenetic mechanisms.

Our findings in this research were very similar to those obtained by other groups. The Scherf laboratory demonstrated the genome-wide distribution of the H3K9me3 epigenetic mark, showing its close concordance with antigenic gene families [90], while the Stunnenberg laboratory established the time-dependence of several key marks including H3K9me3, H3K4me3, and H2A.Z. Collectively, these studies as well as our data offer compelling experimental evidence for the essential role that epigenetic mechanisms play in gene expression of the malaria parasite.

The spatial correlations between epigenetic marks, as well as the known movement of the *var* locus during activation, suggest that the spatial biology of the nucleus is closely linked with epigenetic silencing or activation. We hypothesized that spatial reconfigurations are occurring locally and globally during *var*

gene activation, and sought to test this hypothesis by generating spatial maps of the nucleus at a genome-wide level. The advent of second-generation DNA sequencing techniques greatly facilitated this approach. We established a chromosome conformation capture assay (3C) in malaria for the first time, and extend it to a high-throughput version suitable for studying the architecture of the entire genome. This is done by direct sequencing of DNA libraries after the ligation step in the chromosome conformation capture procedure. The limited resolution and low throughput of fluorescence in situ hybridization studies meant that global studies of the *P. falciparum* had previously been technologically infeasible.

After validating the assay, we generate maps of chromosome folding on a genome-wide level. We generate maps using two different enzymes and show the strong reproducibility of the technique under various conditions. Using HindIII, our atlas of genomic interaction has a resolution of approximately 25 kilobases, while that increases to roughly 5KB when MboI is used. These are the first genome-wide interaction maps of this type for malaria, and because the small size of the malaria genome facilitates high-resolution digestion, are the first maps of such fine resolution in any organism. These maps confirm known interactions in the genome, including ribosomal DNA clusters and heterochromatic foci at the periphery of the nucleus, in addition to identifying a large number of new interactions. After establishing the basic folding principles of the genome, we then go on to study how the chromosomes reconfigure during antigenic variation. We demonstrate that the chromatin surrounding the DNA loosens, and also that an active locus fails to form more than a small handful of new and specific contacts. The driving force which reconfigures the locus is not known, however, we speculate that entropic mechanisms may play a role in relocating more flexible portions of the chromatin fibre to a distinct region of the nucleus by free energy.

Overall, our work can be divided into three main contributions. First, our work offer computational tools to biologists studying malaria gene expression.

Second, we used these tools to establish a model in which the circulating, ring-stage parasite burden in each patient infection is a mixture of asexually developing and sexually committed parasites with distinct gene expression patterns. Third, we have created global maps of chromosome folding which highlight the spatial patterns of the *Plasmodium* nucleus and its changes upon *var* gene activation. All three of these areas are of major interest for follow-up, which we discuss below.

6.2 Follow-Up and Future Work

6.2.1 Statistical Models of Gene Expression

In this thesis, we have provided several methods that permit estimation of temporal progression and lineage commitment, and for one of these methods performed extensive validation. This provides benefit to the analysis of expression studies in two major ways. The first is that it can be used as a 'check' or 'verification' that the samples under study are of equivalent age and lineage. The expression profiles between groups can be then be compared. The second benefit is that the estimated time can be directly incorporated into the analysis procedure. There are several good methods and software implementations that allow for the analysis of time-series microarray data. Knowing the true time for the individual points in the time series will improve statistical power and reduce the number of false positives and false negatives.

One area in which additional improvements could be made is to directly incorporate the periodic pattern of malaria expression profiles into a statistical method to detect differential gene expression. Having such a tool, an experimenter could sample a number of time points and then rank genes which are different in terms of their amplitude as well as their phase. This is important because currently fairly rudimentary methods are used to assess differential expression [56,87], which likely

underestimate the true expression plasticity of the genome and reduce the power to detect regulatory regions.

In addition to better statistical tools, more high-quality datasets are needed. The dataset of Bozdech [9] and colleagues is a great example of the far-reaching and often unanticipated benefits that come from having high-quality, high-resolution datasets. Datasets that offer absolute quantitation along with high resolution over time and lineage development are needed, particularly in difficult-to-culture stages such as mosquito and liver stages.

6.2.2 Ring-stage Gametocytes

Applying stage-estimation algorithms to patient malaria datasets generated a discernible and frequently substantial signal of gametocyte transcripts in nearly all patient samples investigated. For the 43 *in vivo* samples, almost all of them were without slide-positive gametocytemia, suggesting that a portion of the ring-stage parasites are committed to the gametocyte lineage and expressing gametocyte-like transcripts. While a number of studies have demonstrated sub-microscopic gametocytemia [104, 131, 132], and its importance in epidemiology and transmission of malaria is appreciated [70, 105], our data suggest that there is also an important population – at least from a gene expression perspective – of microscopically observable gametocytemia that is indistinguishable by morphology from asexual ring-stage parasites.

This finding should be confirmed by replication in future studies of gene expression in parasites sampled *in vivo* and *ex vivo*. Another question is to what extent the substitution of the mature gametocyte profiles for the committed, but morphologically indistinguishable gametocyte profiles is acceptable. The expression patterns of early, committed gametocytes should be measured using cell sorting based on fluorescently-labelled early gametocyte proteins.

6.2.3 Chromosome Folding and Epigenetic Regulation

The coupling of second-generation sequencing to classical epigenetic techniques such as chromatin immunoprecipitation and chromosome conformation capture has dramatically increased our ability to research nuclear events. In many ways we are only beginning to exploit the potential of these types of technologies when coupled together. The future holds many opportunities as well as some obstacles which need to be addressed. We summarize some of those opportunities and difficulties here.

We have studied the folding of the genome at a single timepoint under conditions of variable *var* gene activations. How do these patterns evolve over the course of the lifetime of a cell? What about during the asexual stage vs. other stages? Is it possible that genes whose expression patterns are strongly covariant are also spatially colocalized? This hypothesis could be tested by generating spatial data from multiple timepoints.

In terms of *var* gene regulation, we have identified local reconfiguration of the chromosome fiber that occurs during switching. This was observed in a single case where an internal *var* gene went from active to silenced locus, and a subtelomeric *var* gene was turned on. Does this hold for other types of transitions, such as subtelomeric to subtelomeri, or internal to internal? What about for parasites that were wild-type and not transgenic? The spatial rearrangements that occur need to be interpreted in the light of switching data, which at the present suggest that internal *var* genes switch off at a much slower rate that subtelomeric *var* genes.

Furthermore, the driving force behind relocation of the active transcript to a transcriptionally permissive region should be identified. The possibility that entropy plays a role could be tested by modulating the fiber density, either pharmacologically or via genetic addition or subtraction of chromatin-modifying enzymes.

Another possibility is that recognition of specific structures or sequences on the heterochromatin or its associated proteins mediates recruitment to the nuclear periphery by an active process. Distinguishing between these mechanisms, and if the second mechanism is correct, identifying the factors that recruit heterochromatin to the nuclear periphery, are of high importance.

In addition to improved temporal resolution and focusing in on the spatial changes at single loci, it will be essential to continue to push the spatial resolution of chromosome conformation capture techniques further. The assay is in theory limited by the enzyme choice, which determines the number of cut sites, and the capacity to sequence a large number of short reads. Using HindIII, we were able to create maps that had a resolution of approximately 25 kilobases, while with MboI, the resolution improved to approximately 5 - 10 kilobases. Nevertheless, it would be important to be able to assess the interactions between individual genes, particularly at promoter, enhancer, and intronic elements that are likely to be looped in complex ways. In order to achieve this, the digestion efficiency of the chromatin needs to be increased using an enzyme that cuts frequently. With screening of different enzymes and further optimization of digestion steps, it is possible that the resolution of the assay can be improved to a point where gene-gene physical interactions are detectable.

In general, we are still in the very early stages of epigenetic research in malaria. Several factors, e.g. PfSIR2A, PfSIR2B, PfHP1, and PfSIP2, have been shown to be involved in epigenetic regulation. A still larger set of molecules has been identified based on homology as participating in chromatin modification. Identifying the enzymatic activities of these proteins, and then testing their function *in vivo*, remains a major task for the near future of malaria research.

The N-terminal tails of core histone proteins are the substrate onto which the epigenetic information is written and encoded. Early modifications have been identified using mark patterns that are conserved between other eukaryoates, yet

there is evidence to suggest that *Plasmodium* has evolved a distinct function for identical modifications as well as a set of marks not observed in other organisms. Describing the repertoire of marks, generating the reagents to study them, and investigating their role in gene expression are major tasks for the future.

6.3 Implication and Outlook

It has been over a century since Ross peered down a microscope and witnessed the dramatic exflagelation of *Plasmodium* gametocytes harvested from the blood of infected individuals. This transformation, like the vast phenotypic diversity of malaria parasites during their complex lifecycle, is mediated entirely at the level of gene expression. Gene expression patterns and the mechanisms that govern them are therefore the keys to understanding and manipulating the parasite. In recent years, the dramatic pace of technological development has given researchers unprecedented ability to monitor the patterns of gene expression for every expressed sequence in the cell. In the course of our research, we have used these technologies to understand the meta-patterns of gene transcription in malaria, to identify genes whose expression is associated with severe disease phenotypes, and to understand the general structure of patient infections at the level of parasite-transcribed RNA.

In the area of epigenetics and chromosome folding, we have created high-resolution maps of epigenetic marks and chromosome folding, delineating the spatial preferences of chromosome loci at a resolution of approximately 5 kilobases, and measuring the global chromosomal refolding events that occur with *var* gene switching. Nevertheless, our research efforts in this area are part of an exciting vein of investigation, now just beginning, into the complex interplay between epigenetics, physical space, and gene expression. A major goal for the future is to define the respective contributions of specific transcription factors, chromatin marks, and spatial affinities. It seems likely that one day in the near

future malaria researchers will be able to generate concrete models specifying the relative importance of these mechanisms for each gene in the genome, providing a platform on which to guide targeted research and discovery. For example, a parasite manipulated to express all of its surface antigens might be a viable approach to vaccination; such a strategy appears to be working in the case of *Giardia* [115, 120].

The impact of malaria around the world remains as devastating today as it ever has been. New treatments and an effective vaccine are badly needed. A precise understanding of the mechanisms that control gene expression, in particular variant antigen expression and switching, are essential for the rational design of a vaccine and the development of new drugs. In 2011, malaria remains a tremendous public health burden for which there are only imperfect control measures and treatments. Reducing deaths and illness from malaria takes coordinated efforts on the part of health workers, scientists, governments, and foundations. It is hoped that the basic research findings presented in this thesis will contribute to the growing body of knowledge that allows the biomedical research community to develop new therapies and prevention measures to reduce the burden of illness and death caused by this destructive disease.

Appendix A

Materials and Methods

A.1 Parasite Culture

Parasites were cultured according to the method of Trager and Jensen [145]. Cultures were synchronized by Plasmagel flotation [111] (to obtain enrichment of trophozoite stages) or sorbitol lysis [79] (to obtain enrichment of ring stage parasites). Where necessary, highly synchronized cultures were achieved using multiple (typically 3) rounds of sorbitol synchronization and/or Plasmagel flotation. Synchronicity was checked by Giemsa stain (see microscopy).

The IT clone and its subclones were used for most experiments in Oxford. The A4 subclone [121] was positively and negatively selected for expression of the BC6 antibody [137]. Other subclones used include 3G8 [63] and R29 [121].

At the National Institutes of Health, the 3D7 subclone [53] was used for some experiments. The strains DCJ, B15C2 and E7C5 [38] were kindly shared by Kirk Deitsch. Strain DCJ was grown without drug to generate strain DCJ off, and grown with 2 μg/mL blasticidin s pressure for 1 month to generate strain DCJ on.

Freezer stocks of all parasites used are stored in triplicate at either Oxford or NIH.

A.2 Selection of Parasites Using the BC6 Antibody

positive selection Parasites were positively selected for expression of the A4var-encoded PfEMP1 by incubating 20 μL of 400 μg/mL BC6 antibody with 50 μL of Protein G-coated magnetic beads (Invitrogen) in 150 μL of 1X PBS / 1% BSA in sterile eppendorf tube at 25 °C for 2 hours. The beads were washed once with 1X PBS / 1% BSA and incubated with Plasmagel purified trophozoites at 25 °C for 30 minutes. The beads containing bound parasites were washed twice with 1X

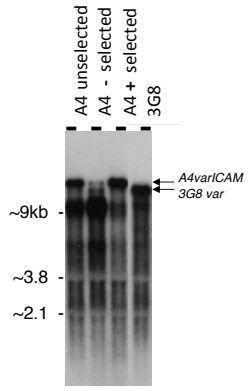

Figure A.1: Northern blot (performed by S. Kyes) showing the results of positive (+), negative (-), and no selection in an A4 clone. The blot has been probed with a conserved VarC probe which recognizes most *var* genes. The quantity of A4var increases with positive selection and decreases substantially with negative selection although there remain a few cells which express the gene even after negative selection. A subclone (3G8) expresses a different *var* and shows an absence of A4var expression.

PBS / 1% BSA, resuspended in a small volume of culture medium, and returned to parasite culture with fresh medium and cells.

negative selection Enrichment of parasite populations for BC6-negative parasites was performed by removal of BC6-positive parasites as follows. BC6 antibody was bound to beads as above. The number of parasitized trophozoites was calculated to be $\frac{1}{10}$ the number of magnetic beads. Incubation of beads and parasitized RBCs together was performed for 2 hours at 25 °C. Magnetic beads were gently pooled at the side of the tube and the supernatant was collected and returned to culture with fresh uninfected red cells and fresh medium.

The success of BC6 selection procedures was monitored by flow cytometry (Section A.3) or Northern blotting (Figure A.1).

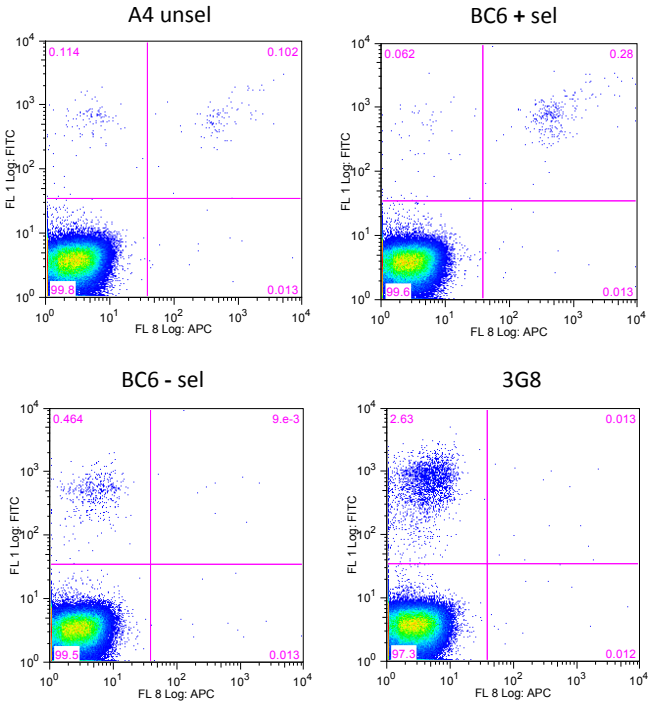

Figure A.2: Flow cytometry analysis of IT cultures using the BC6 antibody. A4 cultures were positively selected (BC6 + sel), negatively selected (BC6 - sel), left unselected (A4 unsel). The 3G8 sub clone, which does not express A4var, was also used. Note that the samples analyzed here correspond to the samples whose RNA was analyzed by Northern blotting above (Figure A.1).

A.3 Flow Cytometry

Primary antibodies were pre-cleared on uninfected red cells by incubating the total antibody used in the experiment with 100 μL of uninfected red cells for 30 minutes at 25 °C. Secondary antibodies were pre-cleared with 100 μL of infected red cells for 30 minutes 25 °C.

FACS was performed in triplicate on 10 μL of cells. Cells were washed twice with 1X PBS / 1% BSA and incubated with primary antibody, suspended in 100 μL 1X PBS / 1% BSA, for 30 minutes at 37 °C. The BC6 antibody was used at a concentration of 5μL of 400 μg / mL per well. Cells were then washed twice with 1X PBS / 1% BSA and incubated for 30 minutes at 37 °C with secondary antibody (goat anti-rabbit conjugated to APC, Invitrogen) resuspended in 1X PBS / 1% BSA along with SYBR green stain diluted to 0.1X. After washing twice with 1 X PBS / 1% BSA, cells were fixed with 1% paraformaldehyde / 1X PBS / 1% BSA and analyzed on a flow cytometer.

A.4 Microscopy

Estimates of parasitaemia were obtained by manually counting 500 cells from Giemsa stained thin films, as per standard methods. For precise measurements of parasite area, a digital imaging system was set up. A robotic stage, controlled from ImagePro Plus 6.1, was configured to tile 8 x 8 frames at 100x magnification. These images were captured onto a digital imaging system and saved in .tiff format. Parasites were identified by segmentation of images by setting RGB thresholds, and several statistics on positively staining parasite area were measured. Greater than 500 parasites per slide were quantified, and the data were exported to R for subsequent analysis.

A.5 Computational Analysis and Data Visualization

Sequence manipulation and analysis were done manually or using scripts written in perl, python, or R. Data analysis and plotting was done either in R or Matlab. Scripts are archived on the Newbold lab server. Sequence display was performed using GBrowse, Artemis, or the Integrated Genomics Viewer (IGV). Circos was used to generate circular genome plots such as those in Figure 5.17.

Microarray data was analyzed using R. The latest stable releases of R and available packages were used (currently R version 2.13.1). Microarray analysis relied on the suite of packages termed `Bioconductor` [55]. For affymetrix microarrays, the .cel files were input into R and the data were normalized using RMA. The .cel files for reference [122] were downloaded from the Winzeler lab website. The .cel files from reference [27] were downloaded from the NCBI Gene Expression Omnibus (GEO) site.

For glass-slide arrays, the data were downloaded from the Derisi lab website as normalized Cy5/Cy3 ratios. The data were log2 transformed, and the smoothed version of expression profiles from [9, 87] were generated by applying locally weighted linear regression (loess) with the span parameter set at 0.3.

Second-generation ("next-generation" or Illumina/Solexa) data were stored as .fastq files and mapped to the genome using SSAHA2 [103], Bowtie or BWA. Aligned reads were stored as .bam files. For calculation of expression values, .bam files were imported into R and counts were computed over relevant genomic regions. Publicly available datasets were downloaded from the Sequence Read Archive (SRA).

For the IT strain, a pseudo-reference genome was generated by 'morphing' the IT SNPs into the 3D7 reference strain using iCORN [109]. The second generation of the IT reference genome was constructed by Thomas Otto and the Pathogen

Sequencing Unit at the Sanger Institute.

Optimization of convex functions (e.g. sections C.3.4 and E.1) was performed in Matlab using CVX.

A.6 Chromatin Immunoprecipitation (ChIP)

Protein-DNA interactions were assayed using chromatin immunoprecipitation [73]. Several methods were tested and substantial optimization was performed. We initially used a cross-linking ChIP ('xChIP', section A.6) approach based on protocols described in [18], but subsequently developed an approach that avoided cross-linking in order to improve the yield of immunoprecipitated chromatin. This 'native' ChIP ('nChIP', section A.6) method was adapted from reference [106].

Native Chromatin Immunoprecipitation (nChIP)

Parasites were isolated from infected red cells by incubation in 0.01% saponin/1x PBS. Nuclei were prepared by incubation of the saponin-lysed pellet in 10mM Tris-HCl, 2.5mM $MgCl_2$, 14mM β-mercaptoethanol 0.5% Nonidet P-40, pH 7.5 and centrifugation in a microcentrifuge for 10 minutes at 2,500 x g. Chromatin was solubilized by micrococcal nuclease digestion as follows: Nuclei were resuspended in micrococcal nuclease digestion buffer (50mM Tris-HCl, pH 7.9, 5mM $CaCl_2$) and divided into two aliquots of equal volume. One aliquot was digested with 0.4 Units per μl of micrococcal nuclease, while the other aliquot was digested with 0.04 Units per μl of the same enzyme, both for 3 minutes. Micrococcal nuclease was inactivated by addition of EDTA at a final concentration of 10 mM. Following digestion, nuclei were centrifuged and the supernatants, which contain the soluble chromatin, were pooled. Two concentrations of enzyme were used because underdigested chromatin is enriched in di-, tri-, tetra- and penta-nucleosomes, while fully digested chromatin is composed primarily of mono-nucleosomes. This diges-

tion procedure was optimized for time and enzyme concentration. All procedures were performed at 4 °C, either on wet ice or in a chilled microcentrifuge.

Immunoprecipitations were performed in IP buffer containing 50mM HEPES, 140mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-Deoxycholate, 5mM DTT. Antibodies were incubated at a 1:40 dilution for 18 hours at 4 °C. Subsequently, Protein G magnetic beads were added at a 10:1 ratio with antibodies. Typically, a 200 μL immunoprecipitation was performed, which included 5 μL of antibody and 50 μL of protein G magnetic beads (Invitrogen). Antibody-bead complexes were washed two times in IP buffer, once in IP buffer + 250 mM NaCl, and once in 10 mM Tris pH 7.4, 1mM EDTA. Protein-DNA complexes were eluted from beads in 50mM Tris-HCl, 10mM EDTA, 5mM DTT, 1% SDS at 65 °C for 15 minutes. Beads were washed once again in 10 mM Tris pH 7.4, 1mM EDTA with 0.67% SDS, 5mM DTT at room temperature, and the eluate from both steps was pooled. DNA was then precipitated by phenol/chloroform with glycogen as a carrier.

Cross-linking Chromatin Immunoprecipitation (xChIP)

For xChIP, parasites were cross-linked with 0.5% formaldehyde / 1x PBS following saponin lysis. Nuclei were isolated by 200 strokes in a Dounce homogenizer, and then fragmented to 300 - 1000bp by sonication for 15 minutes at 30 second intervals (total sonication time, 7.5 minutes) in a Branson sonifer set on 'high'. Soluble chromatin was isolated by centrifugation for 15 minutes at 15,000 x g. Immunoprecipations were then performed as in nChIP (section A.6). xChIP was found to yield lower amounts of chromatin relative to nChIP, primarily due to inefficiencies in cell lysis following cross-linking. However, both procedures will continue to be employed, as xChIP has the advantage of being able to measure weaker protein-DNA complexes and is preferable for nucleosome localization assays which require precise nucleosome immobilization in position and rotational setting before immunoprecipitation.

Forward ID	Forward Sequence	Reverse ID	Reverse Sequence
A4_R29_locus_1_F	caatcaccccacaccacacc	A4_R29_locus_1_R	tggcatctttatcatcctcaccac
A4_R29_locus_2_F	agaaagttggagggggtgagg	A4_R29_locus_2_R	ctttcaacacatgtgccaccc
A4_R29_locus_3_F	gcgccatagtaaaaaccgcac	A4_R29_locus_3_R	acatttccacccccaatgtcg
A4_R29_locus_4_F	acccaacacaaccctaatggc	A4_R29_locus_4_R	gtcggttgttgtgtcggtttg
A4_R29_locus_5_F	tggcaaaacaccaacagaatgg	A4_R29_locus_5_R	gcgcaattgctctttgtttcc
A4_R29_locus_6_F	taaaactgatgtagcgccccc	A4_R29_locus_6_R	gcaatggatcctaatgccaaagc
A4_R29_locus_7_F	accttatgggaaaccaaaaccccc	A4_R29_locus_7_R	ttgtacattggtggggttggc
A4_R29_locus_8_F	aggaacaggacgggtatcagg	A4_R29_locus_8_R	tgcagaagaagatggtggtgc
A4_R29_locus_9_F	atagtttgcgacgtcgtggag	A4_R29_locus_9_R	aggggtggagctagtcgtaac
A4_R29_locus_10_F	gacttacacgtactcgccacc	A4_R29_locus_10_R	tccgaatttgcgaagaaaccg
A4_R29_locus_11_F	aattagcctcgtttggggctc	A4_R29_locus_11_R	cccagcacaaaagcattacgc

Table A.1: Primer pairs used to assess immunoprecipitated DNA in ChIP assays via qPCR

Primers

The primers used to asses the quantity of immunoprecipitated DNA at the A4/R29 locus in the A4 clone are given in Table A.1. All reactions were set up using both primers at a concentration of 100 nM. Real-time PCR conditions are given in Section A.10.

A.7 Chromosome Conformation Capture

Chromosome conformation capture was performed as in reference [33] with modifications for malaria. Parasite pellets were cross-linked with 1% paraformaldehyde (EM grade – Electron Microscopy Sciences) for 10 minutes at 25 °C. Cross-linking reactions were quenched with the addition of glycine to a final concentration of 125 mM and incubated for 10 minutes at 25 °C. Parasites were isolated from infected red cells by incubation in 0.1% saponin/1x PBS. Nuclei were prepared by incubation of the saponin-lysed pellet in 10mM Tris-HCl, 2.5mM $MgCl_2$, 14mM β-mercaptoethanol 0.5% NP-40, pH 7.5, and protease inhibitors (complete EDTA free protease inhibitors, Roche) at 4 °C for 10 minutes. Nuclei were collected via centrifugation in a microcentrifuge at 4 °C for 7 minutes at 2,500 x g. Nuclei were

then resuspended in digestion buffer (NEBuffer 2 (50mM NaCl, 10mM Tris-HCl, 10mM $MgCl_2$, 1mM DTT, pH 7.9) for *Hind*III, *Mbo*I, and *Eco*RI, or *Dpn*II buffer (50mM Bis-Tris-HCl, 100mM NaCl, 10mM $MgCl_2$, 1mM DTT, pH 6.0 for *Dpn*II). SDS was added to a final concentration of 0.3% and the nuclei were incubated for 30 minutes - 1 hour at 37 °C with shaking at 900 RPM. Triton X-100 was then added to a final concentration of 2% and shaking continued for a further 30 minutes - 1 hour. Nuclear clumping was sometimes observed at this stage and was reduced by pipetting up and down to mix the samples. Restriction enzyme was then added (500 units *Hind*III and *Eco*RI, or 400 units *Mbo*I added over 2 - 3 time points) and cells were digested at least 16 hours.

The following day, restriction enzymes were inactivated by addition of SDS to 1.6% final concentration, and incubation at 65 °C for 25 minutes. Each 500 μL reaction was diluted into 8 mL of ligation buffer containing 1% Triton X-100. T4 DNA ligase was added in the amount of 10 Units / tube (Invitrogen) and reactions were incubated at 16 °C for four hours, followed by a further 30 minutes at room temperature. After ligation, 500 μg of proteinase K (Sigma) was added and tubes were incubated at 65 °C overnight. The following morning a further 500 μg of proteinase K was added and incubated for 2 hours at 50 °C.

DNA was purified using phenol-chloroform extraction and precipitated using 0.1 volumes of sodium acetate and 2.5 volumes of 100% ethanol.

HiC

HiC was performed as 3C with the following modifications. After > 16 hours of digestion with a restriction enzyme, but prior to inactivating that enzyme, 1.5 μL of 10mM dATP, dGTP, dTTP was added to each reaction, along with 37.5 μL of biotin-14-dCTP (Invitrogen). Ten units of Klenow (NEB) were then added and the samples incubated at room temperature for 45 minutes. During the ligation step, 40 - 50 units of T4 DNA ligase (Invitrogen) were used.

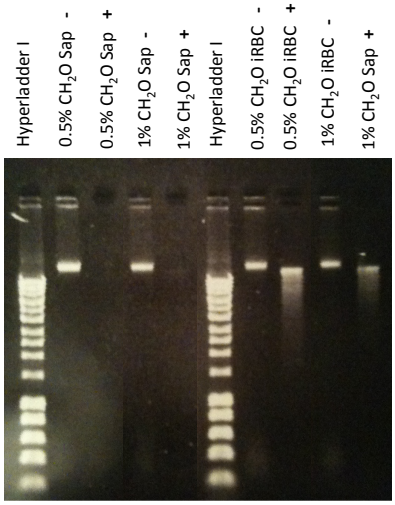

Figure A.3: Initial testing of methods to cross-link and digest chromatin showed poor recovery of DNA when the cross-linking was performed on saponin-lysed pellets. A representative gel is shown: genomic DNA is recoverable in all cases before the addition of enzyme (samples without enzyme marked by a minus sign). When cross-linking saponin-lysed pellets (sap), DNA is poorly recoverable; however, when cross-linking is performed on in-tact cells (iRBC), recovery of digested chromatin is consistent at different concentrations of formaldehyde. The suspected reason for this effect is that nuclei tended to clump excessively in the restriction buffer when cross-linking was performed on saponin-lysed parasites, leading to unreliable recovery during nuclei isolation and DNA extraction.

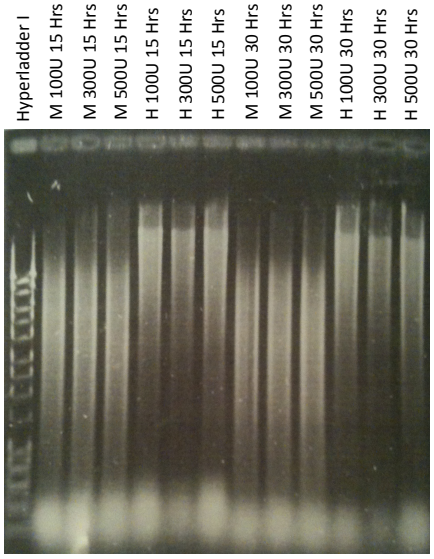

Figure A.4: Efficient and reliable digestion of cross-linked DNA using 1% formaldehyde. Increasing digestion time and concentration shows a modest but consistent increase in digestion efficiency. This is not visible on a gel but can be discerned using quantitative PCR (c.f. Figure 5.2).

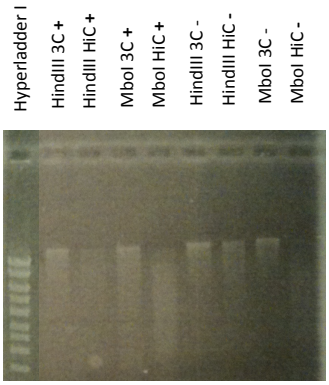

Figure A.5: 3C and HiC libraries in the IT clone generated using both *Hind*III and *Mbo*I. The + denotes BC6 selected lines while the - indicates an absence of selection. The improved ligation efficiency of the 3C samples (which use a sticky-end ligation) is apparent, as DNA migrates at higher molecular weight in the 3C samples relative to the HiC samples.

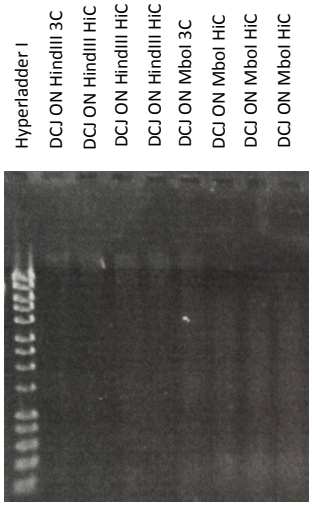

Figure A.6: 3C and HiC libraries in the DCJ ON clone generated using both *Hind*III and *Mbo*I. The 3C and HiC libraries depicted here behave similarly to those in Figure A.5, demonstrating the reproducibility of the procedure between experiments and across strains. The gel contains little DNA because only 1/20 of the final library was run to conserve sample.

Primers

Primers used to validate the 3C libraries are given in Table A.2. Primers used to estimate the digestion efficiency of cross-linking chromatin are in Table refT:3Cdigestionprimers. The *Mbo*I validation primers are in Table A.4 (Forward and Reverse pairs were used to estimate digestion efficiency).

Primer ID	**Primer Sequence**
3CMal12_1	ctgaagcgttcgattgtccg
3CMal12_2	ttggtgtggtgtgtggttca
3CMal12_3	ccaccttccttaccagagga
3CMal12_4	tggccacagttccaaatcca
3CMal12_5	tgtaacgttctacaaacttgcttct
3CMal12_6	agcactatagttacaagacattccat
3CMal12_7	accttatcagtcagctttccagt
3CMal12_8	tcagcggatgctgaagcaag
3CMal12_9	aatgcaccacccctaaagc

Table A.2: Primers used to assess 3C data on chromosome 12 using *Eco*RI as a restriction enzyme.

Primer ID	Primer Sequence
PFL0020w_F	ggtaaatgtaaccccacagaaacg
PFL0020w_R	tccacttttgtcctattttaaaacgc
PFL0075w_F	agcctcatgatgttttgttccac
PFL0075w_R	tggtgatggaactagagctttttatg
PFL1010c_F	ccttaattttgcttgctcttacatcac
PFL1010c_R	atttctcaaaataaaccggtaacaagg
PFL1490w_F	tctcaatccatggaacatgacc
PFL1490w_R	tcgtaggttgtgtttgttgcac

Table A.3: Primers used to assess the efficiency of chromatin digestion at *Eco*RI sites.

Primer ID	Primer Sequence
*Mbo*I_12_1L	tagccaccatgcaccatacg
*Mbo*I_12_1R	agcaggcttatcagttatattttgg
*Mbo*I_12_2L	tggccatgttaaggctgcac
*Mbo*I_12_3L	acctttaccacatgtattgattctac
*Mbo*I_12_4L	actttcaccacccataacataaaag
*Mbo*I_12_5L	tcctctttctttctttgaagttcg
*Mbo*I_12_6L	gttctactaacctattcattccttctg
*Mbo*I_12_7L	tcacctaatctgagacctccttc
*Mbo*I_12_8L	tttttctgctcatttggtgcattc
*Mbo*I_12_9L	tttgttatgcttactgctctgc
*Mbo*I_12_10L	agatgatgtttcacatccttcatttc
*Mbo*I_12_10R	ttagtgcagattctatgcaagtg
*Mbo*I_12_11L	tcaaacgttggaagaaaaatatttagg
*Mbo*I_12_12L	tggagctgttcagcttttcatc
*Mbo*I_12_13L	tcaacattaccacttattttgctttg
*Mbo*I_12_14L	tgcatgcccataccttgtacc
*Mbo*I_12_14R	gagcatatgcatctatgaccttaac

Table A.4: Primers used for 3C analysis using *Mbo*I and to assess the efficiency of chromatin digestion at *Mbo*I sites.

A.8 Fluorescence *In Situ* Hybridization (FISH)

FISH was performed as in reference [90]. Parasites were harvested and hemoglobin removed using 0.15% saponin at 4 °C. Parasites were washed twice in 1X PBS at 4 °C and then fixed overnight at 4 °C on a rocking platform. After fixation cells were washed twice with 1X PBS and stored up to two weeks at 4 °C in 1X PBS. Fixed cells were deposited on a microscope slide and allowed to adhere for 30 minutes in a humidified chamber at room temperature. Parasites were then incubated with 0.1% Triton X-100 for 5 minutes and washed twice with 1X PBS. Slides were then pre-incubated with 1X hybridization solution (2X SSPE, 50% deionized formamide (deionized and stored at −20 °C). Hybridization was performed overnight in 1X hybridization solution plus 0.5 - 2 μL probe (see below). After hybridization, slides were washed in 2X SSC / 50% formamide for 30 minutes at 37% and mounted with vectashield containing DAPI.

Probe generation Probes were generated from 300 ng of PCR amplified, gel-purified PCR product at least 2 KB in length. Probes were synthesized using the fluorescein or biotin High-Prime kit (Roche). Probe synthesis was performed overnight. Probes were precipitated using 2.5 μL of 5M LiCl and 2.5 volumes of 100% ethanol, washed once in 70% ethanol, resuspended in 40 μL deionized water, and stored at −20 °C.

A.9 High Throughput Sequencing

High throughput sequencing was performed at the Sanger Insitute in Hinxton, UK using Illumina Genome Analyzer 2 machines. Both ends of the DNA molecules from a paired-end read library were sequenced with a read length of 54 to 75 bases depending on the library. Library construction and sequencing were performed according to current protocols of the Pathogen Sequencing Unit.

A.10 Quantitative Real-Time Polymerase Chain Reaction

Quantitative real-time polymerase chain reaction (qRTPCR or qPCR) was performed on a RotorGene 6000 with a liquid handling robot and used to quantify cDNA and genomic DNA samples. Primers were designed using Primer3 [http://frodo.wi.mit.edu/primer3/], and incubated at a final concentration 100nM (unless otherwise specified) with SYBR mix (Quantace) diluted to a final concentration of 1X, along with template DNA at 100 ng or less. Cycling conditions were as follows: 58 °C for 30 seconds, 68 °C for 30 seconds, 95 °C for 20 seconds. Between 40 and 50 cycles were performed. Typical reaction volume was 10 μL. Absolute quantitation was performed using standard curves which were run in duplicate from serially diluted genomic DNA at known concentrations. Standard curves were run internal to each reaction where possible. Estimates of concentration were obtained by linear regression of genomic DNA concentration on the crossing threshold values, done either using Rotorgene Software or R. The first dilution of genomic DNA was frequently removed from the computation of reaction efficiency, as the high concentrations of EDTA used to store the DNA were found to inhibit the *Taq* polymerase and consequently overestimate reaction efficiency. When done in R, robust linear regression was used from the MASS library, to reflect the case that occasionally reactions fail entirely at low concentration. An example run giving raw fluorescence during cycling and melt is shown in Figure A.7.

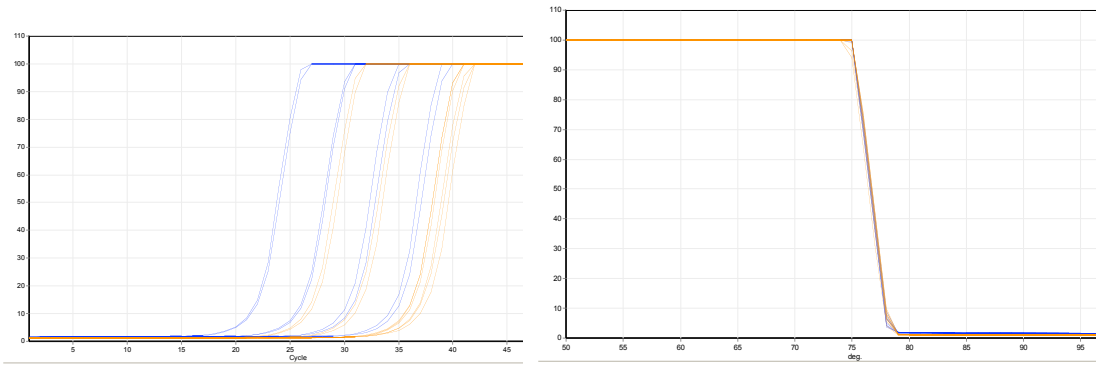

Figure A.7: Raw fluorescence vs. cycle number *(left panel)* is shown for an example PCR reaction containing standard curves (blue) and immunoprecipitated DNA fragments (orange). Melt curves *(right panel)* of fluorescence vs. temperature demonstrate that a single product is amplified in the reaction. Here, reactions have been run in duplicate.

Appendix B

Notation, Terminology, and Expression Spaces

B.1 Notation and Terminology

Our primary focus is on the biology of *P. falciparum*, but in analyzing large datasets we often invoke mathematical procedures or arguments. In this section, we describe the notation used. Most of the mathematical details have been confined to appendices and can be either be read or skipped without losing the main thread of the thesis.

A vector is a list of items, also called components, $(x_1, x_2, ..., x_n)$, along with rules for combining different lists. Addition of vectors is by addition of their components, and multiplication by a scalar c multiplies each component by c. For example, a list could be the x, y, and z (in this case, $n = 3$) coordinates of a point in space, or the expression values of n genes measured by a microarray experiment (in this case, $n \approx 5,000$). In general our lists contain a finite set of real numbers, which we say form real vector spaces of finite dimension. We denote n-dimensional vector space defined over real numbers as $\mathbf{R}^n$. Vectors are denoted as x or y; a vector from an indexed set is labeled as x_i. The same notation is occasionally used to denote the ith element of the vector x; when this occurs it will be made explicit in the adjacent text. We assume that the vector space comes with an inner product, usually defined as $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$ and which we typically write as $x^T x$. This inner product induces the norm $\|x\| = \sqrt{x^T x}$. Defined in this fashion, $\mathbf{R}^n$ is the n-dimensional generalization of 3-dimensional Euclidean space. Distances in such spaces measured by the Euclidean distance the formula $\text{dist}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$.

We sometimes work with the so-called p-norms, or ℓ_p norms, defined as $(\sum_i |x_i|^p)^{\frac{1}{p}}$ and denoted $\|x\|_p$. The ℓ_2 norm is the Euclidean norm defined above. The "counting" norm, $\|x\|_0$ is defined as the number of non-zero entries in the vector x. In particular, we find use for the ℓ_1 norm as a sparsity heuristic for the counting norm.

We often work in vector spaces whose elements, i.e. vectors, are not n-dimensional real vectors. In microarray experiments, where we have n experiments in which p genes are measured, we sometimes work in the space of n by p matrices. We are typically thinking in that context about several vectors in $\mathbf{R}^p$ measured and stacked by row in an n by p dimensional matrix.

In distance geometry and the reconstruction of chromosome folding patterns, we also work in matrix spaces. There, we use the Frobenius norm of a matrix X, $\|X\|_F$ by $\mathrm{Tr}(X^T X)$; Tr denotes the trace operation, or summation of the diagonal elements. The matrix, A, is called positive semidefinite if the inequality $zAz \geq 0$ holds for all vectors z; if the inequality is strict then A is called positive definite. The vector spaces of positive semidefinite and positive definite matrices are denoted $\mathbf{S}_+$ and $\mathbf{S}_{++}$ respectively. A convex cone is a set particularly useful in optimization problems. A set C is a convex cone if, for all $x, y \in C$, $\alpha x + \beta y \in C$ for $\alpha, \beta > 0$. Important convex cones include the non-negative orthant and the cone of positive semidefinite matrices. The notation $\succeq$ defines an inequality with respect to a cone. For a vector x, the non-negative orthant is assumed and $x \succeq 0$ means all elements $x_i > 0$. For a matrix, the positive semidefinite cone is assumed, and $A \succeq 0$ is used to denote that the matrix A is positive semidefinite, and $A \succeq B$ means $A - B \succeq 0$.

We define a number of convenience operations for manipulating matrices. The operator $\mathrm{diag}(X)$, when acting on a matrix $X \in \mathbf{R}^{n \times n}$, returns a vector of length n whose elements are the diagonal elements of X. When acting on a vector $x \in \mathbf{R}^n$, the operator returns a matrix whose diagonal elements are the elements of x. The vector whose elements are all equal to 1 is defined $\mathbf{1}$. Using that notation, the sum of the elements of a vector can be written $\sum_i x_i = \mathbf{1}^T x$, and the matrix formed by repeated columns of x is given by $x^T \mathbf{1}$.

When discussing results from polymer theory in the Appendix D, our conventions are slightly different as we follow conventions in that literature. In particular,

vectors typically refer to real points in a Euclidean 3−dimensional space and are denoted by bold typeface such as $\mathbf{x} \in \mathbf{R}^3$. In most other places in the thesis we would write that as $x \in \mathbf{R}^3$. Also, the brackets $\langle\ \rangle$ refer not to inner product but to an ensemble average, or what is known in probability as an expected value.

B.2 When $p \gg N$

A single microarray yields a high-dimensional vector of responses, $x \in \mathbf{R}^p$, where p is typically on the order of 6,000 for malaria. In a typical microarray experiment, we measure n conditions, where n is in the range $10 - 100$. Stacking these measurements together into an $n \times p$ data matrix, X, gives a compact representation of the information obtained from the experiment. In an experiment such as this one, X has more columns than rows or is known colloquially as a "fat" matrix.

This sort of arrangement is in fact fairly unusual in statistical applications. In a typical statistical experiment, such as (say) using regression to predict the lifespan of a subject based on blood pressure and weight, one would collect more a large number of subjects (n) and measure p covariates (in this case, p would equal 2) in these n individuals. Yet often in genomics applications, the reverse is true; p is larger than n, and frequently, as in the case of microarrays, much larger. For example, with microarrays, there are typically thousands of variables and tens of observations for each variables. To see how this is a problem, consider the case of trying to predict hours post invasion (HPI) by fitting a "vanilla" linear regression. In this application we model y as a linear function of X with coefficients β such that

$$y = X\beta + \epsilon.$$

Equivalent to assuming an independent and identically distributed Gaussian distribution for the error terms, ϵ, we choose β so as to minimize the Euclidean norm

of the residual, $\|y - X\beta\|_2$. This is a quadratic function whose minimum can be found by setting the derivatives with respect to β equal to zero, a procedure which leads to the well-known "normal equations":

$$X^T y = X^T X \beta.$$

Assuming that $X^T X$ is invertible (equivalently, that X has full column rank), then the least-squares estimate,

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

But in this case X does not have full column rank, and the procedure fails. Equivalently, imagine someone asking you to build a statistical model to predict lifespan based on blood pressure and weight, and using as input data the single data point of a person who is 200 pounds, has blood pressure $\frac{130}{90}$, and lived to 78 years of age. One needs more data.

Perhaps the problem lies with linear regression and we can try fitting a different type of model. Instead of estimating a continuous time parameter as one might do using regression, we instead assume that an observed sample comes from one of the 48 time points, and try to choose the best one. We can choose to directly model the distribution that generated the samples. Using Gaussian Discriminant Analysis, we can fit a classifier in which we assume that there are k classes of samples (in this case $k = 48$). Each class has a probability density which is Gaussian with mean, μ_k and covariance matrix, Σ, which we assume might assume to shared among the classes in order to simplify the problem:

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}.$$

The parameters which must be fit in such a model are the class means and the

elements of the covariance matrix. This model is closely related to the one we ultimately use for stage estimation, described in section C.3.1. In attempting to fit this model, we are faced with the similar problem: Computing the class density requires inversion of a covariance matrix $\Sigma \in \mathbf{R}^{p \times p}$, whose parameters we estimate from the data. We can estimate the elements of Σ using the sample covariance between variables, but in that case Σ will be singular. It is impossible to model class density as a multivariate Gaussian since do not have enough data to estimate a non-singular covariance matrix. We must either acquire more data, make assumptions about the structure of the covariance matrix, or use a different method.

In fact, it gets worse. If X is an $n \times p$ matrix collected with real data, we would expect that it is either full row-rank or full-column rank. However, as we show in figure B.1, explained further in section2.5, in malaria gene expression the variables are frequently correlated, and the asexual development transcriptome is approximately rank k, where $2 < k < 10$, but is certainly less than n. The strongly correlated structure of A means that not only do we not have enough data to estimate all the elements of $\beta \in \mathbf{R}^p$ where $p \approx 6000$, we don't really have enough data to estimate more a coefficient vector of much more than length $2-10$ without additional constraints on degrees of freedom.

This is a general problem of working in high-dimensional places, of which microarrays are an archetypal example. In large part because of applications in genomics and other areas of biomedical such as radiological imaging and functional MRI, progress in this area has been swift and led to a number of possibly unexpected outcomes with far-reaching applications such as "compressed sensing" in MRI. A number of innovations proposed that are based in large measure on methods used to control effective degrees of freedom, such as regularization of the ℓ_2 norm ("ridge regression"), the ℓ_1 norm (the "lasso"), or related methods such as the Dantzig Selector and the Elastic Net. These methods succeed in part be-

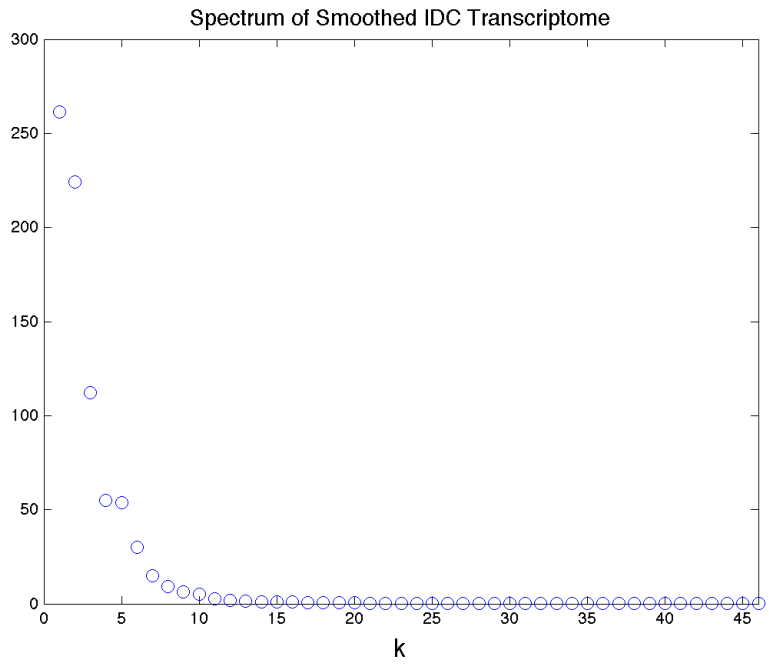

Figure B.1: The spectrum of A showing that its approximate rank is between 2 and 10. The values plotted are elements of diag(Σ), the singular values of A, the smoothed IDC transcriptome matrix measured in [9]. The approximate rank is the number of non-zero elements of diag(Σ), explained further in section 2.5. A biological interpretation of rank would be the number of uncorrelated, discernable regulatory patterns that are present in the expression set. For example, the fact that the matrix can be well approximated by another matrix of rank 3 suggests that there are approximately 3 uncorrelated regulatory transcriptional "modes". That does not mean that there are only three relevant transcription factors, because the action of transcription factors is likely to be correlated; however, it does suggest that regulation of most genes is not as complex as might have initially been thought, since it can be mathematically approximated by addition and subtraction of small number of overall patterns.

cause while the problem of finding the best subset of parameters is a non-convex, hard combinatorial one, good heuristics based on norm minimization exist and can often be readily solved by convex programming techniques.

The overarching point – which is a general one we hope to make in this thesis – is that biology has moved into the $p \gg n$ domain, a territory in which it is likely to reside for some time. There are fundamental limitations, imposed by the geometry of these spaces, on the types of models we can fit. Consequently, time spent studying the geometry of these spaces makes applying and understanding data analysis in biology easier and clearer.

B.3 Singular Value Decomposition and Principal Component Analysis

A basic result of linear algebra is that every real matrix $X \in \mathbf{R}^{n \times p}$ can be factored into the form

$$X = U\Sigma V^T.$$

This decomposition, in which X has n rows and p columns, U is $n \times p$ with orthogonal columns, Σ is an $n \times n$ diagonal matrix, and V is $p \times p$ with orthogonal columns, is unique for every matrix and is known as the singular value decomposition or SVD. The SVD is closely related to principal components analysis (PCA), in the following way. If the columns of X have been mean-centered, then the factorization can be written as

$$X = (U\Sigma)V^T$$
$$X = SV^T.$$

In this case the columns of V are known as the principal component coefficients (or simply the principal components), and the columns of S are termed the scores. The SVD has a number of interpretations:

1. The number of non zero elements in diag(Σ) is the rank of X. The rank of a matrix is a measure of the dimension of the space defined by linear combination of its rows and columns.

2. As the best rank-constrained approximation for a matrix with respect to minimizing the Frobenius of the residual matrix: $\hat{X} \approx U\tilde{\Sigma}V^T$ to a matrix X where $\tilde{\Sigma}$ contains the first k singular values in the diagonal of Σ. Choosing $\tilde{\Sigma}$ in this fashion minimizes the quantity $\|X - \hat{X}\|_F$ over all possible $\hat{X}$ have rank k. It is actually perhaps surprising that a closed-form solution is available for this rank-constrained optimization problem. Equality constraints on rank often introduce non-convexity into the feasible set and are notoriously difficult to handle (c.f. section E.1).

3. The columns of U and the rows of V^T form an optimal basis for the columns and rows of X. This basis is optimal in the sense that the orthogonal directions of the basis vectors give the direction of maximum variance in the data matrix.

4. The elements of diag(Σ), also called the spectrum of singular values, gives an "approximate" or functional rank of the data matrix. Because X is observed with noise, it is typically full rank (i.e. rank$(X) = min\{n, p\}$), but for functional purposes X may lie in an affine set of much lower dimension.

5. The columns of U and V are found by eigendecomposition of the matrices XX^T and X^TX, which are positive semidefinite matrices and therefore have real eigenvalues. In microarray applications, the columns of U are sometimes

called the "eigengenes" and the columns of V are sometimes called the "eigenassays" or "eigenarrays" [1, 2].

6. The eigenvalues corresponding to the directions in U correspond to the percent of variance "explained" by each component.

Appendix C

Stage Estimation Theory and Algorithms

C.1 Abstract

In Section 2.3 we introduced the need for a method to account for the temporal development of the culture but deferred the problem itself because of its complexity. We take it up here. The general outline is as follows: we first describe the problem and give a formal definition. We then solve the problem by several related approaches. We show the intimate relationship between approaches and also how they are constrained by the underlying geometry of the $p \gg n$ scenario of section B.2. Mathematical details are confined to this section and the reader more interested in the results as applied to *P. falciparum* biology can skim this section or skip to chapter 3.

C.2 Background and Problem Definition

The malaria intra-erythrocytic development cycle transcriptome is a 48-hour cycle that repeats in the blood of the human host during malaria infection. The transcriptome comprises the set of all full-length mRNA molecules that are synthesized from DNA templates during asexual development. Because there is an underlying time parameter we can think of the IDC as a continuous map $f : \mathbf{R} \to \mathbf{R}^p$. The biology of *P. falciparum* dictates that f is periodic: $f(x + T) = f(x)$ where T is $\approx$ 48 hours. An "expression profile" as it is here called is a sample of f, i.e., a measurement p genes. In a typical time course experiment, f is sampled n times (typically at uniform intervals) and these samples are used to form an estimate, $\hat{f}$, typically by imposing smoothness constraints on f. Stage estimation is a request for the inverse mapping, f^{-1}, that is, given a sample in the p-dimensional space of mRNA measurements, what was the time value that generated that sample.

The problem is made difficult for several reasons:

1. we don't know that f is invertible (since the function is periodic, we can

only expect that in the best case it is invertible within a period)

2. we observe a noisy realization of the p-dimensional vector

3. we do not know the structure of this noise, precisely we know neither the distributional form of the noise nor that the joint noise distribution can be factored as a product of independent noise terms

4. we don't have access to true underlying function f, only an an approximation $\hat{f}$, and

5. the observed expression profiles are contaminated by hidden sources of variation such as differences in synchronicity between cultures and the presence of sexual stage parasites in the population.

It should be clear from the complexity of the problem that we will have to make simplifying assumptions in order to solve it. There are several simplifying assumptions that can be made; most focus on the distribution of the noise and the absense of hidden variation.

Datasets The methods we describe are designed to deal with the stage estimation problem in the abstract and are meant to be independent of the method used to gather the expression profiles, whether they come from microarrays, RNA sequencing, qRT-PCR, or some other method. Nevertheless, a major issue in stage estimation is the form of the noise distribution and we cannot be sure that performance will be the same on all datasets. While the methods were developed for microarrays, in practice we have found that they perform well on expression datasets from RNA sequencing as well and would likely work for qRT-PCR.

Several malaria expression profiling datasets are available in the literature or in our lab. Datasets we have tested include the asexual expression profiles generated using two-color glass slide arrays [9, 10, 26, 87], single-color affymetrix

arrays [21, 27, 28, 85, 122], RNAseq [110], and sexual stages with both Affymetrix-type [152] and two-color arrays [135]. These datasets provide a rich collection of test and reference sets on which to test stage estimation algorithms.

C.3 Methods for Estimating Stage

C.3.1 Maximum likelihood estimation

In section B.2, we discussed a classifier that models the within-class probability density directly as an example of the constraints imposed by the situation $p \gg n$. Here we use a version of that classifier in which the within-class covariance matrix is assumed to diagonal and the same for all timepoints. This was first suggested by Chris Holmes and he supervised its implementation, which was done by Avi Feller and me and published in [85] as well as in Feller's master's thesis [42]. Our goal here is to give an overview of the method, to extend it, and add to the discussion in [42] in a non-overlapping way.

We want to specify $P(y|t_i)$, the probability of observing a particular set of reference observations when a sample is at given timepoint. We now make two strong assumptions: The first is that the data are generated by a multivariate Gaussian. In this case we can write down a form for the probability of a given observation,

$$P(y|t_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}. \tag{C.1}$$

If we make a further assumption that the error distributions among the genes are independent. If that is the case, in general the likelihood factors to the form

$$L(t_i) = \prod_j P(y_j|t_i), \tag{C.2}$$

which in our case is

$$p(y|t_i) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{(y_i - \mu_i(t))^2}{\sigma_i^2}}. \tag{C.3}$$

A graphical description for a single gene, i.e. one of the factors in the likelihood, is given in figure C.1.

The estimate of sample age is then the the value of t_i which maximizes that expression:

$$\arg\max_{t \in [0,48]} \prod_i p(y|t_i). \tag{C.4}$$

Implicit in this description is that there is some "true" parameter t_i which generated a noisy realization of the sample. That is the philosophical justification for maximum likelihood estimation.

The two technical assumptions above concern the nature of the noise which corrupts the measurements; the assumption os Gaussian errors is principally one of convenience though it is a frequent assumption in statistics and theoretical support is provided by the central limit theorem. The conditional independence of the error terms given $t = t_i$ is reasonable from a technical perspective though it is unequivocally false in a biological context because genes are organized into biological pathways and regulated by common transcription factors. In fact the success of the method convinces us that the violations of this assumption are not so severe, and that there is limited differential expression in vivo. The assumption of conditional independence given $t = t_i$ may also fail is there are unmeasured variables in our sample which are different from those in the reference, such as differences in synchrony or gametocytes. This is in fact the case and we correct this problem to account for hidden gametocytemia in samples. The assumption then becomes conditional indpependence given $t = t_i, \alpha = \alpha_i$.

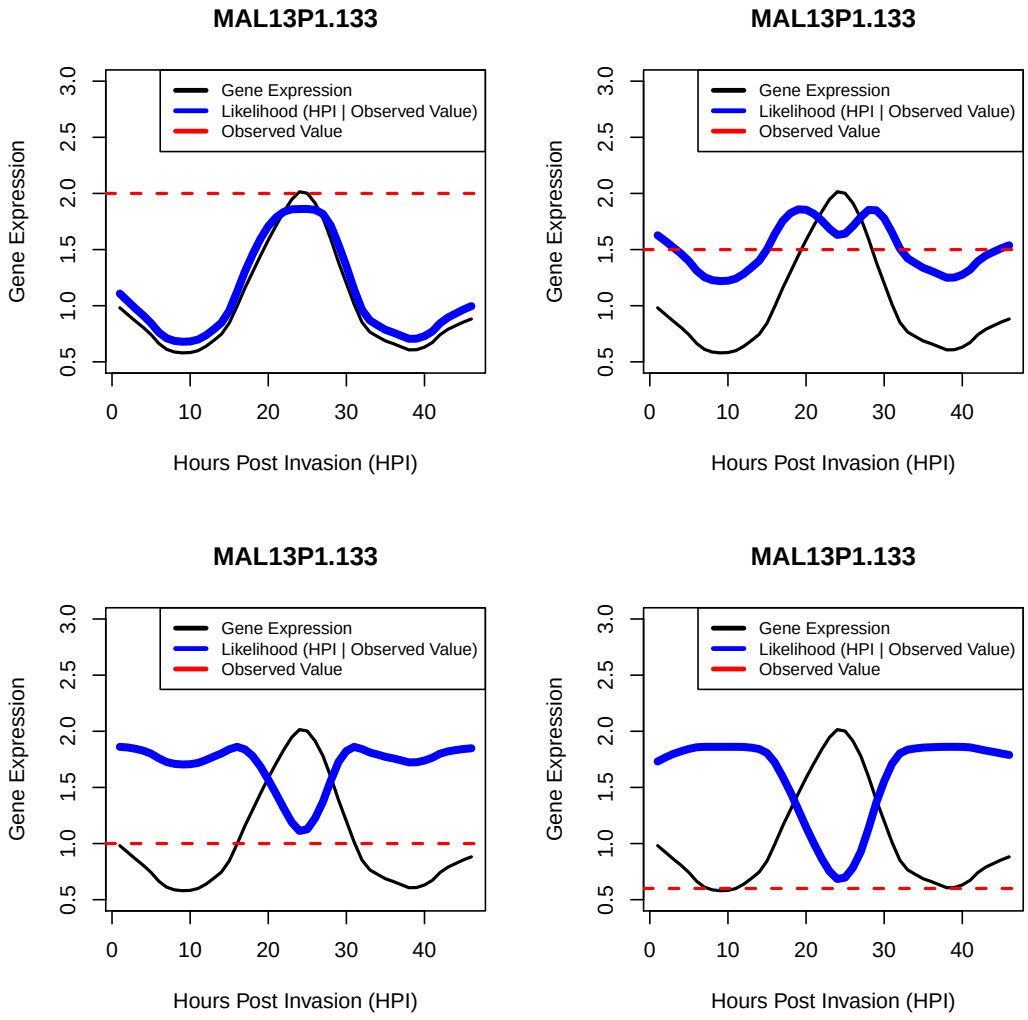

Figure C.1: Stage estimation works by obtaining likelihood curves describing the probability of obtaining the observed value at each cycle along the reference set, given an underlying time-evolving reference gene expression function. This is depicted graphically for four different observed values above. The black line gives gene expression through the asexual lifecycle of MAL13P1.133 as measured in the HB3 clone in ref [9]. In each panel, a dashed red line depicts a measured gene expression value, and the solid blue line gives the corresponding likelihood density. To obtain an estimate for the entire sample, the product of individual likelihood curves for all genes in the reference set is computed.

The sample likelihood function under these assumptions derives from a single Gaussian probability distribution, with mean zero, for each gene. We must estimate the variance of the distribution for each of the genes. We do this by calculating the sample variance of the distribution of residuals from the loess fits as well as by calculating the variance in expression at a given timepoint between the three reference clones from reference [87]. Since the variances of the sum of independent random variables is equal to the sum of the variances, we add the sample variances together to come up with a gene-specific variance, σ_j^2, which parameterizes the likelihood function. The model can expressed compactly as a multivariate normal with a unique mean vector, μ_i, for each class C_i and $\Sigma = \text{diag}(\sigma_1, ..., \sigma_j)$:

$$y = \mathcal{N}(\mu_i, \Sigma). \tag{C.5}$$

We showed in [85] and demonstrate below that this approach yields estimates for sample temporal progression which are accurate and robust. We have tested and confirmed that estimates derived by this procedure are stable when measurement device is a glass-slide array, an Agilent-type photolithographic array, or a result of a direct sequencing of cDNA molecules. We have also confirmed that the estimates are stable due to changes in a subset of genes.

C.3.2 Geometry

The maximum likelihood estimate above has a straightforward geometric interpretation. For a given value of σ, the term in the exponent of the likelihood function above is equivalent to a squared ℓ_2 distance from the curve $\mu(t)$. While in general $\mu(t)$ is a curve in p dimensions where p is on the order of thousands, which cannot be visualized, we can take the case of $p = 3$ and the operation of the algorithm is clear. Figure C.2 shows the expression profiles of 3 genes from the IDC reference set, which together define a curve in $\mathbf{R}^3$, shown in the middle

panel. A test expression profile, $y \in \mathbf{R}^3$ defines a point in this space and is shown in green. The test profile y is to be classified and we do so by calculating the distance between y and all points on $\mu(t)$, and then assigning it to the closest timepoint,

$$\arg\min_t \ \mathrm{d}(y, \mu(t)). \tag{C.6}$$

The solution to this optimization problem is in general not convex except for very special types curves, but it can be easily obtained by evaluating $\mathrm{d}(y, \mu(t))$ along a discretization of points along the t coordinate. Thinking of stage estimation in this way gives an intuitive description of what the algorithm does, and it also clarifies the relationship of our approach to that of PlasmoDB, which uses correlation to measure the relatedness of expression samples. It also leads to generalizations since any distance metric can be used.

ℓ_p distance. The usual "Euclidean" or ℓ_2 distance function

$$(\sum_j (x_j - y_j)^2)^{\frac{1}{2}}, \tag{C.7}$$

which corresponds to a Gaussian probability density with circular contours, is known to lack of robustness, and the ℓ_1 distance function

$$\sum_j |x_j - y_j|, \tag{C.8}$$

sometimes called the "Manhattan" distance or "taxicab" distance, can be a better choice if a robust algorithm is desired. The ℓ_1 distance function is much less sensitive to outliers, which is demonstrated in figure C.3. A single entry in the observed vector y was set to the extremely large value of 500. The ℓ_1 distance between the test vector y and the reference vectors x_i remains increased but largely

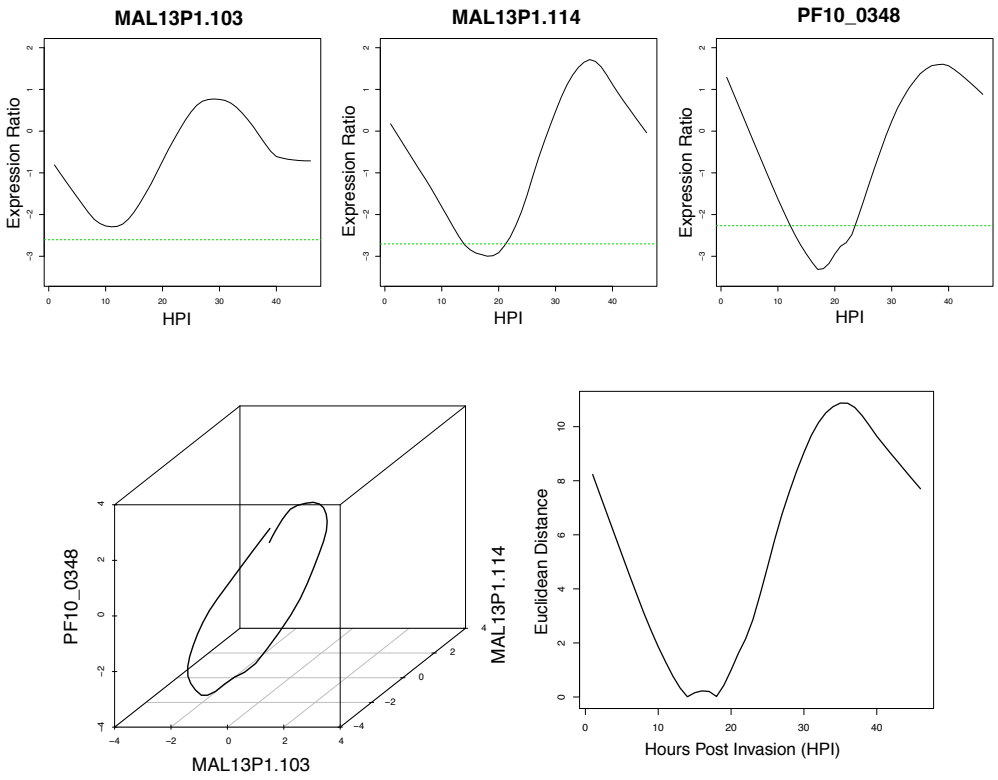

Figure C.2: Stage estimation is distance minimization to a curve in $\mathbf{R}^p$, the p-dimensional space of real numbers. In this example, $p = 3$. In the top panel, individual components of $\mu(t) : \mathbf{R} \to \mathbf{R}^p$ are shown as functions of a single variable, $\mu(t) : \mathbf{R} \to \mathbf{R}$. Individually the curves represent a one-dimensional, circular manifold which is embedded in a larger space, in this case $\mathbf{R}^p$ where $p = 3$ (bottom left panel). The ℓ_2 distance of a test point (green) from this manifold is shown in the bottom right panel.

unchanged. The Euclidean or ℓ_2 distance on the other hand is severely affected. As the glass-slide array data in particular are known to be noisy, robustness in stage estimation is important. The residuals between observed and loess smoothed values for the 3D7 reference set from [87] are shown in figure C.4.

Least Angle. A related approach is to use the Pearson correlation as a proxy for (inverse) distance. This is the approach taken by PlasmoDB, i.e. to find timepoints of maximum correlation with the test sample. The Pearson correlation coefficient for two vectors x and y is defined as

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{C.9}$$

which if the vectors x and y have mean zero, can be written more compactly as

$$\frac{x^T y}{\|x\|\|y\|}, \tag{C.10}$$

which is defined to be the angle, θ between two vectors,

$$cos(\theta) = \frac{x^T y}{\|x\|\|y\|}, \tag{C.11}$$

and suggests the relationship between correlation and Euclidean distance. For vectors of length r (in general the relationship depends on the length of both vectors but we restrict ourselves for the moment to vectors of length r), Euclidean distance squared is $\|x - y\|^2 = 2r^2 - 2x^T y$, while the correlation is $\frac{x^T y}{r^2}$. We then have

$$\rho(x, y) = 1 - \frac{d(x, y)^2}{2r}, \tag{C.12}$$

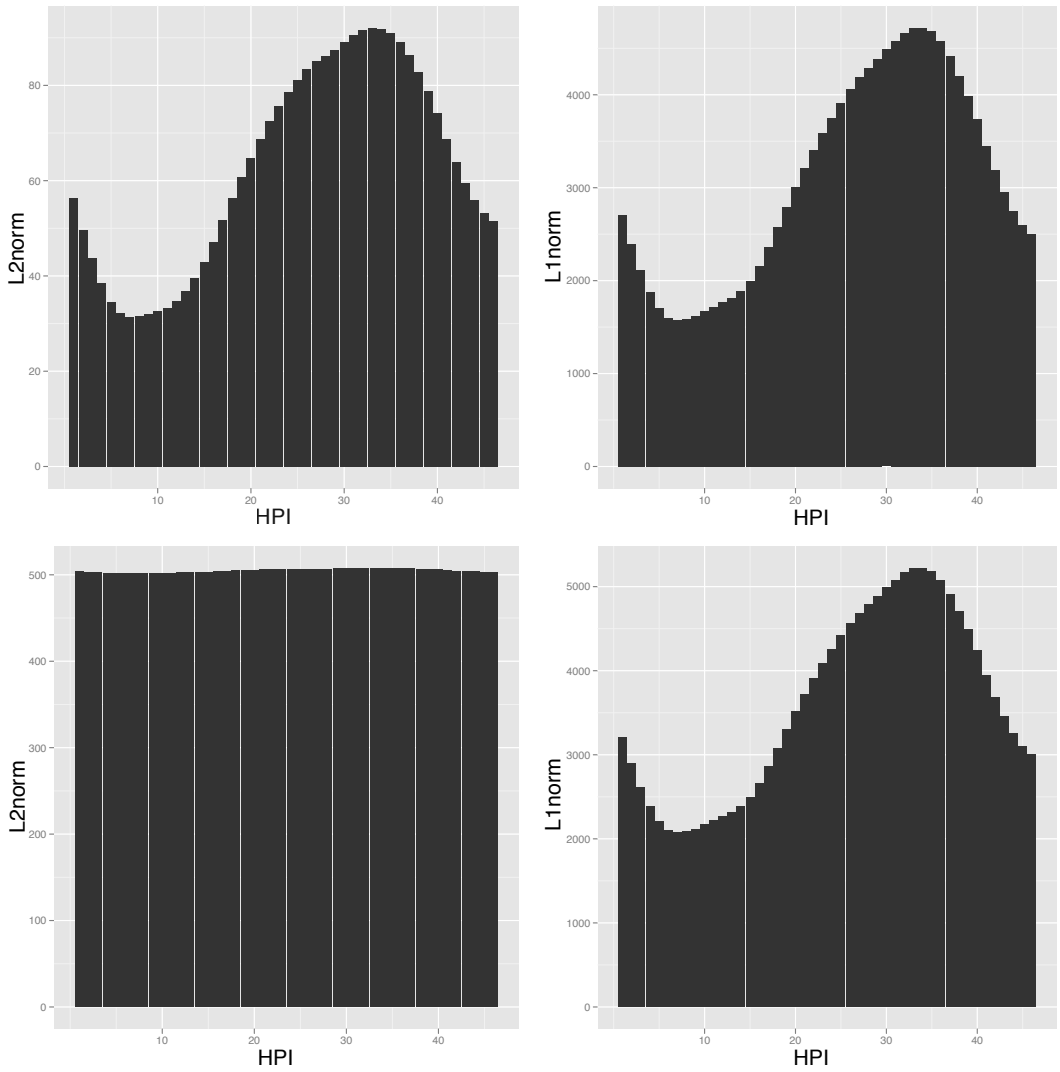

Figure C.3: Distance to cluster center, Euclidean (left) and Manhattan (right) distance metrics between sample 12 in the Dd2 set from reference [87] vs. the Hb3 reference set from [9].Distance to cluster center, Euclidean (left) and Manhattan (right) distance metrics between sample 12 in the 3D7 set from reference [87] vs. the Hb3 reference set from [9] with a single gene measurement set at the extreme value 500. The Manhattan distrance metric shows increased robustness when the error distribution can take on extreme values.

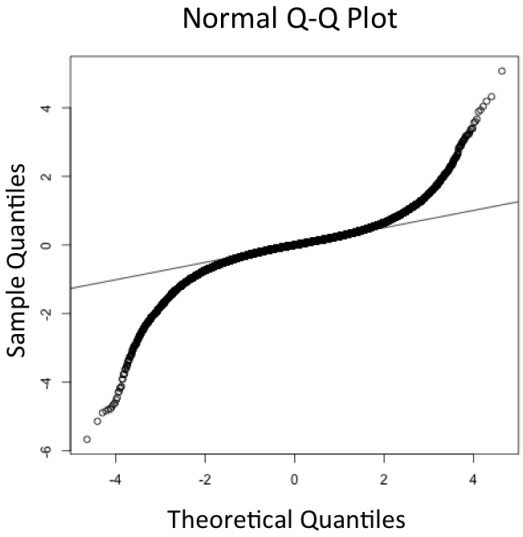

Figure C.4: Plot showing the quantiles of the distribution of residuals for the 3D7 reference series from [87] vs. theoretical quantiles drawn from the normal distribution. Strong deviations from normality are observed in the tails, underscoring the need for robust estimates.

an inverse relationship. For vectors of fixed r, correlation will be maximized when distance is minimized.

A condition (definition) of most normalization algorithms is that $\|x\| = \|y\|$. Therefore, the approach of PlasmoDB for aligning time points can also be written as a minimization problem:

$$\begin{aligned} \underset{y \in \mathbf{R}^p}{\text{minimize}} \quad & \theta \\ \text{subject to} \quad & \theta = \arccos\left(\frac{x^T y}{\|x\|\|y\|}\right) \end{aligned} \tag{C.13}$$

or what might be called the "least-angle" solution for vectors in $\mathbf{R}^p$.

Mahalonobis distance. The maximum likelihood approach implemented (section C.3.1) corresponds to minimizing a diagonal version of Mahalonobis's distance, which we explain here. Some genes may be biologically more variable than others, and groups of genes may covary together; in that case two points which

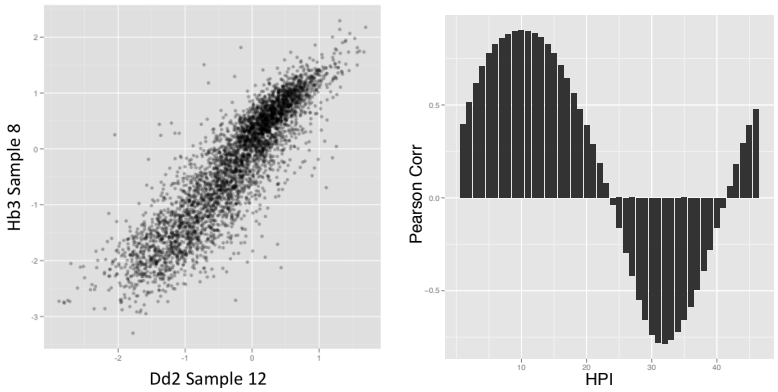

Figure C.5: Scatterplot of 3 hours post invasion samples from [9] vs. sample 12 from the 3D7 series [87] (left). The transparency of the points has been reduced to prevent overplotting. A Pearson correlation coefficient, r, can be calculated for pairs of vectors; in this case $r = 0.88$. The Pearson correlation coefficient can be used as a measure of similarity between samples as in the approach outlined above (right).

are close together in the natural time coordinate might end up, after the addition of biological and experimental noise, far away from each other along certain axes in p-dimensional Euclidean expression space. In geometric terms, the biological and experimental noise introduces anisometries into the expression space. Ideally we would like our measure of distance to account for this possibility. The distance measure introduced by Mahalanobis [94],

$$\mathrm{d}_{M}(x, y) = (x - y)^{T} \Sigma^{-1} (x - y), \tag{C.14}$$

is a natural metric which accounts for these anisometries with Σ the covariance matrix between x and y. As in C.2, we are interesting in the distance between a test sample y and a periodic curve $\mu(t) : \mathbf{R} \to \mathbf{R}^{p}$, which means we want to solve:

$$\arg\min_{t} \mathrm{d}_{M}(y, \mu(t)), \tag{C.15}$$

where d_{M} is defined as above and we assuming that Σ does not vary with t, i.e. $\Sigma(t) = \Sigma$.

We again run into the $p \gg n$ problem of section B.2, because we need an enormous amount of data to estimate a non-singular Σ. The Mahalonobis distance is the metric used in the exponent of a multivariate Gaussian, so we are not established anything new except changing our persective slightly. From this point of view, the constraint introduced in the maximum likelihood section that Σ to be diagonal, also known as the conditional independence assumption or the naive-Bayes assumption, has a geometric interpretation as correcting for anisometries in the expression space. For diagonal Σ, the metric can be written componentwise,

$$d(x, \mu(t)) = \sqrt{\sum_i \frac{(x_i - \mu_i(t))^2}{\sigma_i^2}}, \tag{C.16}$$

which is a Euclidean distance weighted by the variance of points along each axis. In other words, we can only correct for scalings in the space along the direction of each axis; the singularity of Σ implies that we do not have enough data to estimate the anisometries in arbitrary directions.

It is worthwhile to note that while the assumption of zero gene-gene covariance is demonstrably false, the approach does work quite well in practice. The implication is that while gene-gene covariances may exist, they are not substantial. In other words even if Σ is not truly diagonal, it can be well approximated by a diagonal matrix. Nevertheless, when information about the regulatory connectedness of genes becomes available, we may be able to relax the dubious assumption of a diagonal covariance and build better stage estimation algorithms. For the moment, though, the algorithms work well in practice.

C.3.3 Incorporating Microscopy Data

In implementing the maximum likelihood approach we have explicitly modeled the probability of a test vector conditional on a given timepoint, $P(y|t)$. What we actually want to know, however, is $P(t|y)$–the probability of the sample belonging

to each time given the observed expression profiles. These two quantities are related by Bayes theorem:

$$P(t|y) = \frac{P(y|t)P(t)}{P(y)}. \tag{C.17}$$

The term $P(y|t)$ is known as the likelihood function of t and is the function that we maximized in section C.3.1. The term $P(t)$ specificies the probability of observing a sample from each class. It is known as the prior distribution. What this equation basically says is that the probability of a given timepoint is proportional to the $P(y|t)$ multiplied by what we know *a priori* about the probability of timepoints. This is important for two reasons.

The first reason is that we almost always do know something *a priori* about which timepoints are more likely than other, since the culture of parasites is almost always examined under a light microscope using a Giemsa stain. We should not discard this information in stage estimation, and can in fact use it to improve the accuracy of our algorithms. Bayes' theorem tells us how. In practice this does not change our estimates much. The priors that we construct from microscopy data (figure C.6) tend to look a lot like the likelihood functions we calculate from expression profiles and thus the two do not substantially influence each other. Nevertheless, permitting the probabilities of indvidual timepoint to reflect what we know from microscopy does lend elegance and completeness to the stage estimation.

The second reason, on the other hand, matters a great deal and concerns generalization of the method. The approach described for estimating stage can be applied to almost any case of expression data which fall along a continuum or into different discrete classes from which a continuous parameter learned or class assignment made. For example, we might want to classify cancers according to prototypical expression profiles, or we might want to estimate the severity of a

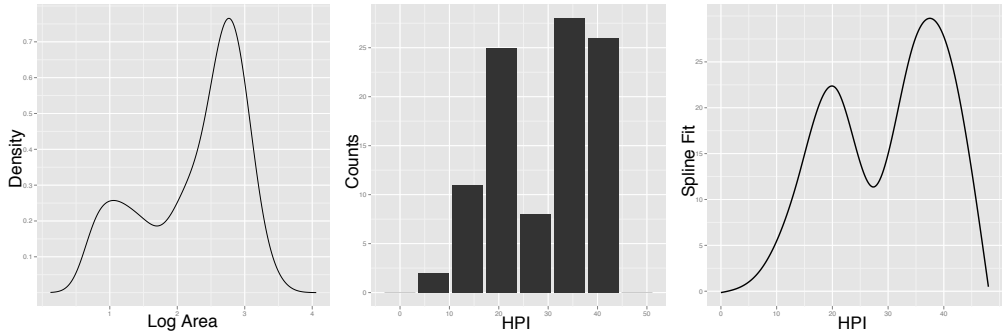

Figure C.6: Two prior distributions for parasite temporal progression. The prior on the left is obtained by measuring the area of 500 parasites on the Giemsa smear. For stages older than 20 hours post invasion, the logarithm of parasite area was shown to relate linearly to parasite age. Therefore for cultures with mature stages only, the distribution of log(area) can serve as a prior predictive density. The prior on the right is obtained in a three step process: 1) By counting the number of rings, early trophozoites, late trophozoites, early schizonts, and segmenting schizonts, 2) Mapping the preceding list to 7, 14, 21, 28, 35, 42 hours post invasion (HPI), respectively, and 3) fitting a smoothing spline to interpolate the count data.

person's diabetes from measured expression profiles of certain tissues. In cases such as these, some values are going to be much more likely than others, and the performance of the algorithm will suffer greatly if the *a priori* probabilities are not accounted for.

C.3.4 Mixtures of Timepoints

In the above sections we have considered the observed vector y to be a noisy realization of one of the rows of the the timecourse matrix A. In general, however, the sample could be a weighted sum of multiple timepoints. It is possible to estimate these weights directly, which we do in this section. Although in some cases this works well, we are introducing greater freedom (precisely, increasing the degrees of freedom) into the problem solution and in general the solution is not as stable as estimating a single timepoint for the sample. Therefore solving for the weights directly cannot be recommended as a front-line method for stage estimation; however, these methods are interesting for theoretical reasons and lead

to informative generalizations.

We introduce some notation to make this more clear. Let A^i be the ith column of the matrix A. Then the equation $y = Ax$ can be written in a fashion which makes explicit the equation as a sum of column vectors:

$$y = \sum_i A^i x_i. \tag{C.18}$$

This is the equation we wish to solve, in particular, we seek to find the correct "amounts" of the A^i to add to the solution. To do this we must specify the individual elements of x, i.e., the relative amounts of the columns of A to be incluced in the solution. In general since we have more equations, and there is noise, we have an overdetermined set of equations and no solution will exist. We therefore ask for an approximate solution such that $Ax \approx y$ or $\|Ax - y\|$ is small.

We typically choose the x that minimizes the norm of the residual. When the norm is Euclidean there is a closed-form solution known as the Pseudoinverse. If A has the singular value decomposition $A = U\Sigma V^T$ then the pseudoinverse, $A^\dagger$ is defined as

$$A^\dagger = V\Sigma^{-1}U^T, \tag{C.19}$$

and the minimization problem

$$\text{minimize} \quad \|Ax - y\|_2 \tag{C.20}$$

is solved by multiplying the test vector y by the pseudoinverse, $\hat{x} = A^\dagger y$. The elements of x contain the relative weights of each timepoint contained in the test vector. We note that this algorithm has the advantage of being–by far–the easiest to implement. The matlab operator \ calculates the pseudoinverse, so the algorithm is implemented in Matlab as with three keystrokes as `A \ y` where `A` is

the matrix of reference samples and `y` the reference sample.

Aside from ease of implementation, this algorithm has few practical advantages and performs poorly because the coefficients (i.e. the relative weights) can be negative and are permitted to become arbitrary large. This can be made slightly more realistic by requiring the coefficients of x to be non-negative and sum to one. The set of points that obey these restrictions is a convex set known as the unit simplex, $C = \{x | \sum_i x_i = 1, x \succeq 0\}$. In that case we must solve a constrained, convex minimization problem:

$$\begin{aligned} \text{minimize} \quad & \|Ax - y\|_2 \\ \text{subject to} \quad & x \succeq 0, \\ & \textstyle\sum_j x_j = 1 \end{aligned} \tag{C.21}$$

Alternative formulations are possible. As in the case above, a different objective function such as the ℓ_1 norm could be used. Furthermore, adding terms such as $\lambda\|x\|$ (regularization terms) or $\sum_j x_j - x_{j+1}$ (smoothness penalties) can prevent overfitting and encourage more realistic estimates while maintaining convexity of the optimization problem. For example we also propose an alternative formulation of the problem which attempts to correct some of the problems above:

$$\begin{aligned} \text{minimize} \quad & \|Ax - y\|_1 + \lambda\|x\|_1 + \kappa \textstyle\sum_j^{n-1}(x_{j+1} - x_j) \\ \text{subject to} \quad & x \succeq 0, \end{aligned} \tag{C.22}$$

A particular advantage of this approach is that it relaxes the obviously unrealistic assumption that a parasite population comes from a highly synchronous single timepoint. From this perspective it is in effect a way to non-parametrically

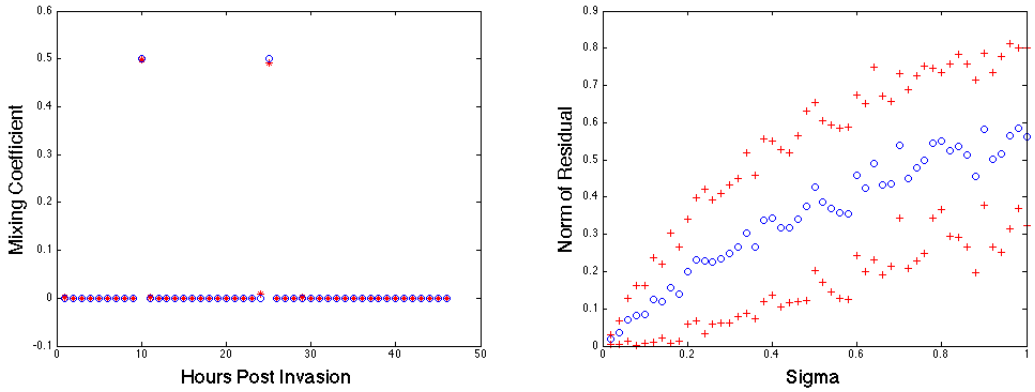

Figure C.7: **Reconstruction by convex optimization.** An alternative approach to stage estimation is to assume that the observed expression profile is a weighted sum of the underlying expression profiles. This contracts to the maximum likelihood approach described earlier (Section C.3.1), which considers the observed expression the noisy realization of a single reference expression profile. In this figure, we establish computational mixtures of timepoints and solve for the relative proportions by convex optimization. The left panel shows the true weighting coefficients as blue circles, and the estimated coefficients as stars. The algorithm is remarkably good at identifying the true weights in this example, as can be shown by the close fit of the red stars inside the blue circles in the left panel. Unfortunately, this approach does not scale well with noise or mixtures of timepoints that are closely related. The plot at right shows how the error increases rapidly as the noise added to the computational mixtures increases.

estimate the synchrony of the culture. Because the reference timepoints themselves composed of a (tightly synchronized) distribution of parasites and are consequently non-negative linear combinations of individual parasite transcriptomes, it will be of greater applicability when the IDC transcriptome of a single parasite is available. Amazingly this seems not unlikely in the near future.

Unfortunately, the lack of complexity of the malaria IDC cycle is actually a hindrance in this setting. The ostensibly 46-dimensional space is well approximated by a lower dimensional space (section 2.5). The matrix is 'ill-conditioned' and the solutions for all 46 x coefficients are unstable. The number of parameters than can be reliably estimated in an unregularized approach is limited by the approximate dimensionality of the IDC timecourse matrix, which in this case is considerably less than 46.

At this time we did not extensively pursue this approach because the maximum likelihood approach worked well for the problem at hand. Nevertheless, this remains an appealing approach for theoretical reasons and may deserve an implementation in the future.

C.3.5 Polar Coordinate Regression

We defined the IDC transcriptome in section C.2 as a function which maps a one dimensional time parameter into a p dimensional space, $f : \mathbf{R} \rightarrow \mathbf{R}^p$. In other words there is a one-dimensional manifold that is embedded in an ambient, high dimensional space, and in some sense our task is to "unroll" it or to find its natural time-coordinate, which lives in a single dimension. The approach of Scholz and Fraunholtz [130] is to develop a neural network that learns (in a statistical sense) one-dimensional component of this manifold. In reference [130], the authors convincingly demonstrate the accuracy and utility of their approach, but it is complex and not straightforward to implement.

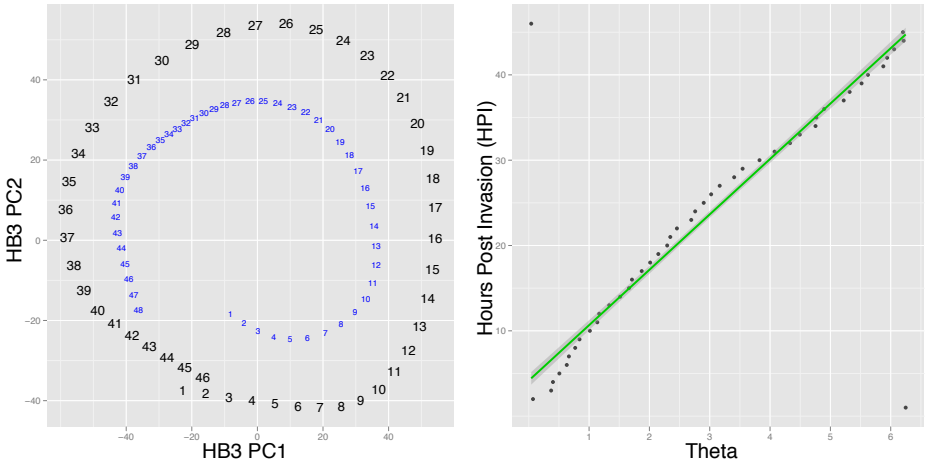

Figure C.8: Projection of Dd2 samples in the plane spanned by the first two principal components of the expression set (left). Regression of theta parameter in the space defined by the first two principal components recovers the time coordinate (right). The movement through the space defined by the first two principal coordinates is circular; therefore, a single regression after conversion to polar coordinates is able to "unroll" the one-dimensional manifold with remarkable success.

What is amazing, however, is that the one-dimensional component is immediately apparent from plotting the principal component scores. This is consistent across different samples (figure C.8). Because the samples trace out a ring, and the first two singular vectors are roughly equivalent in magnitude, the periodic component circular; the time coordinate can be retrieved by conversion to polar coordinates and univariate linear regression of θ:

$$t = \beta_0 \theta + \epsilon.$$

We implemented this on the unsmoothed data from [9]. The method works nicely but does encounter problems at the very beginning and end of the lifecycle. It is not clear whether these problems are due to experimental noise or a feature of the algorithm, as it may simply be that the final timepoint contains too much contamination from rings that have reinvaded to make accurate stage estimation possible.

C.4 The Gene Subset Selection Problem

If it is possible to accurately estimate the age of malaria infections using the expression of all of the approximately 6000 genes in the genome, is it possible to do it with fewer? And if so, which ones? This is an important question because it may not be possible to analyze large numbers of samples using microarrays or RNAseq, but it would be possible to use qRT-PCR. An analogous situation occurs in breast cancer: expression signatures have been used to sub classify tumors and these data have been used predict which individuals derive benefit from extensive chemotherapy regimens [40]. However, full-expression signatures are not better than specifically chosen subsets of genes, a finding which prompted Genomic Health to develop a 21-gene assay known as Oncotype DX [24], which is now routinely used in the clinic.

This is an instance of what is known as the variable selection or feature selection problem. For statistical purposes we want to drop or downweight genes that do not contribute meaningfully to the estimates since otherwise they will contribute noise and reduce the accuracy of our estimates. There are other advantages as well. Not only does using fewer variables for stage estimation have advantages in terms of cost of measurement and for statistical reasons, it also is important in understanding which genes are contributing most to the stage estimation of individual samples at different stages. These 'sentinal' or 'marker' genes would likely be important for biological processes during those stages.

We therefore set out to select the best subset of genes to use in a qRT-PCR-based assay to estimate sample age. For a subset of 25 genes, there are $\binom{6000}{25}$ possible choices of which genes to include, a number which is approximately 1×10^{70}. This is known as the feature selection problem, and it is a difficult problem because it is infeasible to test all subsets and choose the one that performs best. We therefore need an approximation which lets us choose a set of genes that will

perform well under a variety of conditions.

selection vs. regularization Closely related to the idea of variable subset selection is the idea of regularization, or setting a penalty term on the size of coefficients. Regularization is intimately connected to the $p \gg n$ problem, because adding a penalty on the size of the solution vector to underdetermined equations yields a unique solution. The basic notion is that we can optimize both for a good fit as well as for small coefficients, and in the process obtain a model that is more stable and parsimonious.

The incorporation of a gene-specific variance is a form of regularization, since the distance measure gives greater weight to genes with low variance. We can take these further. The basic idea of all of our heuristic solutions is to only include genes in the model which differentiate between timepoints. Genes that vary substantially across the lifecycle will differentiate between timepoints; we can therefore consider weighting further by measures of between-timepoint variability. We suggest three:

1. **high amplitude** A simple and appealing heuristic approach is to consider the expression curve for each gene to be sinusoidal of frequency $w = \frac{1}{2\pi T}$ hours where $T \approx 48$,

$$f_i(t) = a_i \sin(wt + \phi_i) + \epsilon. \tag{C.23}$$

   In that case, we have a two-parameter description of the fit for each gene. We choose those genes with large amplitudes, since they contribute more curvature to the likelihood functions, and we choose genes evenly distributed over the values of ϕ. Soft or hard thresholding is possible, i.e. genes could be weighted by amplitude and phase or genes below a minimum amplitude could be removed.

2. **between-timepoint variance** An alternative, closely related to amplitude but which does not assume a sinusoidal pattern for genes, is to calculate the overall variability of the gene during the lifecycle. Genes which vary little do not contribute much to discriminating between classes and therefore can be downweighted or removed.

3. **evenly sampling the principal ring** We have seen already that the vast majority of the variation in transcription during the intraerythrocytic lifecycle takes place in an affine set of dimension 2. Principal component analysis provides a way to capture the variation along axes that define a coordinate system for this set. The rotated space becomes a reduced dimension space because the coordinate axes have been chosen in such a way as to maximize the variation in the dataset along the subsequent principal components. The full SVD factorization

$$A = U\Sigma V^T \tag{C.24}$$

describes the columns of A as linear combinations of the columns of U whose relative weights are given by ΣV^T, and likewise the rows of A are linear combinations of the columns of V in amounts given by $U\Sigma$. In the reduced space

$$A = U\tilde{\Sigma}V^T, \tag{C.25}$$

space, the same holds true; for example a rank 2 approximation of A expresses the columns of A as a weighted sum of the 1st and 2nd column of U. In this sense, both genes and samples can be plotted in the same two-dimensional space. We show examples of this in figure C.4. These figures are known as biplots. We have already pointed out (Section 2.5) that the

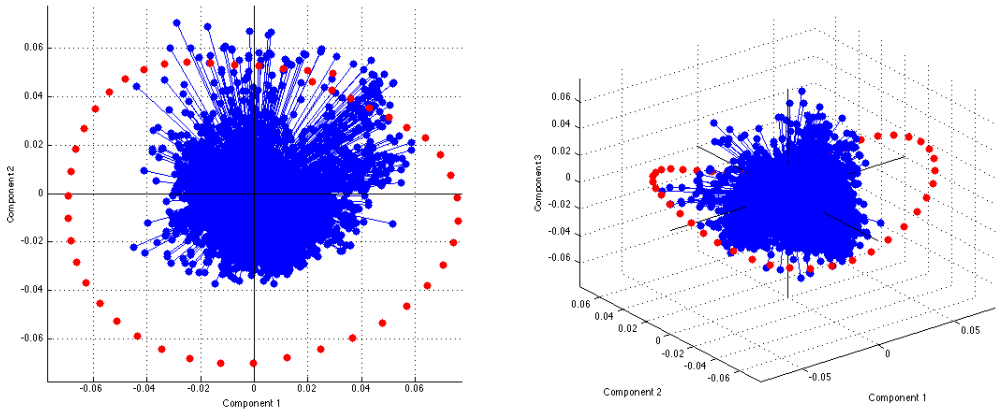

Figure C.9: Biplots take advantage of the diagonilization property the SVD to plot samples and genes in a reduced-dimension space. Here, biplots are shown in 2 and 3 dimensions. Samples are colored as red dots and genes are colored as blue dots. Genes which are important in individual samples can be identified by their proximity in these spaces.

samples map out a ring in principal component space. Genes that are important at each timepoint will point, in this space, in the same direction as samples. Therefore if we want to identify sentinel or marker genes which are characteristic of each timepoint, we can select genes based on their coordinates in this space. We chosen 25 genes in this fashion in figure C.4. These genes accurately represent the principal sources of signal in at each of the timepoints, and we therefore think they make a good choice for genes to include in the qPCR subset.

C.4.1 A qPCR Subset

While this is a problem that could benefit from further study, we used the approach of evenly sampling the "principal ring" to choose a subset. As discussed above, a feature of the SVD is that genes and samples can be plotted in the same space; by this method, the genes which are important at individual time points can be selected. Selecting ones which are equally distributed around the circumference

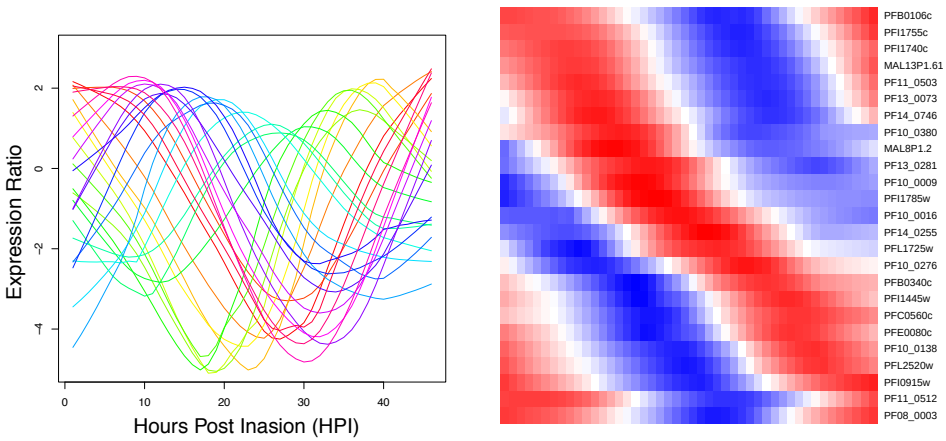

Figure C.10: A subset of genes chosen for qPCR by evenly sampling the sample ring from Figure C.9. The expression pattern and identity of the genes is shown, both as smoothed lines (left panel) and as a heatmap (right panel).

of the ring ensures that there will be genes informative for predicting the HPI regardless of the age of the reference samples. Choosing genes close to the exterior ring ensures that they will individually be informative, at least for time points close by their position on the ring.

Figure C.9 shows SVD biplots of genes and samples in 2- and 3-dimensions. We selected a set of 25 genes by this method and display their identities and expression patterns in Figure C.10. While we did not have the chance to evaluate the performance of this subset rigorously in the lab or computationally, we suspect they would perform well, and the method of gene choice is readily scalable to reducing or increasing the number of genes selected.

Appendix D

Polymer Theory

D.1 Introduction

A polymer is a repeated collection of subunits called monomers which are chemically joined together. For example, Figure D.1 shows a picture of a polymer composed of 10 monomers. Polymers are a fascinating class of molecules which occupy central roles in almost every aspect of modern life. A short list of well-known substances includes substances as different as Silly Putty, car tires, erasers, Pyrex, Gore-Tex, and many more. There can be no doubt that we live in the polymer age. In biology, polymers occupy an even more central position: DNA, RNA, proteins, and many other naturally occurring substances, are polymers. It is not an exaggeration to say that every living cell is a miniaturized polymer chemistry factory.

The macroscopic properties of materials derive from the statistical properties of the positions and momenta of their molecular constituents. This is also true for polymers, which as a class of materials show great diversity as well as an overarching universality. The variety in the form, texture, and physical properties of polymers is a direct result of their molecular architecture as long strings of atoms. There are a very large number of configurations available for the positions and momenta of gas molecules contained in a given volume; the same is true for polymers, although the space of configurations is substantially different between the two cases, and this is direct result of the molecular backbone.

In general, the theory of polymers is a highly developed branch of science in which experimental observations closely match theoretical predictions. This holds true for many aspects of polymer science, including the chemical properties of polymers, their static conformations as well as their dynamics. Since we are studying the folding of chromosomes, our interest is primarily in the static conformations of polymers.

The statistical description of polymers as chain molecules was introduced by

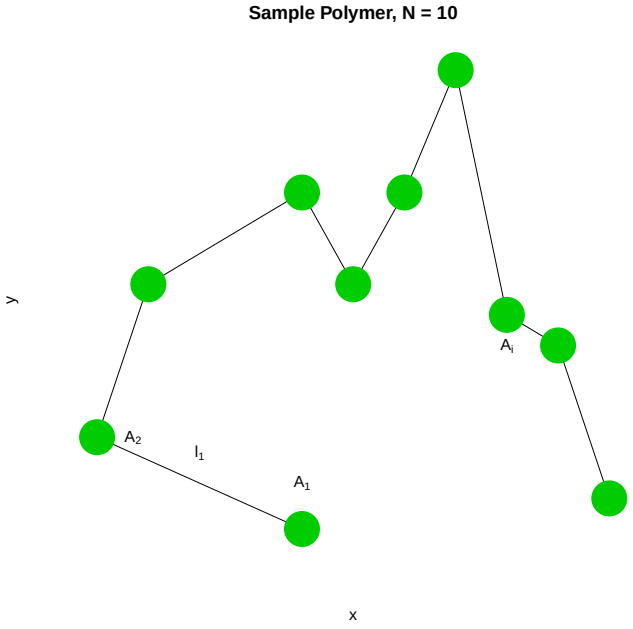

Figure D.1: A polymer with 10 monomers. This figure introduces common notational conventions in the polymer literature, which we try to mirror in this section. The monomers are denoted A_i and the first bond length is labeled ℓ_1

Kuhn in 1934 [72]. He introduced the close correspondence between the conformation of a polymer and a statistical random walk, a model which has been central to nearly all aspects of polymer theory ever since. The principles governing the conformational statistics of polymers were subsequently expanded in several directions, notably with the rotational isomeric model (due to Volkenstein [146]), the wormlike chain model (due to Kratky and Porod). Much of the research in this field was summarized and extended by Flory in the books [44, 45].

Because we need them as background at points in Chapter 5, we include standard results from the polymer literature in this Appendix.

D.2 notations and conventions

I try to follow standard notation in the polymer literature which are due to Flory and known as the Flory convention. As discussed in section B.1, the conventions

used for polymer conformation are slightly different from those used in the rest of the thesis. The main difference is that vectors are denoted by bold lowercase letters and, unless otherwise specified, are elements in real, three-dimensional Euclidean space, $\mathbf{R}^3$.

Polymers are composed of repeating subunits of monomers separated by a bond length, typically denoted ℓ. The conformation of a polymer (frozen in time) can be completely specified by the coordinates of its monomers.

In the plane defined by A_i, A_i, and A_{i+2}, and we set subunits A_{i-1} and A_i along the x-axis, then the angle between the horizontal and A_{i+1} defines the bond angle θ_i. The torsional angle, φ_i is the angle defined by rotating A_{i+1} around the A_{i-1}—A_i axis. The positioning between monomers can be summarized by a bond vector, $\mathbf{r}_i$, which is frequently given in spherical polar coordinates $\mathbf{r}_i = (r_i, \theta_i, \phi_i)$ or in cartesian coordinates $\mathbf{r}_i = (r_{xi}, r_{yi}, r_{zi})$. If the bond length between monomers is uniform, $l = r_i = \sqrt{r_x^2 + r_y^2 + r_z^2}$. The end-to-end distance, R_n of the polymer is the vector sum of the n bond vectors:

$$R_n = \sum_{1}^{n} \mathbf{r}_i$$

For an individual polymer frozen in time, $\mathbf{R}_n$ is a single distance value; however, in general for a large number of ideal chain polymers at a single time or a single ideal chain over a long time, $\mathbf{R}_n$ is studied as a random variable. The entire distribution of $\mathbf{R}_n$ values is considered, with particular interest paid to summary statistics such as the mean end-to-end distance (typically denoted $\langle \mathbf{R}_n \rangle$) or mean squared distance ($\langle \mathbf{R}_n^2 \rangle$).

D.3 Polymer Models

D.3.1 Ideal Chains

The starting point for all polymer conformation theory is the 'ideal' chain. This model, although highly simplified, forms the basis for more realistic theory and is very informative. The ideal chain model is so useful because it speaks directly to the universal characteristics that polymers share, independent of their chemical characteristics.

> "At first glance it seems to be a ridiculously crude model, almost a caricature: real polymer molecules live in a continuous space and have tetrahedral (109.47 °C) bond angles, a non-trivial energy surface for the bond rotation angles, and a repulsive hard-core monomer-monomer potential.
>
> In spite of these rather extreme simplifications, there is now little doubt that the self-avoiding random walk is not merely an excellent but in fact a *perfect* model for some (but not all!) aspects of the behavior of linear polymers in a good solvent. This apparent miracle arises from universality, which plays a central role in the modern theory of critical phenomena." – Alan Sokal

The ideal chain model is the simplest and most basic of all polymer models. It conceives of a polymer as connected monomers which do not interact (i.e. they do not repel or attract each other; they may in fact freely pass through each other. The bond length is typically held fixed) with each other when separated by a sufficiently large chain length. Interactions between the chains and the solvent are also not considered. Although no chains actually behave as ideal chains at their true length scales, at an adjusted length scale, an equivalent ideal chain, with altered bond length known as the Kuhn length, is a good model.

Each polymer has unique chemical and physical properties that determine the way it interacts with itself and the solvent it is in; yet in general, polymers also have general properties as a class of molecules, a fact which is due to their shared composition as chains of monomers. By multiplying the length scale of each polymer by a unique constant which describes local, chemical properties, one can convert any polymer to an 'equivalent freely jointed chain' with an effective bond

length known as the Kuhn length. This is an amazing fact which gives diffusion models pride of place in polymer physics.

For DNA, the Kuhn length (twice the persistence length in the worm-like chain, see below) has been estimated at 1200 bp.

D.4 Probability Distribution of End-to-End Distance

D.5 The end-to-end distance, R

We can consider summary statistics for chains under basic assumptions which then serve as a model for more realistic situations. For an ideal chain diffusingly isotropically (i.e. without a directional preference), $\langle \mathbf{R}_n \rangle = 0$ by symmetry (for every chain whose end-to-end distance is $-R_n$, there is an equally probable one whose end-to-end distance is R_n). What this tells us is that the probability distribution end-to-end distance R, for polymers of N monomers, $P(R; N)$, is centered at zero. We write this equivalently as $\langle R \rangle = 0$ (physics notation – read, "the ensemble average of R is 0", or $E[R] = 0$ (statistical notation – read "the expectation of the random variable R has expected value 0).

The statistical notation in particular indicates clearly that there is an entire distribution of values for R. What can we say about this distribution? We can calculate its variance, which (for a distribution with mean 0) is given by $E[X^2] = \langle \mathbf{R}_n^2 \rangle$ on the other hand is not zero and will be equal to the ensemble average over a large number of chains. Since we can consider the possible positions of invidual bond vectors, this reduces to the average angle between all possible bond vectors

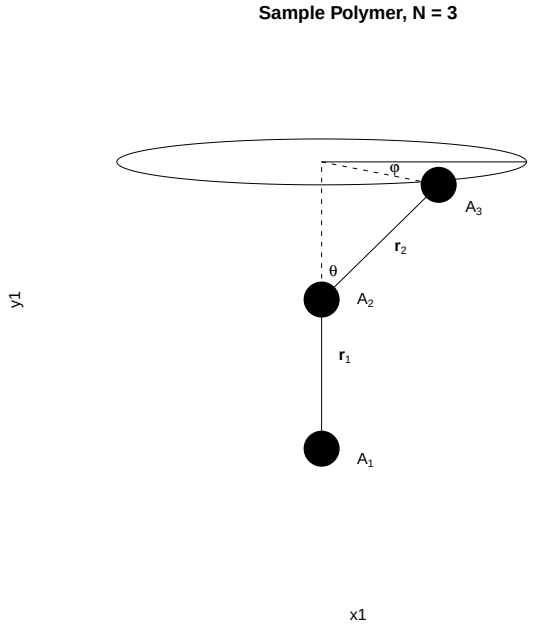

Figure D.2: A sample polymer of length 3 showing bond vectors, the torsional angle φ and the bond angle, θ. Monomers are in black, and the bonds between them are drawn as solid black lines.

in an individual chain, which (for equal bond lenghts, ℓ)

$$\ell^2 \sum_i^n \sum_j^n \langle cos(\theta_{ij}) \rangle$$

D.5.1 freely jointed chain

For the case of a chains with uncorrelated bond angles, the $\langle cos(\theta_{ij}) \rangle = n$ since $cos(\theta_{ij}) = 0$ for $i \neq j$ and 1 for $i = j$. In this case $\langle \mathbf{R}_n^2 \rangle = n\ell^2$.

This case can be improved by assuming that only *distant* bond angles are uncorrelated. In that case $n\ell^2$ is a lower bound and the above double sum converges to

$$C_n n \ell^2$$

where C_n is a constant that is known as Flory's characteristic ratio; its value is $>$ 1 for all real polymers.

D.5.2 Freely Rotating Chain

The torsional angle, φ_i, is the main source of flexibility in most polymers, as the bond angle is typically fixed. Whereas above we calculated var[$\mathbf{R}$] under the assumption that bonds were freely jointed, with their bond angles, θ_i, either uncorrelated or only locally correlated.

In the freely rotating chain, all θ_i are set equal to some θ, and the torsional angles, φ, are permitted to rotate freely, so that on average all torsional configurations are equiprobable.

In this case, the equation for var[$\mathbf{R}$] becomes

$$\sum_{i}^{n}\sum_{j}^{n}\langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle = \sum_{i}^{n}\sum_{j}^{n} cos(\theta)^{|i-j|}$$

is then an exponential decay in the bond vector correlation between subunits A_i and A_j. The entire expression reduces, after some algebra, to an expression:

$$\langle \mathbf{R}^2 \rangle = n\ell^2 \frac{1+cos(\theta)}{1-cos(\theta)}$$

The relationship $C_\infty = \frac{1+cos(\theta)}{1-cos(\theta)}$ provides a link between the freely-jointed model and the freely rotating model.

D.5.3 Worm-Like Chain

The worm-like chain model is used for stiff polymers which typically have bond angles close to 0. This model is particularly relevant to our case because it is a frequently used model for double stranded DNA.

The model is derived by using a small angle approximation for cos(θ):

$$\cos(\theta) \approx 1 - \frac{\theta^2}{2}$$

for $0 < \theta << 1$.

For the worm-like chain, a quantity called persistence length, denoted ℓ_p is used to describe the length scale at which correlations between bond vectors persist. Think, for example, of a garden hose. At small length scales, the hose is stiff like a rod; at larger lengths, however, the garden hose appears flexible like a spaghetti. These properties are well-modeled by the worm-like chain.

For the worm-like chain, the expression for var[$\mathbf{R}$] derived for the freely-jointed chain is written

$$\ell^2 \sum_i^n \sum_j^n \cos(\theta)^{|j-i|} = \ell^2 \sum_i^n \sum_j^n e^{\frac{|j-i|}{\ell_p}}$$

, where $\ell_p = \frac{-\ell}{ln(\cos(\theta))} \approx \frac{2\ell}{\theta^2}$.

In order to get a closed-form expression, the summation is replaced by integration, leaving an expression for var[$\mathbf{R}$]:

$$\langle \mathbf{R}^2 \rangle \approx 2\ell_p R_{max}$$

D.6 Probability Distributions

We have so far modeled polymers, which do not have a deterministic structure, as random variables. We have briefly surveyed some models for ideal chains that permit closed form formulations of the expected value and variance of these random variables. In general, the probability distribution function of a random variable contains much more information than can be expressed by these two quantities, and we seek expressions for the full distribution function.

In general, a closed form solution is available only for simplified cases. In particular, this means essentially the "random walk" model of movement on a lattice of points, and its approximation in the limit by diffusion or Brownian motion.

A simple random walk on a lattice is a stochastic process, $S(t)$, defined such that $S(t+1) = S(t) + q$, with $q = \pm 1$. A 'symmetric' random walk occurs when $P(q = 1) = 0.5 = P(q = -1)$. This is not the only way to simulate a random walk (an alternative is to set p equal to a random sample from the uniform probability distribution on the interval [-0.5, 0.5]). The fact that only +1 or -1 can be added to the walk at each epoch, t, gives it the character of occuring on a lattice.

The probability for each sample path in a random walk on a lattice can be computed exactly as follows. In a symmetric walk all paths are equally probable. The expression $\binom{n}{r}$ is shorthand for $\frac{n!}{(n-r)!r!}$, and gives the number of ways of choosing r objects from a set n. In general, after n epochs of t, there will be r moves upward and $n - r$ moves downward. Each point on the lattice is the final destination after n moves up and $n-r$ down. Therefore y, the vertical position on the lattice at time t will be equal to $r-(n-r)$. Solving for y in the combinatorial identity, yields $\frac{t!}{(\frac{t+y}{2})!(\frac{t-y}{2})!}$. Since each path has probability $\frac{1}{2^t}$, the probability distribution

$$P(t, y) = \frac{1}{2^t} \frac{t!}{(\frac{t+y}{2})!(\frac{t-y}{2})!}$$

This is an exact probability distribution for the paths of symmetric random walks on a lattice but it is unweildy to evaluate for large y because of the factorial. The approach typically taken is to approximate it using a Gaussian distribution, and (after some algebra) the one specified is

$$P(t, y) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$$

A few remarks are in order about this probability distribution. First, it as a Gaussian with mean, $\mu = 0$, and variance $\sigma^2 = t$, $N(0, t)$. For the derivation of the standard symmetric random walk, The variable t began as a discrete valued variable taking values on the integers, with the interpretation that it stood for the number of monomers in the chain. t however can also be interpreted as the

time parameter of a single particle undergoing diffusion. In the case where t is a continuous parameter, the particle is undergoing Brownian motion in the sense described the Weiner process.

This provides a second view of the Gaussian approximation used above, which is as the solution to a partial differential equation governing diffusive motion. In this case, the above Gaussian is the solution (under the initial condition that the particle begins as a point mass at the origin) to

$$\frac{\partial P}{\partial t} = -\frac{1}{2}\frac{\partial^2 P}{\partial x^2}$$

a well known partial differential equation first formulated and solved as the heat equation.

The convergence to a Gaussian for large n is striking. A particularly surprising feature of the derivation is that the probability distribution, which started out on a lattice lacking rotational symmetry, gained this symmetry asymptotically. This is sometimes called an emergent symmetry.

In Chapter 5, we are measuring contact affinity between different sections of chromosomes. Suitably normalized against the total number of observed interactions, this can be converted into contact probability. By establishing good polymer models and cataloguing their assumptions, we can study expected contact probability under these models either computationally or – where the models are solvable – analytically. Ultimately, these models can help us interpret the data from GCC and HiC experiments as shown Chapter 5.

D.6.1 Contact Probability and Scaling

One of the things that makes the intra-chromosomal interaction data so useful to us is that we can derive an analytical expression for how often polymers will contact themselves as a function of linear distance. The simplest way to do this is

to discretize space into a 'lattice', i.e. a set of coordinate axes whose constituent points are evenly spaced at some interval. Then a good model for a polymer (a 'random walk polymer') can be constructed by taking random steps on this lattice in the x, y, and z directions. We then ask what is the probability of returning to the origin after a given number of steps. It is possible to derive a closed-form expression for this quantity which can guide our experimental observations, as follows:

On a one-dimensional lattice, the probability of return (i.e. contact) is equal to taking N steps up and N steps down after a total of 2N steps. In multiple dimensions, we can consider the walks to be independent; the probabilities then multiply. Using Stirling's approximation for the factorial function, it is then straightforward to show that for large N, we expect the probability of making a contact to follow $P \sim N^{-d/2}$ where d is the dimensionality of the walk. In a three-dimensional space, we expect contact probability to scale as the number of steps to the $-3/2$ power. This is a 'scaling law' and is most easily identifiable on plot with two logarithmic axes.

I have plotted the theoretical probability distribution for self-contact for a random-walk polymer, on both linear and logarithmic axes, in figure D.3. The three curves correspond to a 1, 2, and 3 dimensional space. Although polymers free in solution are clearly diffusing in three dimensions, retaining the dimensionality of the space as a free parameter in the contact probability distributions gives us some insight into what is different about the behavior of chromosomes as polymers in our data. The empirical probability distributions, in both linear and log-log format, are displayed for a single chromosome in Figure 5.4, and for all fourteen chromosomes on doubly logarithmic axes in Figure 5.5.

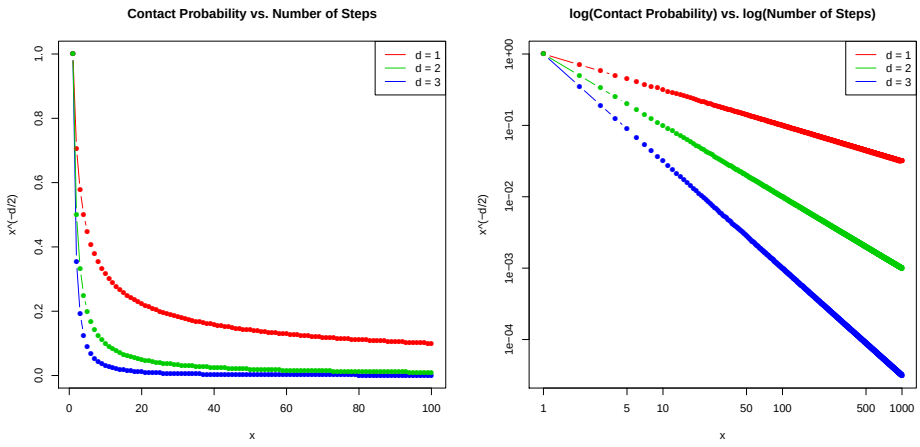

Figure D.3: Linear (left) and doubly logarithmic (right) plot of return probability on a d-dimensional showing the power-law form of the distribution. As mentioned in Section 5.4, for a three-dimenionsal ($d = 3$, blue curve), we expect a scaling exponent of $-3/2$.

D.7 The Use of Simulation

Simulation can be used to calculate the properties of polymers, such as the probability of contact, when analytical expressions are not available.

We ran some simple simulations of polymer configuration by parameterizing two sorts of brownian motion in three dimensions: regular brownian motion, $W(t)$, and a Brownian motion forced to converge to a different point in space. This is often referred to as a Brownian bridge, $B(t)$. The algorithm used to simulate $W(t)$ is in this case:

$$W(t_{i+1}) = W(t_i) + \texttt{uniform[0,1] - 0.5}$$

where `uniform[0,1]` denotes the uniform probability density function on the interval $[0, 1]$. The Brownian Bridge conditioned to hit a point Q was simulated as:

$$B(t_{i+1}) = W(t_i) - tW(t_n) + Qt.$$

Some example polymers from these simulations are shown in Figure D.4. The average distance from one end of the polymer to the midpoint of the polymer is shown in Figure D.5. These simulations, though rudimentary, show some interesting and biologically relevant features. The midpoint of the freely diffusing polymer shows increasing average distance from its end which increases as a function of the square root of the number of monomers, N. This is consistent with theory: because the midpoint of the chain is itself just the end of a Brownian motion which is independent of the monomers ahead of it. The variance of the probability distribution scales with t, so its standard deviation scales with $\sqrt{t}$.

The anchored Brownian motion on the other hand shows categorically different behavior. Polymers tethered between two points are by nature more 'stretched'. The probability distribution in the middle of the chain has smaller variance. The polymers are not free to wander around the space in the same manner as the unanchored ones. As the chains get larger, the difference between constrained chains and unconstrained chains is hard to distinguish; however, the rate of increase in end-to-midpoint distance considered as a function of change length appears to be linear. This suggests that anchored polymers, as malaria chromosomes are, will behave differently in the nucleus from unanchored ones. Statistically, the position of the chain at its midpoint is no longer independent from that of the chain at its endpoint, as it is in the unconstrained case.

The fact the diffusion of malaria chromosomes is anchored may explain the unusual scaling relationship at large distances 5.4. Conducting rigorous simulations to establish whether the decreased likelihood of forming an interaction can be due to the anchoring of portions of malaria chromosomes by heterochromatin foci is a point of planned future research.

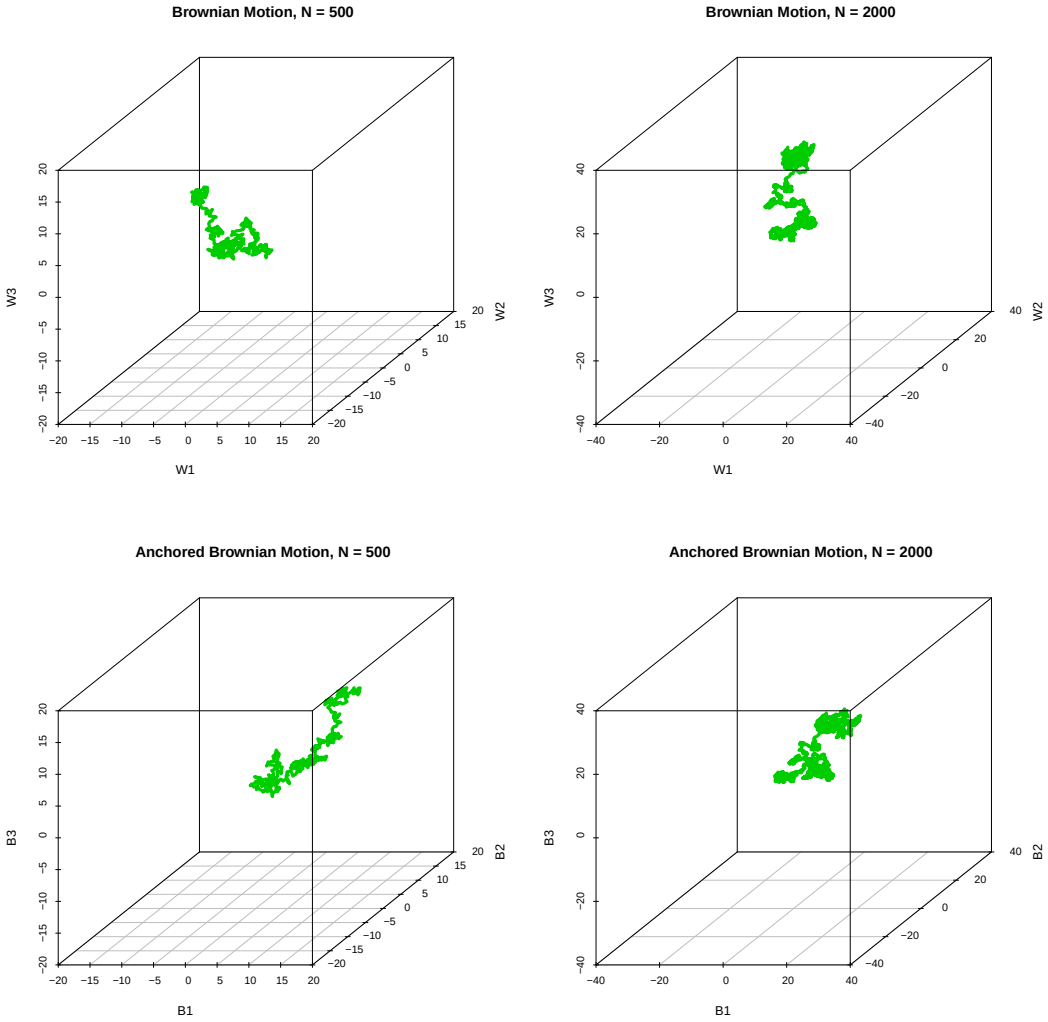

Figure D.4: Single sample paths for simulated Brownian Diffusions and Brownian Bridges for polymers of 500, and 2000 Kuhn monomers, corresponding to approximately 600 and 2400 KB of DNA, respectively. This shows the qualitatively different behavior for anchored Brownian motion ("Brownian Bridge") for polymers in which the ratio between the number of monomers and the distance between anchors is small.

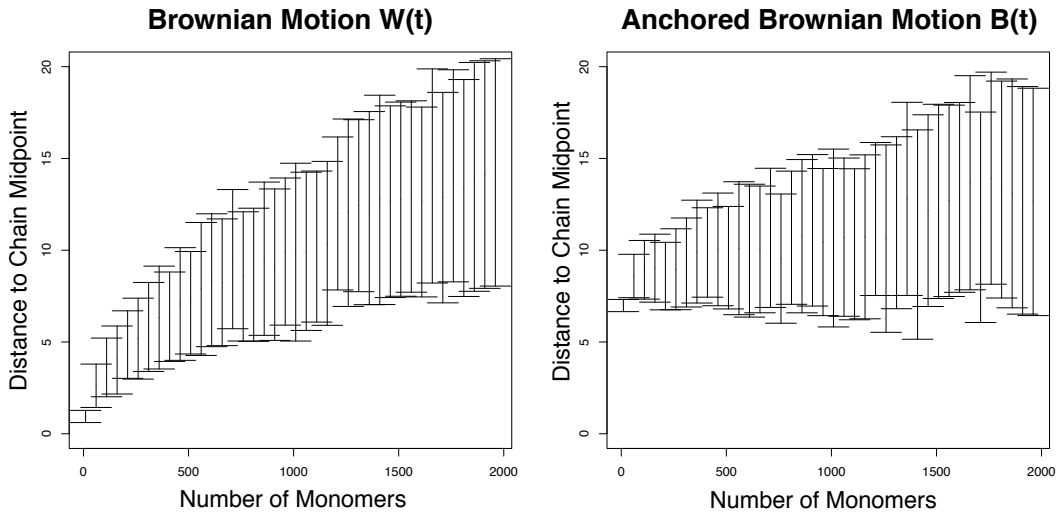

Figure D.5: Euclidean distance between origin and midpoint of chain (100 simulations per chain length)

Appendix E

Structure Reconstruction

E.1 Structure Reconstruction

Ideally we would like a way to display the HiC data that is both simple, intuitive, and informative in a way that contact matrices are not. Since there is clearly some relationship between contact probability and distance, by making reasonable assumptions about this relationship we should be able to convert contact probability into a set of distances. This is turn will help us interpret the data. Our brains are not trained to think about matrices of contract probability; they are however good at analyzing images.

In doing this, there are two main challenges: 1) we must identify (i.e. guess, measure, derive, or divine) the appropriate conversion between contact probability and average spatial distance. Then 2) find a way to create a realizable structure in three dimensional space that is as close as possible to the 'distances' that we have measured.

In terms of specifying the relationship between contact probability and distance, my first approach has been to use a similar function to what others have used. This mapping is shown in Figure E.1. We expect, both from past studies performing both FISH and 3C, as well as from scaling arguments in polymer physics, for the relationship to follow a power law, e.g. $y = \frac{1}{x}$ or $y = \frac{1}{x^2}$ for y distance and x contact probability. At the moment, mainly because I had to choose something to get started, the function I have used is essentially a discretized version of a power law with some informed guesses at parameters. Once we have 3D distance measurements working with FISH, we should be much better able to estimate the true form of this function.

Once a mapping between contact probability and mean spatial distance has been specified, we can transform the contact probability matrix into distance matrix. A distance matrix is symmetric, has zeros along the diagonal, and strictly non-negative entries. The original contact probability matrix and transformed

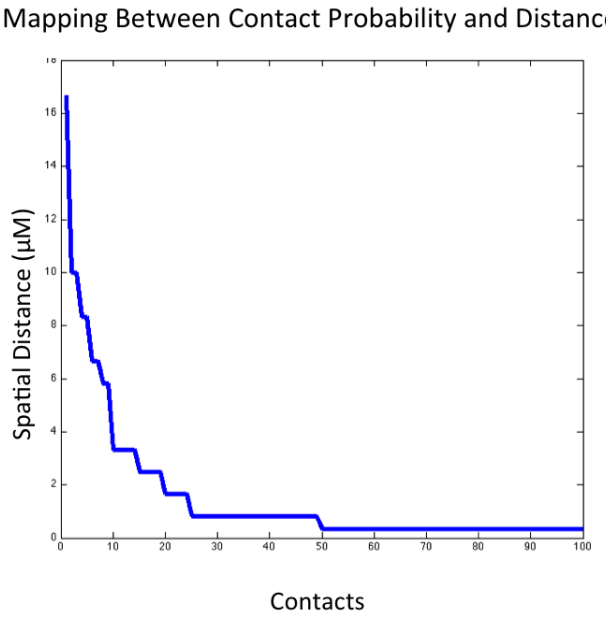

Figure E.1: The mapping relating contact probability to distance. Converting a contact probability matrix, A to a distance matrix Δ requires a function relating the number of contacts between two loci to their average spatial distance. This is not known for malaria chromosomes, and would be prohibitive to experimentally measure since the number of FISH experiments, and time required, would be enormous. We therefore used a function similar to that used for yeast chromosomes [35] in our modeling efforts.

distance matrix for chromosome 2 HiC data from the MboI library are shown in Figure E.2. The left and right halves of Figure E.2 clearly relate to each other; at this point I think it is fair to say that we haven't done anything too unreasonable to the data.

We can give a mathematical formulation of the second part of the problem as follows. We represent each chromosomal segment as a bead, which corresponds to some interval (in, e.g., kilobases). We then face the challenge of arranging the beads in space in such a way as to match the observed spatial distance (obtained from transforming the contact probability matrix). We denote each bead as x_i. In this case the x_i's are vectors with three components ($x_i = [x_i^{(x)} x_i^{(y)} x_i^{(z)}]$) and correspond to points in a three-dimensional space. The (Euclidean) distance between

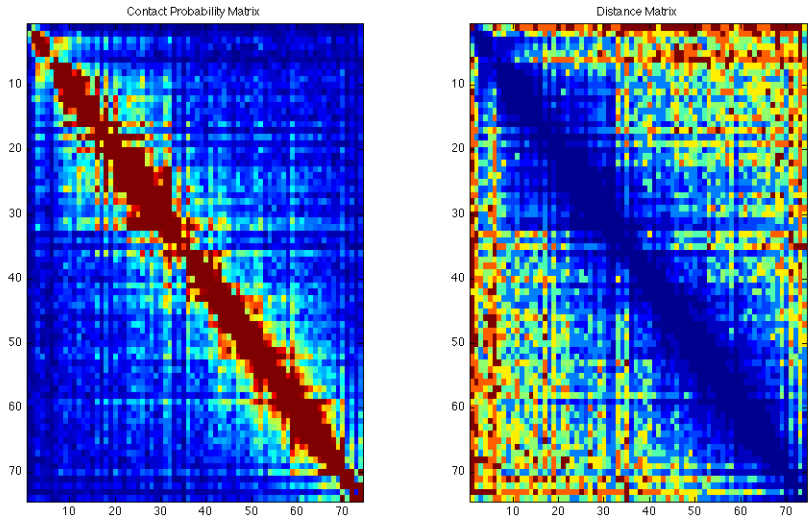

Figure E.2: Observed Contact Probability Matrix, A, (left) and Estimated Distance Matrix, Δ, (right) Chromosome 2. The individual elements are related by $f(a_{ij}) = \delta_{ij}$, where the function f is that given in Figure E.1.

points x_i and x_j is then given by:

$$dist(x_i, x_j) = [(x_i^{(x)} - x_j^{(x)})^2 + (x_i^{(y)} - x_j^{(y)})^2 + (x_i^{(z)} - x_j^{(z)})^2]^{1/2}$$

when written componentwise. More compactly we can write

$$dist(x_i, x_j) = \|x_i - x_j\|.$$

If we know the distance measurements exactly, then we simply place all the x_i in space to satisfy these distances. This involves writing down, and solving, a set of algebraic equations. An alternative approach is to formulate the problem as an optimization problem. Though the first is more elegant, here we choose the second because it more easily accommodates the transition to the next step when distances are known with error. Letting δ_{ij} denote the known distances, one

formulation of such a problem would be as follows:

$$\text{minimize} \quad \sum_{i,j} \quad (\|x_i - x_j\| - \delta_{ij})^2 \tag{E.1}$$

In this case we are minimizing a function that computes the sums of the squared differences between the distances in our 'test' configuration and the true distances. If the distance measurements are correct, then the optimal value of this objective function should be zero.

Nevertheless, for the structures we are attempting to solve, the true distances are not known. Further, if we computed an 'average' distance for a folding polymer, the set of distances would almost certainly not correspond to something that can be faithfully represented in a 3-dimensional space. Therefore the optimization approach proves useful and we accept a non-zero optimal value of our objective function.

This is a very logical problem to solve in our case. One drawback is that, in solving this problem, the points tend to cluster at in an unrealistic, unphysical blob. In real life, there are constraints on how close together two beads can be, how far apart they can be, as well as a number of other constraints that have to do with the chemical and physical properties of DNA polymers. Therefore, at the very minimum to make the structure more realistic, we need to include these constraints in the optimization problem. In this case we put lower and upper bounds, denoted b_l and b_u on the distances between neighboring points.

The rewritten problem looks as follows:

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i,j} \quad (\|x_i - x_j\| - \delta_{ij})^2 \\ \text{subject to } & \|x_i - x_j\| \geq b_l, for\ |i-j| = 1 \\ & \|x_i - x_j\| \leq b_u, for\ |i-j| = 1 \end{aligned} \tag{E.2}$$

for $x_i, x_j \in \mathbf{R}^3$, δ_{ij} is entry in (observed) distance matrix

At this point the problem has been reformulated to be more physical, but now there is a computational problem. The lower bound makes the problem 'non-convex'; specially the feasible set over which we are searching is now not a convex set. Non-convex optimization problems are notoriously difficult; in practice, a non-convex problem essentially means that we can no longer find the global minimum in a reasonable amount of time. For convex problems, on the other hand, when a solution is found it is certified to be globally optimal.

E.2 Distance Geometry

At this point we have two options: We can either solve the problem in non-convex form or search for a convex reformulation or transformation. Problem E.2 is basically intractable because of the non-convexity; such problems are known to be NP-hard, meaning that "solution" of the problem involves accepting a locally optimal solution for which we can say almost nothing about its global optimality. This is the traditional approach taken, for example, by Duan et al for the yeast genome [35], but it is our view an approach with frustrating drawbacks. In particular, the "solution" is sensitive to initial conditions as well as the search algorithm used, and there is way to know how far from optimality the local "solution" is.

The second approach involves searching for a convex reformulation of Problem E.2. As we show below, such a reformulation is possible using results from distance geometry. We develop that here.

Convexity

In optimization, convexity divides computationally hard problems from straightforward ones. Optimization problems that can be proved convex are certifiably solvable in polynomial time, whereas those that are not require either that one

accepts a locally optimal solution or an exceedingly large computation time (in many problems, this is functionally infinite). It is known that problems involving Euclidean distance are closely related to convex sets and functions, largely because the set of euclidean squared distance matrices forms a convex cone in the vector space of symmetry matrices, $\mathbf{S}$. We give some definitions for convex sets and functions, and describe the relevant theory for euclidean distance matrices.

A set $\mathcal{C}$ is convex if

$$\alpha x + (1-\alpha)y \in \mathcal{C}, \tag{E.3}$$

for all $\alpha \in [0,1]$, and a function, $f : \mathbf{R}^n \to \mathbf{R}$ is convex if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y). \tag{E.4}$$

In other words, convexity of a set means that any weighted sum of two points remains inside the set, while convexity of a function means that a weighted sum of points calculated in the domain of the function is less than or equal to a weighted sum of points after the function acts. A convex optimization is one which minimizes a convex function over a convex set. Standard form [8] for a convex optimization problem is

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \\ & h_i(x) = 0, \end{aligned} \tag{E.5}$$

where $f_0(x)$ are $f_i(x)$ required to be convex functions, and the $h_i(x)$ are affine functions. This ensures that the problem is convex.

Non-convexity enters problem E.2 through the feasible set – in particular,

the quadratic constraints on the lower and upper bounds of the distances in the polymer backbone make the feasible set non-convex, which are not permitted by equation E.5. We therefore develop a reformulation of the problem, using results from distance geometry.

Distance Geometry and Euclidean Distance Matrices

By a well-known procedure in distance geometry we are able to transform these quadratic constraints on distances into linear constraints on a matrix of inner products, known in linear algebra as a Gram matrix $G = X^T X$. The relationship between G and X can be seen as follows:

$$\|x_i - x_j\|^2 = x_i^2 + x_j^2 - 2x_i^T x_j. \tag{E.6}$$

Equality E.6 relates the distances between vectors to their lengths and the angle between them, and is sometimes called the law of cosines because of the relation

$$\cos(\theta) = \frac{x_i^T x_j}{\|x_i\|\|x_j\|}. \tag{E.7}$$

If D is a matrix which contains the squared distances between points, then using equation E.6 we can write:

$$D = \text{diag}(D)^T \mathbf{1} + \mathbf{1}\text{diag}(G)^T - 2G. \tag{E.8}$$

or, written component wise:

$$d_{ij} = g_{ii} + g_{jj} - 2g_{ij} \tag{E.9}$$

This characterization is due to Young and Householder [151], and is a part of a body of results, first established by Schoenberg [128], that relate the cone

of euclidean distance matrices, **EDM** to the positive semidefinite cone, $\mathbf{S}_+$. In other words, X describes a set of points in Euclidean space if $X^T X \succeq 0$, and the related distance matrix whose elements are $d_{ij} = \|x_i - x_j\|$, is a Euclidean distance matrix. A more direct condition on the distance matrix, D, was proved by Schoenberg in 1935. The Schoenberg criteria for $D \in \mathbf{EDM}$ is:

$$(1/2)VDV \preceq 0 \tag{E.10}$$

where V is the geometric centering operator,

$$I - (1/n)\mathbf{1}\mathbf{1}^T.$$

In other words, D is a euclidean distance matrix if it is negative semidefinite after geometric centering.

Because cones are particularly suited for optimization, and conic programming techniques are well developed, we can exploit that relationship to minimize our function over the cone of euclidean distance matrices. The relationship between $\mathbf{S}_+$ and **EDM** permits the use of semidefinite programming techniques to identify an optimal solution. Furthermore, we note that since equation E.8 is affine in G and can be used to express the distance constraints as affine inequalities. In fact, because of the relationship in equation E.7, constraints involving angles can be incorporated into the optimization problem; these take the form of linear inequality constraints on the Gram matrix, maintaining problem convexity. The inclusion of angle data is especially useful in incorporating polymer models, such as the worm-like chain model, into the optimization problem.

$$\text{Let} D = \text{diag}(G)^T \mathbf{1} + \mathbf{1}\text{diag}(G)^T - 2G$$

$$\begin{aligned} \text{minimize} \quad & \|D - \Delta\|_F \\ \text{subject to} \quad & d_{ij} \geq b_l \ \text{ for } |i-j| = 1 \\ & d_{ij} \leq b_u \ \text{ for } |i-j| = 1 \\ & G \succeq 0 \end{aligned} \tag{E.11}$$

for Δ is observed distance matrix.

We solve problem E.11 by a conic optimization method known as semidefinite programming ("SDP"). We use SeDuMi, a freely available semidefinite program solver that calls an interior point algorithm. A full explanation of semidefinite programming theory and solver implementations can be found elsewhere. Here it is enough to say that such solvers are capable of efficiently minimizing convex functions over convex constraint sets, and producing solutions which are guaranteed to globally optimal. One downside of these solvers is that they are capable of handling only small to medium sized problems, and for euclidean distance problems have a practical maximum problem size of approximately 200 'points'. Results for several structures are shown below. Structures larger than 200 beads, including the full genome, present difficulties and addressing these issues will require a bigger computer or reducing the resolution of the solution. Nevertheless, with the problem formulated in this way, finding the 'best' reconstructed structure is simply a matter of computing.

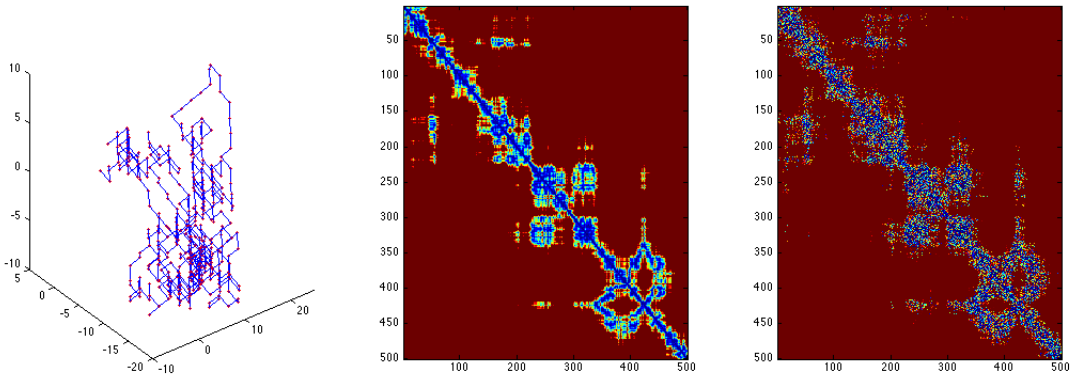

Figure E.3: A random-walk polymer (left) and its distance matrix, Δ, (middle) plus additive Gaussian noise (right)

E.3 Chromosome Structures

The positions for all beads x_i, from the chromosome 2 distance matrix in Figure E.1, are shown in the left panel of Figure E.4. Distance measurements obtained from the reconstructed structure are in the right side of Figure E.4. While not perfect, it is clear that the reconstructed does capture the essence and some of the nuance of the original spatial distance data in Figure E.1. Once the bead positions have been calculated, we can render the structural model by fitting a continuous, smooth function (a B-spline) through the beads. The three dimensional structure can then be examined using Pymol or a similar molecular graphics program, in the same way as a protein crystal structure. Examples of the computed structure of chromosome 3 and the first half of chromosome 12 are shown in Figures E.5 and E.6.

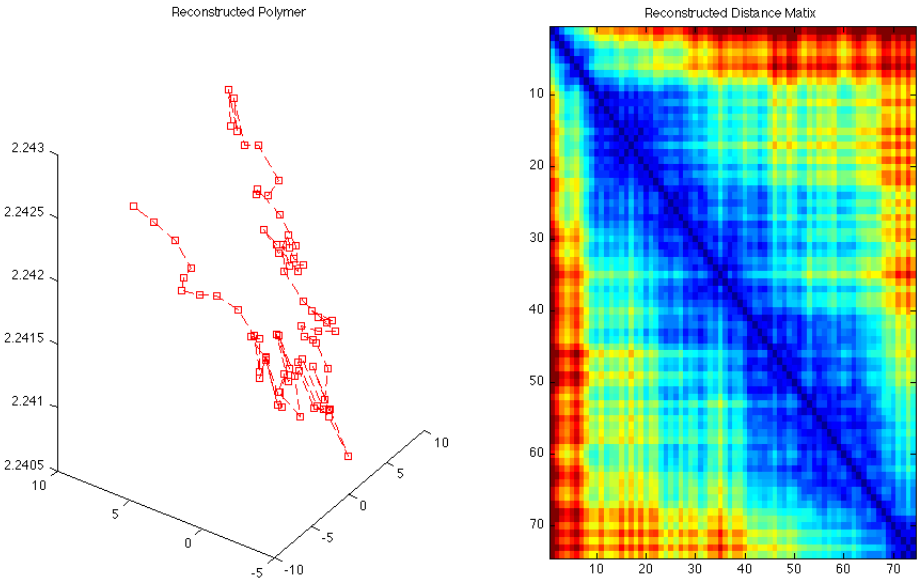

Figure E.4: Best-Fit Structure (left) and Corresponding Distance Matrix (right).

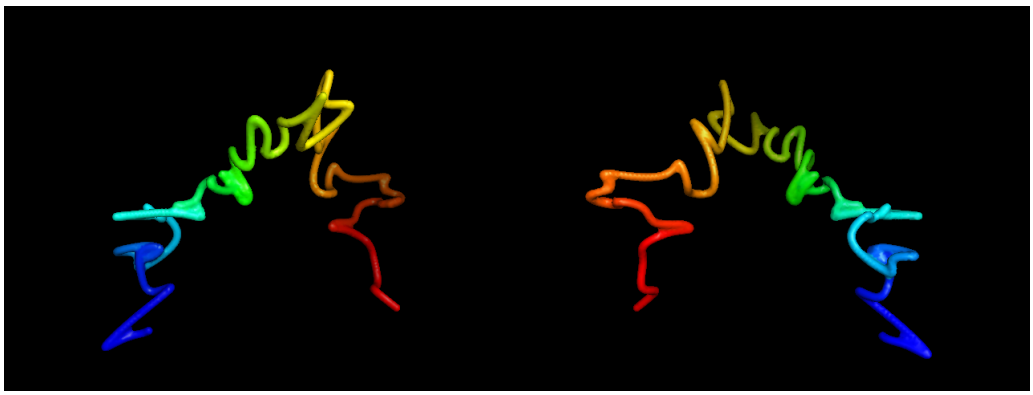

Figure E.5: Solution of Chromosome Structure by Convex Optimization: Chromosome 3, front and back. The reconstructed coordinates have been smoothly interpolated by a spline and rendered in PyMol.

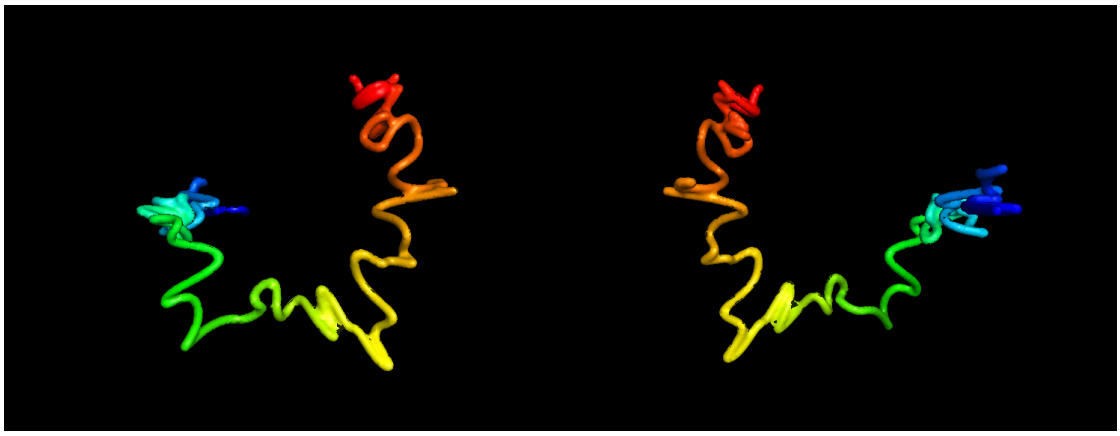

Figure E.6: Solution of Chromosome 12, atoms 10 to 135 Structure by Convex Optimization, front and back. As above in Figure E.5, the reconstructed coordinates have been interpolated and the resulting structure is rendered in PyMol.

Bibliography

[1] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–10106, Aug 2000.

[2] Orly Alter, Patrick O Brown, and David Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A*, 100(6):3351–3356, Mar 2003.

[3] S. Balaji, M. Madan Babu, Lakshminarayan M Iyer, and L. Aravind. Discovery of the principal specific transcription factors of apicomplexa and their implication for the evolution of the ap2-integrase dna binding domains. *Nucleic Acids Res*, 33(13):3994–4006, 2005.

[4] Ziv Bar-Joseph, Georg Gerber, Itamar Simon, David K Gifford, and Tommi S Jaakkola. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc Natl Acad Sci U S A*, 100(18):10146–10151, Sep 2003.

[5] D. I. Baruch, J. A. Gormely, C. Ma, R. J. Howard, and B. L. Pasloske. Plasmodium falciparum erythrocyte membrane protein 1 is a parasitized erythrocyte receptor for adherence to cd36, thrombospondin, and intercellular adhesion molecule 1. *Proc Natl Acad Sci U S A*, 93(8):3497–3502, Apr 1996.

[6] D. I. Baruch, B. L. Pasloske, H. B. Singh, X. Bi, X. C. Ma, M. Feldman, T. F. Taraschi, and R. J. Howard. Cloning the p. falciparum gene encoding pfemp1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, 82(1):77–87, Jul 1995.

[7] A. R. Berendt, D. L. Simmons, J. Tansey, C. I. Newbold, and K. Marsh. Intercellular adhesion molecule-1 is an endothelial cell adhesion receptor for plasmodium falciparum. *Nature*, 341(6237):57–59, Sep 1989.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge, 2004.

[9] Zbynek Bozdech, Manuel Llins, Brian Lee Pulliam, Edith D Wong, Jingchun Zhu, and Joseph L DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biol*, 1(1):E5, Oct 2003.

[10] Zbynek Bozdech, Jingchun Zhu, Marcin P Joachimiak, Fred E Cohen, Brian Pulliam, and Joseph L DeRisi. Expression profiling of the schizont and trophozoite stages of plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biol*, 4(2):R9, 2003.

[11] K. N. Brown and I. N. Brown. Immunity to malaria: antigenic variation in chronic infections of plasmodium knowlesi. *Nature*, 208(5017):1286–1288, Dec 1965.

[12] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12):4164–4169, Mar 2004.

[13] P. C. Bull, B. S. Lowe, M. Kortok, C. S. Molyneux, C. I. Newbold, and K. Marsh. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nat Med*, 4(3):358–360, Mar 1998.

[14] Peter C Bull, Caroline O Buckee, Sue Kyes, Moses M Kortok, Vandana Thathy, Bernard Guyah, Jos A Stoute, Chris I Newbold, and Kevin Marsh. Plasmodium falciparum antigenic variation. mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Mol Microbiol*, 68(6):1519–1534, Jun 2008.

[15] Richrd Brtfai, Wieteke A M Hoeijmakers, Adriana M Salcedo-Amaya, Arne H Smits, Eva Janssen-Megens, Anita Kaan, Moritz Treeck, Tim-Wolf Gilberger, Kees-Jan Franoijs, and Hendrik G Stunnenberg. H2a.z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by h3k9ac and h3k4me3. *PLoS Pathog*, 6(12):e1001223, 2010.

[16] Tracey L Campbell, Erandi K De Silva, Kellen L Olszewski, Olivier Elemento, and Manuel Llins. Identification and genome-wide prediction of dna binding specificities for the apiap2 family of regulators from the malaria parasite. *PLoS Pathog*, 6(10):e1001165, 2010.

[17] Thanat Chookajorn, Ron Dzikowski, Matthias Frank, Felomena Li, Alisha Z Jiwani, Daniel L Hartl, and Kirk W Deitsch. Epigenetic memory at malaria virulence genes. *Proc Natl Acad Sci U S A*, 104(3):899–902, Jan 2007.

[18] Thanat Chookajorn, Ron Dzikowski, Matthias Frank, Felomena Li, Alisha Z Jiwani, Daniel L Hartl, and Kirk W Deitsch. Epigenetic memory at malaria virulence genes. *Proc Natl Acad Sci U S A*, 104(3):899–902, Jan 2007.

[19] Gordon C. Cook and Ali, editors. *Manson's Tropical Diseases*, chapter Malaria. Elsevier, 2008.

[20] Peter R Cook and Davide Marenduzzo. Entropic organization of interphase chromosomes. *J Cell Biol*, 186(6):825–834, Sep 2009.

[21] Alfred Corts, Celine Carret, Osamu Kaneko, Brian Y S Yim Lim, Alasdair Ivens, and Anthony A Holder. Epigenetic silencing of plasmodium falciparum genes linked to erythrocyte invasion. *PLoS Pathog*, 3(8):e107, Aug 2007.

[22] Richard M R Coulson, Neil Hall, and Christos A Ouzounis. Comparative genomics of transcriptional control in the human malaria parasite plasmodium falciparum. *Genome Res*, 14(8):1548–1554, Aug 2004.

[23] A. F. Cowman, D. Galatis, and J. K. Thompson. Selection for mefloquine resistance in plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. *Proc Natl Acad Sci U S A*, 91(3):1143–1147, Feb 1994.

[24] Maureen Cronin, Chithra Sangli, Mei-Lan Liu, Mylan Pho, Debjani Dutta, Anhthu Nguyen, Jennie Jeong, Jenny Wu, Kim Clark Langone, and Drew Watson. Analytical validation of the oncotype dx genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. *Clin Chem*, 53(6):1084–1091, Jun 2007.

[25] Cheston B Cunha and Burke A Cunha. Brief history of the clinical diagnosis of malaria: from hippocrates to osler. *J Vector Borne Dis*, 45(3):194–199, Sep 2008.

[26] Erica L Dahl, Jennifer L Shock, Bhaskar R Shenai, Jiri Gut, Joseph L DeRisi, and Philip J Rosenthal. Tetracyclines specifically target the apicoplast of the malaria parasite plasmodium falciparum. *Antimicrob Agents Chemother*, 50(9):3124–3131, Sep 2006.

[27] J. P. Daily, D. Scanfeld, N. Pochet, K. Le Roch, D. Plouffe, M. Kamal, O. Sarr, S. Mboup, O. Ndir, D. Wypij, K. Levasseur, E. Thomas, P. Tamayo, C. Dong, Y. Zhou, E. S. Lander, D. Ndiaye, D. Wirth, E. A. Winzeler, J. P. Mesirov, and A. Regev. Distinct physiological states of plasmodium falciparum in malaria-infected patients. *Nature*, 450(7172):1091–1095, Dec 2007.

[28] Johanna P Daily, Karine G Le Roch, Ousmane Sarr, Daouda Ndiaye, Amanda Lukens, Yingyao Zhou, Omar Ndir, Soulyemane Mboup, Ali Sultan, Elizabeth A Winzeler, and Dyann F Wirth. In vivo transcriptome of plasmodium falciparum reveals overexpression of transcripts that encode surface proteins. *J Infect Dis*, 191(7):1196–1203, Apr 2005.

[29] S. J. Darkin-Rattray, A. M. Gurnett, R. W. Myers, P. M. Dulski, T. M. Crumley, J. J. Allocco, C. Cannova, P. T. Meinke, S. L. Colletti, M. A. Bednarek, S. B. Singh, M. A. Goetz, A. W. Dombrowski, J. D. Polishook, and D. M. Schmatz. Apicidin: a novel antiprotozoal agent that inhibits parasite histone deacetylase. *Proc Natl Acad Sci U S A*, 93(23):13143–13147, Nov 1996.

[30] Herbert M Gilles David A Warrell, editor. *Essential Malariology, 4th Edition*. Arnold, 2002.

[31] K. W. Deitsch, M. S. Calderwood, and T. E. Wellems. Malaria. cooperative silencing elements in var genes. *Nature*, 412(6850):875–876, Aug 2001.

[32] Job Dekker. The three 'c' s of chromosome conformation capture: controls, controls, controls. *Nat Methods*, 3(1):17–21, Jan 2006.

[33] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb 2002.

[34] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, Oct 1997.

[35] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C. Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, May 2010.

[36] Manoj T Duraisingh, Till S Voss, Allison J Marty, Michael F Duffy, Robert T Good, Jennifer K Thompson, Lucio H Freitas-Junior, Artur Scherf, Brendan S Crabb, and Alan F Cowman. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in plasmodium falciparum. *Cell*, 121(1):13–24, Apr 2005.

[37] Ron Dzikowski, Matthias Frank, and Kirk Deitsch. Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. *PLoS Pathog*, 2(3):e22, Mar 2006.

[38] Ron Dzikowski, Matthias Frank, and Kirk Deitsch. Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. *PLoS Pathog*, 2(3):e22, Mar 2006.

[39] Ron Dzikowski, Felomena Li, Borko Amulic, Andrew Eisberg, Matthias Frank, Suchit Patel, Thomas E Wellems, and Kirk W Deitsch. Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites. *EMBO Rep*, 8(10):959–965, Oct 2007.

[40] Cheng Fan, Daniel S Oh, Lodewyk Wessels, Britta Weigelt, Dimitry S A Nuyten, Andrew B Nobel, Laura J van't Veer, and Charles M Perou. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med*, 355(6):560–569, Aug 2006.

[41] A. P. Feinberg and B. Vogelstein. A technique for radiolabeling dna restriction endonuclease fragments to high specific activity. *Anal Biochem*, 132(1):6–13, Jul 1983.

[42] Avi Feller. "temporal estimation in the malaria lifecyce". Master's thesis, Oxford University Department of Statistics, 2008.

[43] Marcelo U Ferreira, Martine Zilversmit, and Gerhard Wunderlic. Origins and evolution of antigenic diversity in malaria parasites. *Curr Mol Med*, 7(6):588–602, Sep 2007.

[44] Paul J. Flory. *Principles of Polymer Chemistry*. Cornell University Press, 1953.

[45] Paul J. Flory. *Statistical Mechanics of Chain Molecules*. Interscience, 1969.

[46] Christian Flueck, Richard Bartfai, Jennifer Volz, Igor Niederwieser, Adriana M Salcedo-Amaya, Blaise T F Alako, Florian Ehlgen, Stuart A Ralph, Alan F Cowman, Zbynek Bozdech, Hendrik G Stunnenberg, and Till S Voss. Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathog*, 5(9):e1000569, Sep 2009.

[47] Matthias Frank, Ron Dzikowski, Borko Amulic, and Kirk Deitsch. Variable switching rates of malaria virulence genes are associated with chromosomal position. *Mol Microbiol*, 64(6):1486–1498, Jun 2007.

[48] Matthias Frank, Laura Kirkman, Daniel Costantini, Sohini Sanyal, Catherine Lavazec, Thomas J Templeton, and Kirk W Deitsch. Frequent recombination events generate diversity within the multi-copy variant antigen gene families of plasmodium falciparum. *Int J Parasitol*, 38(10):1099–1109, Aug 2008.

[49] L. H. Freitas-Junior, E. Bottius, L. A. Pirrit, K. W. Deitsch, C. Scheidig, F. Guinet, U. Nehrbass, T. E. Wellems, and A. Scherf. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of p. falciparum. *Nature*, 407(6807):1018–1022, Oct 2000.

[50] Lucio H Freitas-Junior, Rosaura Hernandez-Rivas, Stuart A Ralph, Dvorak Montiel-Condado, Omar K Ruvalcaba-Salazar, Ana Paola Rojas-Meza, Liliana Mncio-Silva, Ricardo J Leal-Silvestre, Alisson Marques Gontijo, Spencer Shorte, and Artur Scherf. Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell*, 121(1):25–36, Apr 2005.

[51] M. Fried and P. E. Duffy. Adherence of plasmodium falciparum to chondroitin sulfate a in the human placenta. *Science*, 272(5267):1502–1504, Jun 1996.

[52] J. P. Gardner, R. A. Pinches, D. J. Roberts, and C. I. Newbold. Variant antigens and endothelial receptor adhesion in plasmodium falciparum. *Proc Natl Acad Sci U S A*, 93(8):3503–3508, Apr 1996.

[53] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G. Mani Subramanian, Chris Mungall, J. Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511, Oct 2002.

[54] PCC Garnham. *Malaria Parasites and Other Haemosporidia*. Blackwell Scientific, 1966.

[55] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[56] Joseph M Gonzales, Jigar J Patel, Napawan Ponmee, Lei Jiang, Asako Tan, Steven P Maher, Stefan Wuchty, Pradipsinh K Rathod, and Michael T Ferdig. Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol*, 6(9):e238, Sep 2008.

[57] A. Yu. Grosberg, S. K. Nechaev, and E. I. Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys. France*, 49:2095–2100, 1988.

[58] Anita Gndr, Carole Rougier, and Rolf Ohlsson. High-resolution circular chromosome conformation capture assay. *Nat Protoc*, 3(2):303–313, 2008.

[59] Rebecca A Haeusler, Matthew Pratt-Hyatt, Paul D Good, Theresa A Gipson, and David R Engelke. Clustering of yeast trna genes is mediated by specific association of condensin with trna gene transcription complexes. *Genes Dev*, 22(16):2204–2214, Aug 2008.

[60] S. M. Handunnetti, K. N. Mendis, and P. H. David. Antigenic variation of cloned plasmodium fragile in its natural host macaca sinica. sequential appearance of successive variant antigenic types. *J Exp Med*, 165(5):1269–1283, May 1987.

[61] R. E. Hayward, J. L. Derisi, S. Alfadhli, D. C. Kaslow, P. O. Brown, and P. K. Rathod. Shotgun dna microarrays and stage-specific gene expression in plasmodium falciparum malaria. *Mol Microbiol*, 35(1):6–14, Jan 2000.

[62] M. Hommel, P. H. David, and L. D. Oligino. Surface alterations of erythrocytes in plasmodium falciparum malaria. antigenic variation, antigenic diversity, and the role of the spleen. *J Exp Med*, 157(4):1137–1148, Apr 1983.

[63] Paul Horrocks and David Muhia. Pexel/vts: a protein-export motif in erythrocytes infected with malaria parasites. *Trends Parasitol*, 21(9):396–399, Sep 2005.

[64] Paul Horrocks, Robert Pinches, Ze Christodoulou, Sue A Kyes, and Chris I Newbold. Variable var transition rates underlie antigenic variation in malaria. *Proc Natl Acad Sci U S A*, 101(30):11129–11134, Jul 2004.

[65] Paul Horrocks, Robert Pinches, Sue Kyes, Neline Kriek, Sarah Lee, Ze Christodoulou, and Chris I Newbold. Effect of var gene disruption on switching in plasmodium falciparum. *Mol Microbiol*, 45(4):1131–1141, Aug 2002.

[66] Paul Horrocks, Eleanor Wong, Karen Russell, and Richard D Emes. Control of gene expression in plasmodium falciparum - ten years on. *Mol Biochem Parasitol*, 164(1):9–25, Mar 2009.

[67] F. J. Iborra, A. Pombo, D. A. Jackson, and P. R. Cook. Active rna polymerases are localized within discrete transcription "factories' in human nuclei. *J Cell Sci*, 109 (Pt 6):1427–1436, Jun 1996.

[68] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, 31(4):e15, Feb 2003.

[69] Suckjoon Jun and Bela Mulder. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. *Proc Natl Acad Sci U S A*, 103(33):12388–12393, Aug 2006.

[70] Stephan Karl, David Gurarie, Peter A Zimmerman, Charles H King, Tim G St Pierre, and Timothy M E Davis. A sub-microscopic gametocyte reservoir can sustain malaria transmission. *PLoS One*, 6(6):e20805, 2011.

[71] Martin Krzywinski, Jacqueline Schein, Inan Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome Res*, 19(9):1639–1645, Sep 2009.

[72] Werner Kuhn. Uber die gestalt fadenformiger molekule in losungen. *Kolloidzeitschrift*, 68:2, 1934.

[73] M. H. Kuo and C. D. Allis. In vivo cross-linking and immunoprecipitation for studying dynamic protein:dna associations in a chromatin environment. *Methods*, 19(3):425–433, Nov 1999.

[74] Dominic P Kwiatkowski. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet*, 77(2):171–192, Aug 2005.

[75] S. Kyes, P. Horrocks, and C. Newbold. Antigenic variation at the infected red cell surface in malaria. *Annu Rev Microbiol*, 55:673–707, 2001.

[76] Sue Kyes, Ze Christodoulou, Robert Pinches, Neline Kriek, Paul Horrocks, and Chris Newbold. Plasmodium falciparum var gene expression is developmentally controlled at the level of rna polymerase ii-mediated transcription initiation. *Mol Microbiol*, 63(4):1237–1247, Feb 2007.

[77] Sue A Kyes, Zoe Christodoulou, Ahmed Raza, Paul Horrocks, Robert Pinches, J. Alexandra Rowe, and Chris I Newbold. A well-conserved plasmodium falciparum var gene shows an unusual stage-specific transcript pattern. *Mol Microbiol*, 48(5):1339–1348, Jun 2003.

[78] Sue A Kyes, Susan M Kraemer, and Joseph D Smith. Antigenic variation in plasmodium falciparum: gene organization and regulation of the var multigene family. *Eukaryot Cell*, 6(9):1511–1520, Sep 2007.

[79] C. Lambros and J. P. Vanderberg. Synchronization of plasmodium falciparum erythrocytic stages in culture. *J Parasitol*, 65(3):418–420, Jun 1979.

[80] M. Lanzer, D. de Bruin, and J. V. Ravetch. A sequence element associated with the plasmodium falciparum kahrp gene is the site of developmentally regulated protein-dna interactions. *Nucleic Acids Res*, 20(12):3051–3056, Jun 1992.

[81] M. Lanzer, D. de Bruin, and J. V. Ravetch. Transcription mapping of a 100 kb locus of plasmodium falciparum identifies an intergenic region in which transcription terminates and reinitiates. *EMBO J*, 11(5):1949–1955, May 1992.

[82] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct 1999.

[83] J. H. Leech, J. W. Barnwell, L. H. Miller, and R. J. Howard. Identification of a strain-specific malarial antigen exposed on the surface of plasmodium falciparum-infected erythrocytes. *J Exp Med*, 159(6):1567–1575, Jun 1984.

[84] Jacob E Lemieux, Avi Feller, Chris Holmes, and Chris Newbold. Reply to wirth et al.: In vivo profiles show continuous variation between 2 cellular populations. *Proc Natl Acad Sci U S A*, 107:E72, 2009.

[85] Jacob E Lemieux, Natalia Gomez-Escobar, Avi Feller, Celine Carret, Alfred Amambua-Ngwa, Robert Pinches, Felix Day, Sue A Kyes, David J Conway, Chris C Holmes, and Chris I Newbold. Statistical estimation of cell-cycle progression and lineage commitment in plasmodium falciparum reveals a homogeneous pattern of transcription in ex vivo culture. *Proc Natl Acad Sci U S A*, 106(18):7559–7564, May 2009.

[86] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

[87] Manuel Llins, Zbynek Bozdech, Edith D Wong, Alex T Adai, and Joseph L DeRisi. Comparative whole genome transcriptome analysis of three plasmodium falciparum strains. *Nucleic Acids Res*, 34(4):1166–1173, 2006.

[88] Stavros Lomvardas, Gilad Barnea, David J Pisapia, Monica Mendelsohn, Jennifer Kirkland, and Richard Axel. Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2):403–413, Jul 2006.

[89] Jose Juan Lopez-Rubio, Alisson M Gontijo, Marta C Nunes, Neha Issar, Rosaura Hernandez Rivas, and Artur Scherf. 5' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Mol Microbiol*, 66(6):1296–1305, Dec 2007.

[90] Jose-Juan Lopez-Rubio, Liliana Mancio-Silva, and Artur Scherf. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe*, 5(2):179–190, Feb 2009.

[91] Mara Jos Lpez-Barragn, Mariam Quiones, Kairong Cui, Jacob Lemieux, Keji Zhao, and Xin-Zhuan Su. Effect of pcr extension temperature on high-throughput sequencing. *Mol Biochem Parasitol*, 176(1):64–67, Mar 2011.

[92] Margaret J Mackinnon, Tabitha W Mwangi, Robert W Snow, Kevin Marsh, and Thomas N Williams. Heritability of malaria in africa. *PLoS Med*, 2(12):e340, Dec 2005.

[93] C. Magowan, W. Wollish, L. Anderson, and J. Leech. Cytoadherence by plasmodium falciparum-infected erythrocytes is correlated with the expression of a family of variable proteins on infected erythrocytes. *J Exp Med*, 168(4):1307–1320, Oct 1988.

[94] P.C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

[95] C. Ben Mamoun, I. Y. Gluzman, C. Hott, S. K. MacMillan, A. S. Amarakone, D. L. Anderson, J. M. Carlton, J. B. Dame, D. Chakrabarti, R. K. Martin, B. H. Brownstein, and D. E. Goldberg. Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite plasmodium falciparum revealed by microarray analysis. *Mol Microbiol*, 39(1):26–36, Jan 2001.

[96] Liliana Mancio-Silva, Qingfeng Zhang, Christine Scheidig-Benatar, and Artur Scherf. Clustering of dispersed ribosomal dna and its role in gene regulation and chromosome-end associations in malaria parasites. *Proc Natl Acad Sci U S A*, 107(34):15117–15122, Aug 2010.

[97] K. Marsh, D. Forster, C. Waruiru, I. Mwangi, M. Winstanley, V. Marsh, C. Newton, P. Winstanley, P. Warn, and N. Peshu. Indicators of life-threatening malaria in african children. *N Engl J Med*, 332(21):1399–1404, May 1995.

[98] S. A. McLean, C. D. Pearson, and R. S. Phillips. Plasmodium chabaudi: antigenic variation during recrudescent parasitaemias in mice. *Exp Parasitol*, 54(3):296–302, Dec 1982.

[99] L. H. Miller. Distribution of mature trophozoites and schizonts of plasmodium falciparum in the organs of aotus trivirgatus, the night monkey. *Am J Trop Med Hyg*, 18(6):860–865, Nov 1969.

[100] Louis H Miller, Dror I Baruch, Kevin Marsh, and Ogobara K Doumbo. The pathogenic basis of malaria. *Nature*, 415(6872):673–679, Feb 2002.

[101] C. Newbold, P. Warn, G. Black, A. Berendt, A. Craig, B. Snow, M. Msobo, N. Peshu, and K. Marsh. Receptor-specific adhesion and clinical disease in plasmodium falciparum. *Am J Trop Med Hyg*, 57(4):389–398, Oct 1997.

[102] C. I. Newbold, D. B. Boyle, C. C. Smith, and K. N. Brown. Stage specific protein and nucleic acid synthesis during the asexual cycle of the rodent malaria plasmodium chabaudi. *Mol Biochem Parasitol*, 5(1):33–44, Jan 1982.

[103] Z. Ning, A. J. Cox, and J. C. Mullikin. Ssaha: a fast search method for large dna databases. *Genome Res*, 11(10):1725–1729, Oct 2001.

[104] Mayke J A M Oesterholt, Michael Alifrangis, Colin J Sutherland, Sabah A Omar, Patrick Sawa, Christina Howitt, Louis C Gouagna, Robert W Sauerwein, and Teun Bousema. Submicroscopic gametocytes and the transmission of antifolate-resistant plasmodium falciparum in western kenya. *PLoS One*, 4(2):e4364, 2009.

[105] Lucy C Okell, Azra C Ghani, Emily Lyons, and Chris J Drakeley. Submicroscopic infection in plasmodium falciparum-endemic populations: a systematic review and meta-analysis. *J Infect Dis*, 200(10):1509–1517, Nov 2009.

[106] Laura P O'Neill and Bryan M Turner. Immunoprecipitation of native chromatin: Nchip. *Methods*, 31(1):76–82, Sep 2003.

[107] World Health Organization. World malaria report 2010. *WHO Library*, 1:1–238, 2010.

[108] Cameron S Osborne, Lyubomira Chakalova, Karen E Brown, David Carter, Alice Horton, Emmanuel Debrand, Beatriz Goyenechea, Jennifer A Mitchell, Susana Lopes, Wolf Reik, and Peter Fraser. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet*, 36(10):1065–1071, Oct 2004.

[109] Thomas D Otto, Mandy Sanders, Matthew Berriman, and Chris Newbold. Iterative correction of reference nucleotides (icorn) using second generation sequencing technology. *Bioinformatics*, 26(14):1704–1707, Jul 2010.

[110] Thomas D Otto, Daniel Wilinski, Sammy Assefa, Thomas M Keane, Louis R Sarry, Ulrike Bhme, Jacob Lemieux, Bart Barrell, Arnab Pain, Matthew Berriman, Chris Newbold, and Manuel Llins. New insights into the blood-stage transcriptome of plasmodium falciparum using rna-seq. *Mol Microbiol*, 76(1):12–24, Apr 2010.

[111] G. Pasvol, R. J. Wilson, M. E. Smalley, and J. Brown. Separation of viable schizont-infected red cells of plasmodium falciparum from human blood. *Ann Trop Med Parasitol*, 72(1):87–88, Feb 1978.

[112] A. Pombo, D. A. Jackson, M. Hollinshead, Z. Wang, R. G. Roeder, and P. R. Cook. Regional specialization in human nuclei: visualization of discrete sites of transcription by rna polymerase iii. *EMBO J*, 18(8):2241–2253, Apr 1999.

[113] R. N. Price, C. Cassar, A. Brockman, M. Duraisingh, M. van Vugt, N. J. White, F. Nosten, and S. Krishna. The pfmdr1 gene is associated with a multidrug-resistant phenotype in plasmodium falciparum from the western border of thailand. *Antimicrob Agents Chemother*, 43(12):2943–2949, Dec 1999.

[114] Ric N Price, Anne-Catrin Uhlemann, Alan Brockman, Rose McGready, Elizabeth Ashley, Lucy Phaipun, Rina Patel, Kenneth Laing, Sornchai Looareesuwan, Nicholas J White, Franois Nosten, and Sanjeev Krishna. Mefloquine resistance in plasmodium falciparum and increased pfmdr1 gene copy number. *Lancet*, 364(9432):438–447, 2004.

[115] Csar G Prucca, Ileana Slavin, Rodrigo Quiroga, Eliana V Elas, Fernando D Rivero, Alicia Saura, Pedro G Carranza, and Hugo D Lujn. Antigenic variation in giardia lamblia is regulated by rna interference. *Nature*, 456(7223):750–754, Dec 2008.

[116] Karla Prez-Toledo, Ana Paola Rojas-Meza, Liliana Mancio-Silva, Nora Adriana Hernndez-Cuevas, Dulce Maria Delgadillo, Miguel Vargas, Santiago

Martnez-Calvillo, Artur Scherf, and Rosaura Hernandez-Rivas. Plasmodium falciparum heterochromatin protein 1 binds to tri-methylated histone 3 lysine 9 and is linked to mutually exclusive expression of var genes. *Nucleic Acids Res*, 37(8):2596–2606, May 2009.

[117] Stuart A Ralph, Christine Scheidig-Benatar, and Artur Scherf. Antigenic variation in plasmodium falciparum is associated with movement of var loci between subnuclear locations. *Proc Natl Acad Sci U S A*, 102(15):5414–5419, Apr 2005.

[118] J. V. Ravetch, J. Kochan, and M. Perkins. Isolation of the gene for a glycophorin-binding protein implicated in erythrocyte invasion by a malaria parasite. *Science*, 227(4694):1593–1597, Mar 1985.

[119] Mario Recker, Caroline O Buckee, Andrew Serazin, Sue Kyes, Robert Pinches, Ze Christodoulou, Amy L Springer, Sunetra Gupta, and Chris I Newbold. Antigenic variation in plasmodium falciparum malaria involves a highly structured switching pattern. *PLoS Pathog*, 7(3):e1001306, Mar 2011.

[120] Fernando D Rivero, Alicia Saura, Cesar G Prucca, Pedro G Carranza, Alessandro Torri, and Hugo D Lujan. Disruption of antigenic variation is crucial for effective parasite vaccine. *Nat Med*, 16(5):551–7, 1p following 557, May 2010.

[121] D. J. Roberts, A. G. Craig, A. R. Berendt, R. Pinches, G. Nash, K. Marsh, and C. I. Newbold. Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature*, 357(6380):689–692, Jun 1992.

[122] Karine G Le Roch, Yingyao Zhou, Peter L Blair, Muni Grainger, J. Kathleen Moch, J. David Haynes, Patricia De La Vega, Anthony A Holder, Serge Batalov, Daniel J Carucci, and Elizabeth A Winzeler. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639):1503–1508, Sep 2003.

[123] C. D M Rodley, F. Bertels, B. Jones, and J. M. O'Sullivan. Global identification of yeast chromosome interactions using genome conformation capture. *Fungal Genet Biol*, 46(11):879–886, Nov 2009.

[124] Ali Salanti, Trine Staalsoe, Thomas Lavstsen, Anja T R Jensen, M. P Kordai Sowa, David E Arnot, Lars Hviid, and Thor G Theander. Selective upregulation of a single distinctly structured var gene in chondroitin sulphate a-adhering plasmodium falciparum involved in pregnancy-associated malaria. *Mol Microbiol*, 49(1):179–191, Jul 2003.

[125] Adriana M Salcedo-Amaya, Marc A van Driel, Blaise T Alako, Morten B Trelle, Antonia M G van den Elzen, Adrian M Cohen, Eva M Janssen-Megens, Marga van de Vegte-Bolmer, Rebecca R Selzer, A. Leonardo

Iniguez, Roland D Green, Robert W Sauerwein, Ole N Jensen, and Hendrik G Stunnenberg. Dynamic histone h3 epigenome marking during the intraerythrocytic cycle of plasmodium falciparum. *Proc Natl Acad Sci U S A*, 106(24):9655–9660, Jun 2009.

[126] A. Scherf, R. Hernandez-Rivas, P. Buffet, E. Bottius, C. Benatar, B. Pouvelle, J. Gysin, and M. Lanzer. Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in plasmodium falciparum. *EMBO J*, 17(18):5418–5426, Sep 1998.

[127] Artur Scherf, Jose Juan Lopez-Rubio, and Loc Riviere. Antigenic variation in plasmodium falciparum. *Annu Rev Microbiol*, 62:445–470, 2008.

[128] I.J. Schoenberg. Remarks to maurice frechet's article, "sur la definition axiomatique d'une classe d'espace distancies vectoriellement applicable sur l'espace de hilbert". *Annala of Mathematics*, 36(3):724–732, 1935.

[129] Stefan Schoenfelder, Tom Sexton, Lyubomira Chakalova, Nathan F Cope, Alice Horton, Simon Andrews, Sreenivasulu Kurukuti, Jennifer A Mitchell, David Umlauf, Daniela S Dimitrova, Christopher H Eskiw, Yanquan Luo, Chia-Lin Wei, Yijun Ruan, James J Bieker, and Peter Fraser. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*, 42(1):53–61, Jan 2010.

[130] Matthias Scholz and Martin J Fraunholz. A computational model of gene expression reveals early transcriptional events at the subtelomeric regions of the malaria parasite, plasmodium falciparum. *Genome Biol*, 9(5):R88, 2008.

[131] Seif Shekalaghe, Chris Drakeley, Roly Gosling, Arnold Ndaro, Monique van Meegeren, Anders Enevold, Michael Alifrangis, Frank Mosha, Robert Sauerwein, and Teun Bousema. Primaquine clears submicroscopic plasmodium falciparum gametocytes that persist after treatment with sulphadoxine-pyrimethamine and artesunate. *PLoS One*, 2(10):e1023, 2007.

[132] Seif A Shekalaghe, J. Teun Bousema, Karaine K Kunei, Paminus Lushino, Alutu Masokoto, Liselotte R Wolters, Steve Mwakalinga, Frank W Mosha, Robert W Sauerwein, and Chris J Drakeley. Submicroscopic plasmodium falciparum gametocyte carriage is common in an area of low and seasonal transmission in tanzania. *Trop Med Int Health*, 12(4):547–553, Apr 2007.

[133] H. E. SHORTT and P. C C GARNHAM. The exoerythrocytic parasites of plasmodium cynomolgi. *Trans R Soc Trop Med Hyg*, 41(6):705–716, May 1948.

[134] Erandi K De Silva, Andrew R Gehrke, Kellen Olszewski, Ilsa Len, Jasdave S Chahal, Martha L Bulyk, and Manuel Llins. Specific dna-binding

by apicomplexan ap2 transcription factors. *Proc Natl Acad Sci U S A*, 105(24):8393–8398, Jun 2008.

[135] Francesco Silvestrini, Zbynek Bozdech, Alessandra Lanfrancotti, Eugenia Di Giulio, Emanuele Bultrini, Leonardo Picci, Joseph L Derisi, Elisabetta Pizzi, and Pietro Alano. Genome-wide identification of genes upregulated at the onset of gametocytogenesis in plasmodium falciparum. *Mol Biochem Parasitol*, 143(1):100–110, Sep 2005.

[136] D. C. Smith and L. B. Sanford. Laveran's germ: the reception and use of a medical discovery. *Am J Trop Med Hyg*, 34(1):2–20, Jan 1985.

[137] J. D. Smith, C. E. Chitnis, A. G. Craig, D. J. Roberts, D. E. Hudson-Taylor, D. S. Peterson, R. Pinches, C. I. Newbold, and L. H. Miller. Switches in expression of plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell*, 82(1):101–110, Jul 1995.

[138] Robert W Snow, Carlos A Guerra, Abdisalan M Noor, Hla Y Myint, and Simon I Hay. The global distribution of clinical episodes of plasmodium falciparum malaria. *Nature*, 434(7030):214–217, Mar 2005.

[139] Irina Solovei, Moritz Kreysing, Christian Lanctt, Sleyman Ksem, Leo Peichl, Thomas Cremer, Jochen Guck, and Boris Joffe. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell*, 137(2):356–368, Apr 2009.

[140] John D Storey, Wenzhong Xiao, Jeffrey T Leek, Ronald G Tompkins, and Ronald W Davis. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A*, 102(36):12837–12842, Sep 2005.

[141] X. Z. Su, V. M. Heatwole, S. P. Wertheimer, F. Guinet, J. A. Herrfeldt, D. S. Peterson, J. A. Ravetch, and T. E. Wellems. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. *Cell*, 82(1):89–100, Jul 1995.

[142] Heidi Sutherland and Wendy A Bickmore. Transcription factories: gene expression in unions? *Nat Rev Genet*, 10(7):457–466, Jul 2009.

[143] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–6572, May 2002.

[144] Bas Tolhuis, Robert Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, 10(6):1453–1465, Dec 2002.

[145] W. Trager and J. B. Jensen. Human malaria parasites in continuous culture. *Science*, 193(4254):673–675, Aug 1976.

[146] M.V. Volkenstein. *Configurational Statistics of Polymeric Chains*. Interscience, 1963.

[147] T. E. Wellems, L. J. Panton, I. Y. Gluzman, V. E. do Rosario, R. W. Gwadz, A. Walker-Jonah, and D. J. Krogstad. Chloroquine resistance not linked to mdr-like genes in a plasmodium falciparum cross. *Nature*, 345(6272):253–255, May 1990.

[148] Dyann Wirth, Johanna Daily, Elizabeth Winzeler, Jill P Mesirov, and Aviv Regev. In vivo profiles in malaria are consistent with a novel physiological state. *Proc Natl Acad Sci U S A*, 106(27):E70; author reply E71–E70; author reply E72, Jul 2009.

[149] Y. Wu, L. A. Kirkman, and T. E. Wellems. Transformation of plasmodium falciparum malaria parasites by homologous integration of plasmids that confer resistance to pyrimethamine. *Proc Natl Acad Sci U S A*, 93(3):1130–1134, Feb 1996.

[150] Y. Wu, C. D. Sifri, H. H. Lei, X. Z. Su, and T. E. Wellems. Transfection of plasmodium falciparum within human red blood cells. *Proc Natl Acad Sci U S A*, 92(4):973–977, Feb 1995.

[151] Gale Young and A.S. Householder. Discussion of a set of points in terms of the mutual distances. *Psychometrika*, 3:19–22, 1938.

[152] Jason A Young, Quinton L Fivelman, Peter L Blair, Patricia de la Vega, Karine G Le Roch, Yingyao Zhou, Daniel J Carucci, David A Baker, and Elizabeth A Winzeler. The plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol*, 143(1):67–79, Sep 2005.