

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The genome of Plasmodium falciparum

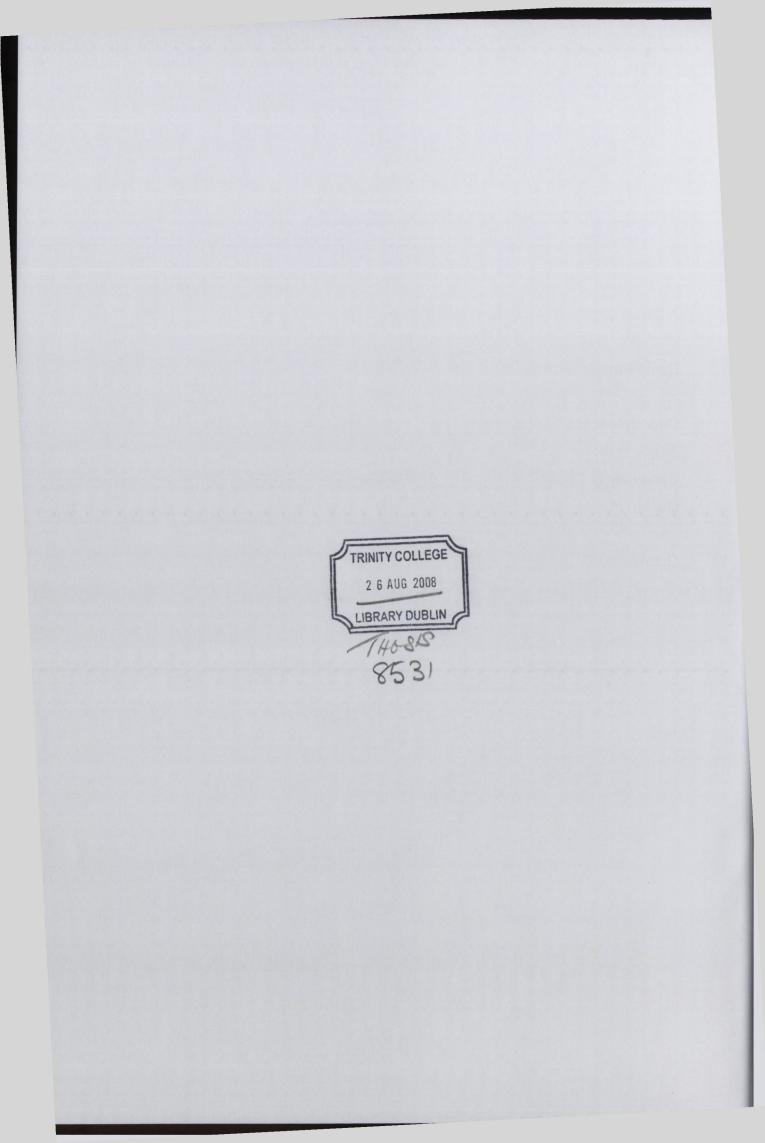
A thesis submitted for the degree of doctor of philosophy by

David Patrick Mitchell

Department of Microbiology Trinity College University of Dublin Dublin

October 2006

i



Declaration

I, David Patrick Mitchell, declare that this thesis has not been submitted as an exercise for a degree at this or any other university.

I further declare that this thesis is entitely my own work with the exception of annotation work on this genome and the development of the gene identification system both of which were carried out in collaboration with Dr. Robert Huestis of Monash University, Melbourne, Australia. This thesis includes some of his unpublished work and his work is acknowledged in this thesis wherever applicable.

The Library may lend or copy this thesis on request.

D-1ittl

David Patrick Mitchell

Summary

The base distribution of the 14 nuclear chromsomes, of the plastid and the mitochondrion of the malarial parasite *Plasmodium falciparum* were studied. Isochores were discovered in the nuclear chromosomes and found to be consistently associated with the presence of erythrocyte membrane protein 1 genes. The centromeres of the nuclear chromosomes were identified by their low GC content. Base distribution within the genome was found to follow a set of mathematical laws which were subsequently found to hold in all genomes.

Serious defects were discovered in the published annotation of the *P. falciparum* genome. These problems were later admitted by the annotators. A novel method of genome annotation is described here and has been used to re annotate this genome. The predictions made were tested against all the published data available to date and found to be in complete agreement. A new method of gene identification was developed which has since been adopted by the Sanger centre. A database was created to investigate this new annotation.

Gene arrangement along the four of the nuclear chromosomes was found to be nonrandom: the reason for this is not known. Genes are on average separated by ~ 1.5 kilobases in the 5' direction and by ~ 300 bases in the 3' direction. In this genome Sybalski's rule - that purines exceed pyrimidines in coding sequences - was found to be the consequence of a linear relationship between complementary bases in the coding sequences.

The frequency of codon occurrence in this genome follows an approximately exponential law. Correspondence analysis of the protein sequences identified the GC content of the first two codon positions and the hydrophobicity of the proteins as the two major trends. Correspondence analysis of the codons revealed only minor trends correlating with the third codon base composition. Base use surrounding the start and termination codons shows features that differ from the bulk of the coding sequences: these seem likely to be involved in translational control.

Intron number per kilobase of chromosome is constant. The AT content of the introns is on average greater than that of the exons. There is no correlation between the GC content of the introns and their surrounding exons. Intron and exon phase distribution is non-random. Evidence of clustering of bases within the introns was found and a putative lariat site for the introns was identified.

Several new biochemical reactions were tentatively identified and their potential as drug targets was discussed. The methods described here are currently being used to explore other genomes.

Acknowledgements

I wish to thank the following people for their assistance.

My supervisor Dr. Angus Bell (Dept. Microbiology, Trinity College, Dublin). My collaborator Dr. Robert Huestis (Dept. Medical Microbiology, Monash University, Melbourne, Australia).

Professor Janet Cox-Singh (Malaria Research Centre, University of Malaysia, Sarawak, Malaysia) for her editorial help.

Dr. Myra Regan (Dept. Statistics, Trinity College, Dublin) for assistance with correspondence analysis.

Dr. Zbynek Bozdech (University of California, San Francisco, USA) for the coordinates of the oligonucleotides.

My mother and father for more reasons than I can list here.

Contents

List of figures	ix
List of tables	xii
Publications arising from this thesis	xiv
Chapter 1: Introduction	1
1.1 Introduction	1
1.2 History of malaria	3
1.3 Epidemiology	6
1.4 Vaccines	10
1.5 Antimalarial chemotherapy	14
1.6 Malaria genome project	21
1.7 Difficulties with the annotation	22
1.8 Organisation of this thesis	22
Chapter 2: Organization of the genome	24
2.1 Introduction	24
2.2 Methods	26
2.3 Results	27
2.4 Discussion	42
2.5 Summary	48
Chapter 3: Annotation methods	49
3.1 Introduction	49
3.2 The MGP annotation methods	49
3.3 Annotation methods used here	55
3.4 Known problems	61
Chapter 4: Gene indentification system	67
4.1 Introduction	67
4.2 Proposed system of gene identification	69
4.3 Discussion	72
Chapter 5: Database files	77
5.1 Introduction	77
5.2 Description of the included files	77
Chapter 6: Gene organization and structure	78

6.1 Summary	78
6.2 Nuclear genome overview	78
6.3 Organisation and spacing of genes	81
6.4 Protein length, predicted pl and hydrophobicity	86
6.5 Nucleotide content of the protein genes	91
Chapter 7: Analysis of codon use	96
7.1 Summary	96
7.2 Methods	98
7.3 Results	100
7.4 Discussion	117
Chapter 8: Intron organization and structure	129
8.1 Summary	129
8.2 Overview	129
8.3 Methods	131
8.4 Results	137
8.5 Discussion	148
Chapter 9: An application of the annotation	152
9.1 Background	152
9.2 Introduction	153
9.3 Methods	157
9.4 Results	157
9.5 Discussion	162
Chapter 10: Summary and future work	164
10. 1 Introduction	164
10.2 Annotation methods	167
10.3 Gene organisation	168
10.4 Analysis of codon use	168
10.5 Intron organisation and structure	170
10.6 An application of the annotation	172
10.7 Conclusions	173
Appendix	175
A.1 Types of errors	175
A.2 Summary of differences between genome	176

versions	
A.3 Probable origin of errors	181
in the P. falciparum genome	
A.4 Examples of errors found	183
A.5 Discussion	189
References	193

List of figures

Number	Title	Page
1.1	Distribution of clinical falciparum malaria	7
2.1	GC content in Chromosome 2	28
2.2	GC content along Chromosome 2	29
2.3	GC content in Chromosome 7	29
2.4	GC content by window in Chromosome 2 plotted against the	31
	GC content of the succeeding window	
2.5	GC skew and position in bases	31
	along the mitochondrial DNA	
2.6	GC skew along the inverted repeat A of the plastid	32
2.7	Position along the inverted repeat B of the plastid	32
2.8	Cumulative GC skew along Chromosome 1	33
2.9	Mean adenosine position and skewness	35
	for nuclear chromosomes	
2.10	Adenosine variance and kurtosis	36
	of the nuclear chromosomes.	
2.11	Number of A and T isostichs in Chromosome 14	36
2.12	Number of A and T isostichs in the mitochondrion	37
2.13	AC and GT tracts in Chromosome 2	37
2.14	Length of the adenosine isostichs (poly A tract)	38
	Against log ₁₀ (number of isostichs) in Chromosome 2	
2.15	Log ₁₀ chromosome length versus intercept	41
	value for short adenosine isostichs	
2.16	Log ₁₀ chromosome length versus	41
	intercept value for long A tracts	
6.1	Distribution of residues per protein	87
6.2	Distribution of log ₁₀ (protein length)	88
6.3	Distribution of the predicted pl values	88
6.4	The numbers of basic and acidic residues	89
6.5	Distribution of the hydrophobicity	89
	(Kyte Doolittle units) per residue	

6.6	Weak (A - T) and strong (G - C) Chargaff	92
	differences in the protein encoding genes	
6.7	Distribution of the normalised	93
	strong (G - C) Chargaff differences	
6.8	Distribution of the normalised	93
	weak (A - T) Chargaff differences	
6.9	A versus T content of the protein encoding genes	94
6.10	G versus C content of the protein encoding genes	94
7.1	Percentage T content in the third codon position	102
7.2	Correlation of hydrophobicity and (A2 – T2)	102
7.3	GC3 and hydrophobicity	103
7.4	Correlation of GC3 and (A2 – T2)	103
7.5	Index and percentage of encoded amino acids	105
7.6	Correlation of the index and log ₁₀ (percentage)	105
7.7	Index of codon use and frequency of codon use	106
7.8	Skree plot for the proteins	106
7.9	Protein plot with respect to the first two axes	108
7.10	Codon plot with respect to the first two axes	109
7.11	Plot of the effective number of codons (N _c)	110
	and GC3s in <i>P. falciparum</i>	
7.12	Effective number of codons and protein length	110
7.13	Correlation of the codon chi square and protein length	111
7.14	Base use in the first codon position	112
7.15	Base use in the second codon position	113
7.16	Base use in the third codon position	114
7.17	Mean base use in the first codon position for	115
	the last 50 codons 5' of the termination codon	
7.18	Base use in the second codon position for	115
	the last 50 codons 5' of the termination codon.	
7.19	Base use in the third codon position for	116
	the last 50 codons 5' of the termination codon	
7.20	T3 content and distance from the termination codon	117
8.1	Number of introns per kilobase of chromosome	131

8.2	Distribution of log ₁₀ (intron lengths)	138
8.3	Distribution of log ₁₀ (exon lengths)	138
8.4	Relationship between the number	139
	of introns and gene length	
8.5	Distribution of intron adenosine content (percentage)	140
8.6	Relationship between intron GC content	140
	and five prime exon GC3 content	
8.7	Relationship between intronic and	141
	three prime exon GC3 content	
8.8	Relationship between five and	141
	three prime exon GC3 content	
8.9	Mean AT content surrounding the splice sites	141
8.10	Relationship between intronic A and T content	142
8.11	Relationship between intron	142
	adenosine and guanine content	
8.12	S values relative to the 5' splice site	147
8.13	S values relative to the 3' splice site	147

List of tables

Number	Title	Page
1.1	Drugs used to treat malaria and	
	their dates of introduction	18
2.1	Length (base pairs) of the P. falciparum chromosomes	25
2.2	Approximate centromere centres	30
2.3	Chargaff's second rule in P. falciparum	34
2.4	Mean, variance, skewness and	
	kurtosis of the adenosine bases	35
2.5	Number of AC and GT tracts in Chromosome 2	38
2.6	Number by length of adenosine	
	Isostichs in Chromosome 2	39
2.7	Coefficient and intercept from regression of	
	observed and expected short adenosine isostichs	40
2.8	Coefficient and intercept from regression of	
	observed and expected long adenosine isostichs	40
4.1	Proposed vocabulary for annotation notes	72
5.1	Values used to determine the relative molecular	
	weight, predicted pl and hydrophobicity of the proteins	76
6.1	Gene content by chromosome	80
6.2	Mean intergenic distances (standard deviation)	
	in base pairs, observed numbers of runs and	
	expected numbers of runs (standard deviation)	83
6.3	Mean intergenic distances (standard deviation) in base	
	pairs organised by chromosome and gene order	83
6.4	Mimimum integenic distances in base pairs	84
7.1	Codon composition in <i>P. falciparum</i>	
	by base and position	101
7.2	Encoded amino acid (percentages)	
	within the P. falciparum genome	104
7.3	Residue values in the first three axes	107
7.4	Base use preceding the initial ATG codon	111

7.5	Percentages of genes grouped by -3/+4 bases	111
7.6	Stop codons by +4 base as percentages	
	of all protein encoding genes	114
8.1	Number of introns per chromosome	130
8.2	Bases lying immediately five prime	
	of the GT splice site	142
8.3	Bases lying immediately three prime	
	of the AG splice site	143
8.4	Exons by phase distribution	143
8.5	Means and standard deviations of	
	the expected and observed run lengths	145
9.1	Partial listing of PEST +ve proteins	158 -161
A.1	Comparison of GenBank MGP genes 2002, 2005 and	176 - 177
	2005 with the current annotation of genes	
A.2	Comparison of GenBank MGP exons 2002, 2004 and	
	2005 with the current annotation of exons	178
A.3	Comparison of PlasmoDB 2005 MGP genes with the	
	current annotation of genes	179
A.4	Comparison of PlasmoDB 2005 MGP exons with the	
	current annotation of exons	180

Publications arising from this thesis

Mitchell D. and A. Bell (2003) PEST sequences in the malaria parasite *Plasmodium falciparum*: a genomic study. Malar. J. 2:16.

Mitchell D. and R. Bridge (2006) A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. 340(1):90-94.

Mitchell D. and R. Bridge (2006) An investigation of genomic base distribution Biochem. Biophys. Res. Commun. 344(2):612-616.

Abstracts

Seven posters and one oral presentation based on parts of this work have been presented in at scientific meetings in the UK, Ireland and Australia.

Other materials

An earlier version of the annotation can be found at the WHO/TDR malaria database site. The web links are:

http://www.wehi.edu.au/MalDB-www/who.html

http://www.wehi.edu.au/MalDB-www/genome.htm

The latter site has a downloadable poster at the 'VBC annotation methods' link describing some of this work.

Two additional papers have been submitted and the referees' reports are awaited.

"And I saw, and behold, a pale horse: and he that sat upon him, his name was Death; and Hell followed after him. And there was given unto them authority over the fourth part of the earth, to kill with the sword, and with famine, and with death" (Revelations 6:8)

Chapter 1: Introduction

1.1 Introduction

Today 7000 people died from malaria: half were children under five (465, 466). Two hundred times this number were infected. Tomorrow is not expected to be different.

Man is the host to hosts to nearly 300 species of parasitic worms and over 70 species of protozoa - some of which derive from parasites of our primate ancestors and some that have been acquired from the animals we have domesticated or come in contact with during our short history (94). The most important of these is malaria which claims the lives of more children worldwide than any other infectious disease (55, 99, 248, 407). This disease is the greatest mass murder in history and most of its victims have been children. In comparison to it legendary murderers of Biblical proportions - Herod's slaughter of the innocents - pale into insignificance. It has been a scourge throughout history and has killed more people than all wars and other plagues combined.

Since 1900 - when malaria accounted globally for one death in ten (74) - the area of the world exposed to malaria has been halved and malaria now accounts for only one death per thousand. Within this time the number of persons exposed has risen by two billion. In part because of this rise in the human population within the last 35 years the incidence of malaria has risen 2 - 3 fold (191) and it remains a threat in 103 countries (194, 291).

Malaria - while found worldwide - is predominantly an African disease. 80% of the cases and 90% of deaths attributable to malaria occur in Africa. 5% of all African children die of malaria and the most common age of death is 4 years old. In endemic areas up to 50% of children are chronically infected with malaria as are as many as 60% of pregnant women (356).

With malaria morbidity as well as mortality is substantial. Chronic infection has been shown to reduce school scores by up to 15% (133, 230) an effect that has been known

for decades: the physician and poet Leipoldt in his capacity as Medical Inspector of Schools wrote in 1921 'School children in the Transvaal infected with chronic malaria were mentally classed as 'feeble-minded' (458). More recent work has shown significant reductions in tests of comprehension, syntax, pragmatics, word finding, memory, attention, behaviour and motor skills in children who suffer from epilepsy as a result of cerebral malaria when compared with age matched controls (73).

Africa is the poorest continent and uniquely it has become poorer rather than richer over the last three decades. War, lack of democracy and the rule of law and poor social policies are responsible for the bulk of this. Disease none the less has played a part and none more so than malaria. The economic burden of malaria control is considerable: Tanzania presently spends over 1% of its GNP (US\$2.2 per capita) - 39% of its health expenditure - on malaria control alone (227). Reduction in the incidence of malaria has been shown to coincide with increased economic output. Estimates have been made suggesting that reducing malaria by 10 per cent in countries severely affected by the disease would stimulate economic growth by 0.3% while its eradication would increase economic growth in these countries by 3.2% a year (154, 417). While these gains may seem modest when compared with rapid economic growth in some countries, a 0.3% improvement in the economy of the countries most affected by malaria where economic growth has been static or declining for decades would be considerable.

Malaria is a febrile illness caused by infection with protozoa of the genus *Plasmodium*. One hundred and seventy five species are currently recognised (31) and at least five species infect man naturally: *P. falciparum* - which alone accounts for 80% of the cases and 90% of the deaths – *P. vivax* – about 10% cases – *P. ovale*, *P. malaria* and *P. knowelsi* (454). Other species – *P. cynomolgi*, *P. semiovale*, *P. brazilianum*, *P. inui*, *P. rodhaini* and *P. schwetzi* - are known to infect man at least occasionally (530). Multiple simultaneous infections are not uncommon and occur in up to 25% of cases (379).

Because of the importance of malaria, attempts have been made since the discovery of the causative organism to culture the parasites but the difficulties have been considerable. *In vitro* culture of *P. falciparum* was first described for the erythrocytic cycle in 1976 (506) and for the mosquito stages in 1993 (449, 526). Genetic transformation - an invaluable tool in research in molecular biology - was first

successfully reported in 1996 (552). Successful in *vitro* culture has yet to be achieved for the other *Plasmodium* species.

In addition there have been considerable difficulties with the control and treatment of malaria itself - and in particular with that caused by *P. falciparum*.

1.2 History of malaria

The earliest known evidence of malaria comes from Egyptian mummies (circa 2000 BC) in whom the spleen is enlarged, a finding highly suggestive of malaria (79). The Ebers Papyrus (1534 BC), the oldest known Egyptian medical text, describes the symptoms of malaria such as enlarged spleen, fever and cites ancient remedies for malaria. Cuneiform writings on clay tablets found in the library of Ashurbanipal (c 690-626 BC) mention periodic and deadly fevers similar to malaria and ascribe them to Nergal, the Sumero-Babylonian god of the netherworld, destruction and pestilence. Malaria may also have been described in the Chinese medical classic Nei Ching, a text written for the Emperor Huang Ti about 2700 BC.

More recognisable descriptions of the clinical features of malaria are found in the writings of the physician Hippocrates who classified the fever types as quotidian (daily fever), tertian (fever on alternate days) and quartan (fever three days apart) as well as noting the poor prognosis associated with the respiratory pattern later described by the physician Kussmaul, professor of medicine at Strasburg (201, 253). Hippocrates also noticed that those who drank the stagnant marsh water had large stiff spleens - a characteristic of this disease - and that fatal dropsy was common among them (200). Furthermore he noted an association with rainy weather but does not appear to have connected this with mosquitoes. Carus (94 - 55 BC), the poet and Epicuran philosopher, suggested in his *De Rerum Natura* published in 50 BC that swamp fever might be caused by infection with a living organism. Columella, a Roman living in Spain in first century AD, suggested in his *De Re Rustica* a connection between fevers with flies. Unfortunately Galen (131 - 210), physician to Roman Emperor Marcus Aurelius (121 - 180), disagreed and held that malaria was due to internal causes. His views held sway almost undisputedly until the sixteenth

century. The parasite has been shown to be present in London in 450 AD (419) and is likely to have been widespread throughout the Roman Empire at this time.

The name malaria meaning bad air (*mal aria*: Italian) comes from the linkage suggested by Lancisi in 1717 linking malaria with the poisonous vapours of swamps. He also suggested the draining of swamps and a possible connection with mosquitoes. This term 'malaria' was later introduced into English in 1740 by Horace Walpole, the author and son of the Prime Minister Sir Robert Walpole, after he had read letters from Italy describing the disease. Rasori a physician living in Parma suggested - with remarkable foresight - the cause of the characteristic paroxysms of malaria. 'For many years,' he wrote in 1816, 'I have held the opinion that the intermittent fevers are produced by parasites which renew the paroxysm by the act of their reproduction which recurs more or less rapidly according to the nature of the species.' Sklotovki in 1871 suggested that malaria was due to an organism that lived in the blood and in 1879 Afanasiev suggested that the dark pigmented lesions found in the blood of malarious patients might the causative agent but did not realise these were parasites (3).

The organism itself was finally seen by a French military physician Laveran, professor of military diseases and epidemics at the École de Val-de-Grâce, on Saturday evening November 6 1880 at a military hospital in Constantine, Algeria when he discovered a microgametocyte exflagelating in the blood of a soldier who had been in Algeria for about a year (263). Laveran subsequently found this organism in 148 out of 192 malaria patients that he examined. Acceptance of this new theory was not immediate and it was not until 1886 that the non bacterial origin of malaria was generally accepted. Welch the professor of pathology at Johns Hopkins, proposed the name *Haematozoon falciparum* for the parasite that Laveran had discovered because this species had the distinct property of forming crescents.

Most of the species infecting humans were soon identified. Golgi in 1885 discovered schizonts and recognised that there were at least two new species on the basis of the different schizogeny and he named these '*vivax*' and '*malariae*.' *P. falciparum* was described independently by Sakharov in 1889 and in 1890 by Marchiafava and Celli the director of the Institute of Hygiene at University of Rome. Marchiafava and Celli also proposed the name 'Plasmodium' in their paper – Laveran having earlier suggested *Oscillaria malariae*. *P. ovale* was described in 1922 by Stephens at the

Liverpool School of Tropical Medicine (476). *P. knowlesi*, first described in India (1931) as a parasite of a long-tailed macaque *Macaca fascicularis*, has been recognised as human pathogen only recently (453).

The asexual life cycle of the parasites was also described by Golgi (1886) and gametocyes and the process of fertilization were described in 1897 by Opie and MacCallum while they were still medical students at Johns Hopkins (363). These latter authors realised that exflagellation was part of the process of fertilization and not a death spasm as previously believed. Romanowsky, a physician in St Petersburg (Russia) described a staining technique that is still in use today (135, 403).

The possibility of an association between malaria and mosquitoes has been recognised for millennia in India – possibly as early as 2000 BC. King took up this theme again in 1883. This was reiterated by Laveran in 1884 and again by Manson (1844 - 1922) in 1894 The hypothesis was finally experimentally confirmed independently by two physicians – Grassi working with *P. falciparum* in Italy and Ross with *P. relictum* in Secunderabad (India) both in 1898 (410). Grassi proposed the existence of an exoerythrocytic stage in the life cycle (172): this was later confirmed by Short, Garnham, Covell and Shute who found *P. vivax* in the human liver (445, 446). *P. knowlesi* was described in 1933 (459) and *P. cynomolgi*, a facultative human pathogen, in 1935 (329). The liver forms of the then known human species were identified over the next decade. That the liver forms could remain infectious for years (hypnozoites) was proposed by in 1982 by Krotoski *et al* (246), a suggestion that is now accepted dogma.

The nomenclature of the parasites for some time was confused. In 1892, Grassi and Feletti, as an honour to Laveran, proposed the genus name *Laverania* (174). Sergent suggested in 1929 the name *Plasmodium praecox* (173) for *P. falciparum* and in 1935 Giovannola proposed that *P. ovale* and *P. vivax* were the same species (165). Only in 1954 at the meeting of the International Commission on Zoological Nomenclature was the present nomenclature agreed.

1.3 Epidemiology

Malaria is transmitted by mosquitoes and all the known human parasites are spread by members of the genus *Anopheles*. Malaria can only survive when there is sufficient contact between humans and female mosquitoes - a hypothesis first proposed by Ross (409). While other mechanisms of transmission are known – blood transfusion, needle stick and organ transplant (29, 58, 545, 549) - these cases are negligible in number compared with the mosquito borne infections.

Approximately one hundred species of *Anopheles* are known to be competent for human malaria transmission and ~80% of human cases are transmitted by members of the *Anopheles gambiae* complex (89). The reproductive capacity of mosquitoes is considerable with each female mosquito laying 200 to 1000 eggs. Depending on the area, as many as five different anopheline species may transmit malaria (139).

Malaria parasites cease to develop in the mosquito when the temperature is below 16° C or above 30° C. *P. falciparum* sporozoites, the infective mosquito stage, can only develop at temperatures above 18° C. Optimal conditions for parasite development occur when the mean temperature is within the range $20-30^{\circ}$ C and the humidity is ~60%. Environmental conditions lying within these parameters are presently found largely in the tropical and sub tropical regions. Inoculation rates vary from almost none to more than a 1000 infective bites per year. Transmission can occur throughout the year or only during a couple of months (362) and heterogeneities are observed between years within the same locale.

While endemic malaria is currently confined to tropical and subtropical parts of the world (Figure 1.1), this was not always so (116): the last indigenous case of malaria in the UK occurred in 1953 and in Australia in 1981. Dublin University itself - and the Houses of Parliament in the UK - are sited on what were once malarious swamps.

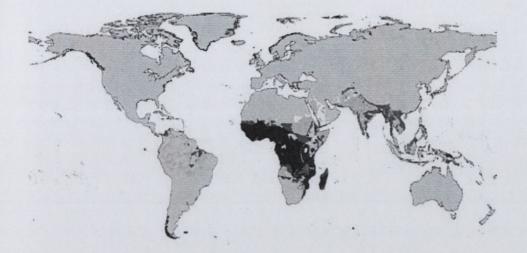


FIG. 1.1. Distribution of clinical falciparum malaria (464). The more heavily shaded areas have the greater prevalence.

Malaria had been a problem in the United States from earliest records and a major epidemic swept the country in the 18th century reaching as far north as Montreal. It was a major source of casualties in the civil war and endemic in the southern states until the 1930's. During the nineteenth century the decline of malaria seen throughout Europe was due to new agricultural practices and changed social conditions. The disappearance of the disease in Europe and North America was due more to those changed conditions than to the use of DDT and other residual insecticides (107).

Non tropical but less developed countries continue to suffer from malaria. Throughout in 1970s and 1980s, the USSR suffered outbreaks of malaria in Azerbaijan and Tadjikistan. Turkey, an EU accession candidate country, is still malarious. Between the years 1950 and 2000, 21 outbreaks of introduced malaria, all caused by *Plasmodium vivax*, have occurred in the US. Fourteen of these occurred in California. Four outbreaks - one each in 1986 and 1988 and two in 1989 - occurred in San Diego County. In the 1986 outbreak 27 cases were identified, 2 in local residents and 25 in a group of mostly Mexican migrant workers. The outbreak in 1988 resulted in 30 cases of malaria being diagnosed, 28 in migrant agricultural workers and 2 in local residents. *Anopheles freeborni*, a competent vector of malaria, was identified as the responsible vector.

Imported malaria remains a problem: there are 30,000 malaria cases are imported into industrialized countries each year (415) with 1500 cases a year in the US alone (339) There have been >52,000 cases in the UK alone since 1953 (251) and an outbreak

occurred recently in Australia (188). Currently there are ~2000 imported cases each year in the UK while malaria appears to be still indigenous in the US albeit at low level (56). Refuge resettlements in addition are presenting problems to countries long unused to malaria (336).

The last 50 years have seen a reduction in malaria exposure: the percentage of the human population at risk has fallen from 80% in 1950 to the current 40% (536) - and this despite an increase in size from 2.5 billion (1950) to 6.1 billion (2000). This reduction in malarial infection has coincided with an unprecedented growth in urbanization throughout the world: 14% in 1900; 30% in 1950; and 47% in 2000. Today there are 40 cities in Africa whose population exceeds one million: 330 (39%) of Africa's 850 million people live in cities. By 2030, 800 million (54%) are expected to be urban dwellers (195). Habitats available in urban environments are lacking in diversity compared to rural areas and relatively few species are able to exploit them. With important exceptions, anopheline malarial vectors have not succeeded in adapting to urban life but malaria can still be a problem where rural pockets exist within towns. Also when compared with rural settings, malaria control in urban areas is easier for a number of reasons - both cultural and practical (113, 182, 345, 355).

Insecticide use has been invaluable in malaria eradication and control. DDT (dichlorodiphenyl-trichloroethane) - a compound originally synthesized in 1874 and subsequently discovered to be usable as an insecticide by the J.R. Geigy chemist Muller in 1939 - was used to free an estimated one billion people (then a quarter of the world population) from malaria. The US began its eradication effort in 1947 by providing \$7 million to a 5 year campaign using indoor residual spraying of DDT. By 1952, the United States was malaria free and the program ended. Encouraged by the success of the US eradication effort the WHO began an 8 year effort in 1957 to eradicate malaria worldwide. In Sri Lanka reported malaria cases fell from an average of 3 million a year in the 1940's to 50 in 1963. Concerns with the use of DDT were raised in the mid 1940's: by the 1970s' its use except for malaria control has almost ceased (229). With the end of the eradication effort, the incidence of malaria skyrocketed. In 1975 China recorded 9 million cases (1963) to 6 million (1975). Worldwide the number of cases in 1975 was two and a half times that of 1961. Resistance to DDT persists in the mosquito species and has proven to be a problem for malaria control (190, 456).

As *A. arabiensis* feeds on cattle as well as humans the use of insecticides on cattle has been explored (183). Insecticides are poured on a monthly basis onto the backs of cattle to control tsetse and ticks. Mosquitos feed principally on the legs of cattle and the effective duration of the insecticide there is only one week. While this remains an option for mosquito control, cost and logistics make this difficult to implement on a wide scale basis.

Use of mosquito repellents is popular but rarely has been properly evaluated. Recent studies suggested these agents may offer limited protection at best (110, 412). The use of impregnated bed nets at night have been advocated for some time and appear to be cost effective at US\$25-38 per person per year (550). Acceptance of their use remains disappointingly low (9) in spite of trials showing a reduction in the death rate in children of 20% (17). Permethrin is the most commonly use insecticide in nets but 'knockdown' resistance (*kdr*) to this insecticide is known particularly in West Africa. The mechanism is thought to be a mutation in a voltage-gated sodium channel gene on chromosome 2 but additional genes may play a role (126, 343, 385). Other insecticides are under investigation (128).

In addition to insecticides, genetic manipulation of the mosquito is under investigation but to date there has been limited success (398). Given the difficulties encountered in bed net use and mosquito control present hopes are largely focused on the development of vaccines and new drugs. Previous work has shown that malaria is unlikely to be eradicated or even controlled with single approaches and that combined intervention with insecticides, bed nets, drainage, oiling of potable water, habitat alteration, case identification and treatment can reduce both its prevalence and incidence (45, 67, 235, 274, 324, 357).

Recently an association has been recognized between outbreaks of malaria and the *El Nino* current, a cold ocean current which appears along the coast of Chile around Christmas time. *El Nino* originates in the Southern Pacific at intervals of 2 - 7 years. This current, first noticed in 1957 by Bjerknes a meteorologist, has wide-ranging consequences for weather around the world including worldwide droughts and floods (40). There is a strong correlation between the occurrence of this current and malaria epidemics in parts of South Asia, South America and Africa, an effect thought to be

mediated by an increased number of mosquitoes (50-52, 241, 273, 372). *El Nino* may also result in lower than expected numbers of cases, presumably again via an effect on the size of the mosquito populations (275). This data is used to model malaria outbreaks (95, 208). The predictions have implications for public heath policies in affected countries.

1.4 Vaccines

It has been estimated that, given expected parasite replication rates during acute falciparum malaria in a human, all available erythrocytes would be invaded resulting in fatality within 10–12 days following appearance of parasites in the blood. In clinical practice, the parasitemia at this time averages only 0.1%, suggesting that innate defense mechanisms contribute to parasite control and host survival before acquired immunity develops (319). In addition to this non-specific immunity repeated infection of an individual with malaria leads to the gradual acquisition of what has since been termed 'naturally acquired immunity'(19). This is a short-lived, partially strain-specific immunity that leads initially to a reduction in death rate and incidence of complications, then with increasing age to a decreased incidence of clinical malaria and ultimately results in the suppression of parasitaemia to low or undetectable levels (160, 486).

Because of its medical importance a malarial vaccine is one of the most sought after goals in the biological sciences. Recognition of this partial immunity gave the initial impetus to the vaccine development. The first trials with non-human primate malarial parasites vaccines began in 1910 when it was hoped that these species could induce cross immunity in a fashion similar to that which had been so successful in viruses.

Almost a century later prospects of a vaccine still appear slim although a recent study has demonstrated partial immunity and reduction in disease severity in children vaccinated with part of a recombinant sporozoite coat protein (RTS,S/ASO2A) (10) The trial involved 2,000 children less than one year old in Mozambique and the results showed \sim 30% protection against infection over the study period. An earlier trial with the same vaccine in adult males had shown considerably less promise (43). This vaccine is the first other than the earlier irradiated sporozoite trial to show useful levels of protection but there is considerable work to do to improve its efficacy to the 90%+ range that will be needed.

Pregnancy in particular presents unique problems for vaccine development. Pregnancy is associated with changes in the immune system (533). It has been suggested that the placenta is in some way 'immunologically naïve', that a specific adaptive immune response in the utero-placental circulation was required and that repeated exposure to the parasite as gestation advances and subsequent pregnancies make these local responses more effective (303). A general suppressive effect has also been suggested (386). A more recent suggestion is that the parasite produces a particular set of *var* genes that are positively selected in pregnancy and that these genes bind to pregnancy specific proteins in the placenta (148, 396). The production of immunity here seems to depend in some fashion on gravidity (347, 351). Evidence also suggests that malaria during pregnancy down regulates the ability of the child to mount a protective response later in life (59).

Evidence that a malaria vaccine is at least potentially possible has been recognised for decades. Passive immunization with immunoglobulins from adults residing in endemic areas gave some degree of protection (416). More promisingly immunization with irradiated sporozoites gave protection for 6 to 10 months (65, 397) but the use of the pooled irradiated sporozoites from thousands of mosquitoes appears to most observers impractical to scale up – although some wish to make this attempt (63). The use of immunoglobulins - aside from the practical difficulties of collecting, storing and distributing this material – is likewise not scalable. Newer techniques – monoclonal antibodies and *in vitro* cultivation of sporozoites – might eventually make these methods viable but this also seems unlikely at present. Current interest resides principally in the use of malarial proteins derived from cloned genes.

Genes from *P. falciparum* were first cloned in 1985 (473). A number of proteins expressed at various stages of the parasite's life cycle are under active consideration as vaccine components including: the pre-erythrocytic proteins - circumsporozoite protein (CSP) and thrombospondin related anonymous protein (TRAP) - for the protection of the uninfected traveller; hepatic stage proteins - liver stage antigen (LSA) 1 (252), LSA 3 and exported antigen 1 (EXP 1) - aimed at boosting cell mediated immunity; the erythrocytic stage proteins aimed at both the intracellular forms - glutamate rich protein (GLURP) - and the merozoite - merozoite surface

protein (MSP) 1, MSP 2, MSP 3 and apical merozoite antigen 1 (AMA 1) and the transmission blocking vaccines (Pfs25) aimed at the mosquito stages (20, 163, 308, 321). There is evidence that these proteins are involved in the immune response (381). Adjuvants based on cloned genes are under active consideration. Interestingly vaccination with DNA plasmids is showing promise (525)(167, 238, 249, 524, 525).

Development of a vaccine for malaria is hindered by the response of the host immune system to malaria (169). Infection has a profound effect on the immune system and frequently induces immune suppression. Dendritic cells represent a vital link between the innate and adaptive immune response to infection and are responsible for presenting antigen, activating naive T cells and enhancing antibody production (142, 277). Dendritic cell maturation is characterized by increased expression of major histocompatibility complex (MHC) class II and co-stimulatory molecules, antigen processing and presentation, and secretion of cytokines that activate immune cells. These cells suffer a maturation defect following interaction with infected erythrocytes and become unable to induce protective liver-stage immunity - an effect that appears to be mediated by parasitized cells binding to CD36, a surface expressed glycoprotein (170, 348, 371, 510-512). The effects on the dendritic cells are likely to be significant because of their importance as antigen presenting cells - a factor crucial to an appropriate immune response to the parasite in the initial immune response to malaria (57). Despite this induced defect these cells make a significant contribution to immunity to malaria (261, 371).

Glycosylphosphatidylinositol (GPI) anchors on the surface of *P. falciparum* infected cells are powerful stumulators of the immune response. Macrophage surface phospholipases play an important role in the clearance of these molecules (244). The Toll like receptors (TLR) types 2 (assisted by the accessory TLRs 1 and 6) and 4 recognise the GPIs as ligand. The further pathways of this reaction involve the MyD88-dependent activation of the ERK, JNK and p38 and NF-kappaB signaling pathways.

Haemozoin (malarial pigment) in spite of being a potent proinflamatory agent (224), known to inhibit macrophage function (231, 290, 341, 460), possibly via the induction of IL-10 (111) and may influence immunoglobulin isotype switching (86). It is known to act as Toll 9 receptor ligand and its effects on chemokine release may be abrogated by chloroquine (87). Other increases in the mRNA levels of various chemokines have

been reported including MIP-1alpha/CCL3, MIP-1beta/CCL4, MIP-2/CXCL2, and MCP-1/CCL2. Haemozoin is also known to affect ERK1/2 phosphorylation, NF-kappaB activation, reactive oxygen species (ROS) generation and ROS-dependent protein-tyrosine phosphatase down-regulation (225).

Macrophage and neurophil chemotaxis is inhibited by the presence of P. falciparum cells (225). Complement, in animal models, plays some role in immunity (498) and levels fall (indicating consumption) during an attack of malaria and to rise again subsequently (448). Infected erythrocytes directly adhere to and activate peripheral blood B cells from nonimmune donors and are abnormally resistant to haemolysis. A protein, translationally controlled tumour, is found in the serum of infected patients and affects basophil responses (286). The interleukin (IL)-12/IL-10 secretion pattern becomes inverted (348) and the presence of uninfected erythrocytes impairs immune recognition of malarial antigens (487). Other complex changes occur in the cytokine levels (285). The var gene products bind Fab and Fc fragments of human immunoglobulins in a fashion similar to protein A of Staphylococcus aureus and these genes, like protein A, are polyclonal stimulators of B lymphocyte (118). This non specific elevation of immunoglobulin has long been known and it is thought that this contributes to the increased incidence of bacterial infections as in multiple myeloma. The circumsporozoite protein binds the complement fragment C3d at it C-terminal flanking sequence and this suppressing the induction of antibodies specific for the Cterminal flanking sequence and the induction of circumsporozoite protein specific IL-4-producing spleen cells (32).

The total parasite load affects the response of the T cells (392). Concomitant infections are also common: infection with schistosomes, HIV and filarial may increase susceptibility to or the severity of malaria (171, 234, 467). Hepatitis B and tuberculosis also appear to be associated with malaria (2) as does a poor response to immunization with meningococcal vaccines (54). Autoantibodies to the central nervous tissue are produced in natural infections in malaria (259) so vaccine induced encephalitis is a theoretical possibility.

The main existing problem with vaccines is that of the blood stages. Existing vaccines can reduce the numbers of parasites by over 90% at the liver stage (28). The human response to vaccines may also differ considerably: the Fulani, Mossi, and Rimaibe are three tribes found northeast of Ouagadougou, Burkina Faso which is hyperendemic

for malaria. The Mossi and Rimaibe are Sudanese negroid populations with a long tradition of sedentary farming and the Fulani are nomadic pastoralists. There is a considerable interethnic difference in the response to malaria with the Fulani being considerably less susceptible (317). This effect seems to be cumulative with additional protective measures: use of bed nets reduced malaria by 18% in the Mossi and Rimaibe but by 42.8% in the Fulani (318). Given the difference in the immune response rates between two sympatic tribes alone, vaccine efficacy may prove to be highly variable.

1.5 Antimalarial chemotherapy

Antimalarial drugs have been employed for centuries. Alcohol and later opium were commonly used to suppress the rigors of the first stage of the malarial paroxysm but they do not reduce the parasite load.

Between 1620 and 1630 the Jesuit missionaries learned of the use of the use of an extract from a particular tree as a treatment for fever while in Loxa (Peru). The natives called the material *quina quina*. The fourth Countess Chinchon and the second wife of the new viceroy who had just arrived from Europe, was taken ill with tertian malaria at Lima in 1638. The Count used an infusion of this bark at the suggestion of his physician del Vego as a last resort to cure his wife. The trees were later classified under the genus *Chinchona* by the physician Linnaeus, professor of medicine at Uppsala, in 1742 (204).

Cinchona bark was brought from Lima to Spain and from there to Rome and other parts of Italy by the Jesuit procurator de Cobo in 1632. Count Chinchon and de Vega brought back additional supplies in 1640. This new medicine was first discussed in European medical literature in *Discours et Aris sur les flus de ventre* (1643) by the Dutch physician Heyden in Antwerp. Thompson (1650) introduced this "Jesuits' bark" to England and its first recorded use there was by Dr Metford of Northhampton in 1656. The Protestant English physicians refused to contemplate the use of this medicine: Oliver Cromwell died in 1658 from malaria after refusing to take the "powder of the devil."

Nonetheless Talbor an English apothecary's apprentice in London, pioneered the use of cinchona bark in the treatment of malaria in England. His secret remedy cured many sufferers in the Fens and Essex marshes before it was administered to King Charles II (1630 – 1685) and the ailing son of France's King Louis XIV (1638 -1715) in 1679. Talbor received an honorary knighthood and was appointed Royal Physician. It was not until after Talbor's death in 1681 that the Louis – with Talbor's prior consent – disclosed the secret remedy. Morton (1696) presented the first detailed description of the clinical picture of malaria and of its treatment with cinchona bark infusion but it was as late as 1775 when Torti wrote that quinine cured only a limited number of fevers - effectively recognising a unique aetiology for malaria (505).

During the eighteenth century demand for bark soared leading to the destruction of 25,000 trees per year by 1795 and threatening survival of the genus (40 species). This demand was not eased when the British launched an attack on the French Emperor Bonaparte in Holland in 1809 which in turn lead to the worst outbreak of malaria in European history. The Bolivian authorities had realized the value of these trees and forbidden their export. In 1865 Ledger and his servant Manuel succeeded in smuggling a pound of cinchona seeds which the Dutch government bought for \$20. With these the Dutch grew 12,000 high-potency trees in Java and dominated the market (97% share) until World War II. Gize (1816) studied the extraction of crystalline quinine from the cinchona bark and the professors of pharmacy in Paris, Pelletier and Caventou, succeeded in extracting two pure quinine alkaloids which they named quinine and cinchonine (365).

These drugs changed the face of history. Quinine allowed the Europeans to colonize the tropics and operate plantations and mines using workers transported from India and China where there was a surplus of labor. This new surplus of manpower can in part be attributed to quinine with population increases resulting from effective malaria treatment in these countries. In spite of the passage of time, intravenous quinine remains the treatment of choice for complicated malaria.

The effect on quinine on protozoa was first noted by the pharmacologist Binz who while working in Bonn discovered in 1867 that quinine was highly toxic to microscopic organisms found in impure water. Attempts to make synthetic antimalarials began in earnest in 1891. The Germans during World War 1 were cut off from the supply of quinine and were unable to supply their colonies or troops in

Africa. The synthesis of antimalarials began in Germany with pamaquine in 1924, primaguine in 1926 (221) and Atebrin (mepacrine) in 1928. This latter compound was developed by Mietzsch, Mauss and Hecht (461). The US alone evaluated 14,000 potential compounds for anitmalarial activity during World War 2 before settling on Atebrin. This drug was used widely throughout the Pacific but it proved deeply unpopular because of the yellowing of the skin it caused. The Germans also developed plasmoquine and resochin (1934) and later sontochin and this latter compound went into widespread use in the North African campaigns (461). Sontochin fell into the hands of the Americans after they captured Tunis in 1943 and with some modifications this drug subsequently became chloroquine. This agent was first thought to be too toxic for use in humans, a view that since has changed. Other drugs developed at this time included plasmoquine (synthesized in 1928 by Schuleman, Schonhofer and Wingler), acriquine (in 1930 by Knunyants and Chelintsev), proguanil (in 1944 by Curd, Davey and Rose). Mao Tse-tung encouraged his scientists to find new antimalarials after seeing the casualties in the Vietnam War. Artemisinin, a sesquiterpene lactone endoperoxide isolated from Artemesia annua (also known as sweet wormwood or Qinghao) was discovered in 1971 based on a medicine described in China in 340AD by the physician Ge Hon (197, 209). This drug became known to Western scientists in the late 1980s and is now recommended as a standard treatment for malaria.

The range of drugs used to treat malaria is limited (Table 1.1) and parental treatment is commonly required (30 - 40%) (61). Initial treatment failure is common – the median overall failure rate is 10% and it ranges from 0 - 47%: much of this is due to resistance (478). With the exception of artemisinin and its derivatives, resistant strains now are widespread. Chloroquine resistant *Plasmodium falciparum*, first described in 1957 along the Thai-Cambodian border, now predominates in Southeast Asia, South America and increasingly in Africa (1). Quinine resistance had been reported in South America before this but was already much less widely used than chloroquine. In 1973 chloroquine finally was replaced by the combination of sulphadoxine and pyrimethamine (FansidarTM) as first line drug for the treatment of uncomplicated malaria in Thailand, a policy that was later adopted by 10 African countries. Resistance, first noted in the 1960s, is now widespread in Asia (463) and South America and is currently spreading in Africa (349, 521, 546).

Drugs in clinical use	Introduced
Quinine	1620s
Primaquine	1926
Chloroquine	1945
Amodiaquine	1950s
Mefloquine	1975 (507)
Halofantrine	1982 (93)
Piperaquine.	1982 (82)
Proguanil	1944
Chlorproguanil	1945
Trimethoprim	1967 (296)
Trimethoprim and sulphonomides	1968 (295)
Pyrimethamine and dapsone	1960s (42)
Sulfadoxine and pyrimethamine	1960s
Chlorproguanil and dapsone	1988 (531)
Atovaquone	1992
Proguanil and atovaquone	1996 (284)
Artemisinin	1970s
Dihydroartemisinin and piperaquine	2002 (104,
	109)
Doxycycline	1970s
Rifampicin	1970s

Quinine and clindamycin	1974 (309)
Rifampicin, isoniazid, sulfamethoxazole and trimethoprim	1995 (145)
Fosmidomycin	2002 (311)
Fosmidomycin and clindamycin	2004 (47)

TABLE 1.1: Drugs used to treat malaria and their dates of introduction

Mefloquine was introduced into clinical practice along Thai-Cambodian border in 1983 in a trial of 60,000 patients. It officially replaced Fansidar for uncomplicated malaria in Thailand the following year. Resistance was described in 1990.

Halofantrine resistance was reported in 1985, only three years after its introduction (92). Pyrimethamine with dapsone is also no longer recommended because of widespread resistance. Artesunate was introduced in Thailand in 1995 for the treatment of falciparum malaria in areas of multidrug resistance, where it is used in combination with mefloquine. Artemisinin resistance has to date has been described only in the laboratory but resistance may be arising already (376).

Quinine and the other alkaloids kill the parasites by crossing the erythrocyte and parasite membranes and accumulating in the acidic digestive vacuole where they prevent detoxication of haematin produced during haemoglobin breakdown by inhibiting its dimerization (527). Chloroquine is known to inhibit the parasite's lactate dehydrogenase but it is not clear how important this mechanism is in killing the parasite (390). Quinine resistance is known to exist but its mechanism is not yet understood (131, 239).

Chloroquine resistance has two known mechanisms: increased export from the cells (560) and failure to transport the alkaloid into the digestive vacuole (75) This latter mechanism arises via point mutations causing loss of function mutations (Met74Ile, Asn75Asp/Gln, Lys76Thr, Ala220Ser and Arg371Ile)(121, 310) in a putative transport molecule (298, 560). This transporter protein is a dimeric integral membrane protein whose C terminus is predicted to be located within the cytoplasm. Choloroquine resistance can be reversed with the use of chlorpheniramine, a

histamine H_1 receptor antagonist, which is to be safe to use in pregnancy (469) and by the antivirial drug amantadine (535). The molecular mechanisms are not known.

The mechanism of mefloquine resistance is not known but it is thought that this maybe be due to an increase in the copy number of a transmembrane transport molecule - the multiple drug resistance gene (mdr) whose product - P-glycoprotein homolog 1 - pumps the drug away from its active site (377).

Trimethoprim, pyrimethamine, chlorproguanil and the other inhibitors of the folate pathway act on the dihydrofolate reductase gene or the dihydropteroate synthase genes: resistance here is due to point mutations in these genes (492) - 16Val , 51Ile, 59Arg and 108The being the usual suspects - and appears to have spread from South East Asia to Africa in the recent past (192, 406). Atovaquone inhibits respiration in the parasite by specifically binding to the ubiquinol oxidation site at center P of the cytochrome bc(1) complex (233). Resistance has been reported due to point mutations (Tyr268Asn and Tyr268Ser) within this gene and at least one other as yet unidentified mechanism (327, 539).

Rifampicin inhibits RNA polymerase by binding to its rpoB subunit and resistance arises rapidly through point mutations (81). Deoxycycline inhibits protein synthesis by preventing the binding of aminoacyl-tRNA molecules to the 30S ribosomal subunit (13). In bacteria resistance to this drug is usually due to an efflux pump (267). Resistance has been found in malaria but the mechanism is not yet known (380). Clindamycin causes dissociation of peptidyl-tRNA from the 50s subunit of the ribosome (501) and resistance was reported soon after its introduction (185). Isoniazid has been used for decades to treat tuberculosis and acts by inhibiting enoyl-acyl carrier protein reductase (120). Resistance to its action appears to arise via mutations in the catalase gene *katG* (69). While isoniazid has been used in malarial clinical trials (166) attention is presently focused on another inhibitor of the same enzyme – triclosan (322). Fosmidomycin acts through inhibition of 1-deoxy-D-xylulose 5-phosphate reductoisomerase (540) and resistance in bacteria appears to be via an efflux mechanism (153).

Artemisinin (and its derivatives) are active against both the asexual and gametocyte stages but has no effect on the hepatic forms. The molecular basis of their action is not yet clear but it is likely there is more than one such mechanism. Artemisinin is concentrated in the digestive vacuole and may interfere with haemozoin formation, a

process that involves the histidine rich protein 2 (228). It also induces abnormalities in the parasite's vacuolar network (8). It is thought that these drugs undergo reductive cleavage by ferrous iron to oxy radicals and other reactive compounds and that it is these radicals that kill the parasite by damaging its membrane (370). It has also been shown that artemisinin inhibits an ATPase in the endoplasmic reticulum of the parasite after activation by iron (122). Artemisinin also binds to the translationally controlled tumor protein but the effect of this action are not known (37).

The use of artemisinins in pregnancy is still under investigation. Known side effects in rats and rabbits include abortions, cardiac malformations and skeletal defects (84). In spite of this than more than 100 first trimester and 600 second and third trimester human cases of malaria in pregnancy have been documented with no known adverse effects.

The single greatest current difficulty with arstemisinin is its cost: at US \$2.40 per treatment course, artemisinin combination therapy is 10 - 20 times more expensive than other existing drugs. People in malaria endemic countries survive on less than US \$15 per month: the typical budget available for antimalarial treatment is US \$1 per person and presently only chloroquine and sulfadoxine-pyrimethamine are widely available at this price. Artemisinin is obtained from plants grown primarily on Chinese and Vietnamese farms which have not kept up with demand. Artemisia farms are now springing up in India and the WHO is currently supporting experiments to grow the plants in east Africa. Over the last decade, Keasling and his colleagues at the University of California, Berkeley have cloned nine genes into Escherichia coli bacteria to make them produce terpenoids, a class of molecules that includes artemisinin. With a few genes borrowed from Artemisia, they hope to be able to produce an artemisinin precursor. In addition a new artemisinin-like drug OZ277 which has been synthesized by Vennerstrom and his colleagues at the University of Nebraska, Omaha (518) and a phase 1 safety trial has just been completed and an India pharmaceutical company (Ranbaxy) is taking this further. Other drugs are also under investigation (7, 541).

The rapidity with which the parasites - and *P. falciparum* in particular - develop resistance is a matter of some concern and it has been suggested that these mutations may arise through errors generated by DNA polymerase (509). Poor compliance with drugs regimes also contributes to this problem (137). Because of the rapidity of the

appearance of resistance, it is advocated that combination therapy, particularly with artemisins, be used wherever possible (243, 472, 480, 481, 538). Unfortunately short courses of even multidrug combinations may not be entirely successful (247) and the need for prolonged treatment is associated with poor compliance. Curiously resistant strains may be replaced by sensitive ones relatively rapidly after the use of drug is discontinued due to the effect resistance mutations on fitness have - possibly as much as 25% and at least some of this is due to an effect on the merozoites (196). This may be of use in public health programmes (262, 265).

1.6 Malaria genome project

In 1990 the Human Genome project was created with the intention of sequencing the human genome by 2005. In 1995 Venter and his colleagues published the sequence of the genome of *Haemophilus influenzae* (136). This paper outlined a new method of relatively rapid genome sequencing and assembly. This paper changed the face of molecular biology as it showed that it was possible to sequence the genomes of entire organisms much more rapidly than had previously been thought. Given the medical importance of *P. falciparum*, the problems with *in vitro* work, the difficulties of vaccine development and the rapidity of the appearance of drug resistance *P. falciparum* was one of the first eukaryotes chosen for sequencing (12). In 1995 a consortium - the malaria genome project (MGP) - was created to sequence the nuclear genome of *P. falciparum* - which began its work in 1996.

Before this project started it was known that the genome was organised into 14 nuclear chromosomes and one chromosome each for the two organelles – a mitochondrion and a plastid. The nuclear chromosomes were known to be linear and to posses a repeated element at either end (358). Fewer than thirty genes had been cloned before this project began (424) and most of these either lacked introns (202, 408, 442, 450) or possessed a single intron (25, 361). It was not known how representative these genes were of the entire genome. Gene expression not unexpectedly was known to vary during the life cycle and emphasis had been placed on antigenically variable genes (213, 214) presumably on the grounds of seeking conserved regions of the encoded proteins that might be prove to be of use as potential vaccine candidates. The prospects for successive vaccine development

seemed poor given the considerable variation and diversity known to exist in antigenic genes (96).

The sequence of the *P. falciparum* mitochondrial genome had been reported in 1992 (130) and that of the plastid in 1993 (157, 547). The sequence of the first chromosome completed (Chromosome 2) was reported in 1998 (156). The sequence of Chromosome 3 was reported in 1999 and the entire genome in 2002 (158, 186, 215). The ~24 megabase genome is extremely AT rich (~80%) and just over 5300 genes were described.

1.7 Difficulties with the annotation

Within a year of the publication of chromosome 2 its annotation was being questioned. Errors, including in frame stop codons, were also later found. In 2001 a revision of Chromosome 2 with considerable experimental confirmation was published using a new method of largely manual annotation (212). Given the success of the results of chromosome 2, it was decided to attempt independently to annotate the entire genome. Considerable progress to this goal had been made by the time the MGP published their findings.

The MGP relied almost entirely on computer analysis for gene prediction (Berriman, 2002, personal communication). After its publication a number of problems were noted in the annotation. These have been since confirmed by the MGP and an effort to correct the mistakes in the published *P. falciparum* genome – a curation project - has recently been announced (35). In this paper the MGP have stated that only half *the P. falciparum* chromosomes were completely sequenced at the time of publication and there were errors that should not been missed. This paper was submitted shortly after a presentation was given at the British Society for Parasitology based on some of the work presented here.

1.8 Organisation of this thesis

The sequence of the *P. falciparum* genome was published in 2002 but the organisation of the bases within the genome does not seem to have previously been studied. This is

the subject of Chapter 2. The annotation of the genome was carried out in conjunction with a second Ph. D. student – Dr. Robert Huestis. An analysis of the programmes used by the MGP and some of the problems with their chosen approach is given in Chapter 3. Also in this chapter are given the methods used for the annotation used here. Detailed comparisons between the two annotations have not been presented here as this work is presented in Dr Huestis's thesis. Chapter 4 describes a new method of gene identification which was developed jointly by the current author and Dr Huestis. Chapter 5 documents the files used in this work. Chapter 6 examines a number of the properties of the genes in *P. falciparum*, while Chapters 7 and 8 examine the properties of the codons and the introns respectively of the genes. Chapter 10 gives a general discussion and describes some future work.

Chapter 2: Organization of the genome

2.1 Introduction

The *P. falciparum* genome is organized into 14 nuclear chromosomes that vary in length from 650,000 bases (650 kilobases) to 3.3 megabases (3.3 million bases) for a genome total of 22.8 megabases. The mitochondrion and the plastid each contain one chromosome respectively of 6 kilobases and 28 kilobases length (Table 2.1). These two smaller chromosomes had been sequenced before the malaria genome project (MGP) was established in 1995. Chromosome 12 was sequenced by Stanford University, USA; chromosomes 1, 3 - 9 and 13 by the Sanger Center, UK (Sanger); and chromosomes 2, 10, 11 and 14 by the Institute for Genomic Research (Tigr), USA. The three centers relied on a shotgun approach to sequencing but differed in their approach to assembly. While Stanford used a system of overlapping contigs, Sanger and Tigr preferred to use software to assemble the sequenced contigs directly. Chromosomes 5 - 8 are very similar in size and as it was not possible to separate reliably these chromosomes on gels before sequencing, this group was termed 'the blob'. Because of these difficulties these chromosomes were cloned and sequenced as a whole which made the subsequent assembly a considerable challenge.

Because the high AT content made discrimination between contigs difficult, assembly in *P. falciparum* was difficult even outside the blob. Scholars working with ancient manuscripts have created a terminology for difficulties of this nature: contamination – an extraneous element introduced from elsewhere: homeoteleuton – an error in the copy confusing two sequences that have the significant similarity at the end: and homeoarchy – an error in the copy confusing two sequences that have the significant similarity at the beginning. All these problems are likely to have afflicted the MGP.

When the MGP published their results in 2002 (158, 186, 216) approximately 1000 contigs had not yet been allocated to a chromosome. Huestis has reduced this number to 59 containing 125 genes by recognizing that the differences between otherwise identical sequences were located in the terminal one hundred or so bases (Huestis personal communication, 2003). The introduction of a small numbers of errors in the ends of long sequences is a common problem with automated sequencing. The existence of these unassembled contigs was omitted in the 2002 papers but has since

been confirmed in a later publication (34). As a result it is now known that the unmapped contigs belong to chromosomes 4, 6 - 8 and 13. It is not known when the assembly will finally be completed.

Mitochondrion	5 949
Plastid	28 426
Chr 1	643 292
Chr 2	947 102
Chr 3	1 060 087
Chr 4	1 204 112
Chr 5	1 343 552
Chr6	1 377 956
Chr 7	1 350 452
Chr 8	1 323 182
Chr 9	1 541 723
Chr10	1 694 360
Chr 11	2 035 238
Chr 12	2 271 483
Chr 13	2 747 326
Chr 14	3 290 837
Total	22 865 041

TABLE 2.1. Length (base pairs) of the P. falciparum chromosomes

Consequently any hypothesis requiring as an assumption that the current assembly is correct should be treated with suspicion. Even with this proviso the available experimental evidence suggests that the assembly while not completely accurate is close to being correct.

While much of the interest in genomes understandably lies in the genes themselves the organization of the chromosomes is itself also of interest. The remainder of this chapter will discuss various facets of their organization and structure. The annotation methods employed in this project and the findings in the genes will be discussed in subsequent chapters.

2.2 Methods

Isochores are long (usually >30 kilobases) regions of DNA with a constant GC content. The mitochondrion and the plastid were considered too short for meaningful isochore analysis. The distribution of the GC content in the nuclear chromosomes was examined in non overlapping windows of 2 kilobases. This size used here was an arbitrary choice. While it is small compared with those in vertebrates it seemed more suitable for these shorter chromosomes. The use of non overlapping windows avoided the problems that have been identified with alternative methods.

DNA is replicated by the insertion of complementary bases one at a time and the nucleotides used are thought to be obtained from a pool within the cytoplasm. Given the nature of this process it seems probable that regions lying close to each other may have a similar GC content. This possibility was examined by regression of the GC content of each window on the succeeding windows. The regression analysis was done with Microsoft's Excel.

In 1968 Chargaff and his colleagues discovered a rule in *Bacillus subtilis*: in single stranded DNA, A = T and C = G. Lobry later created a set of statistics - the GC content and the 'GC' and 'AT skew' of the sequence - to measure deviations from this rule (279). The formulae used were:

GC content = (G + C)/ N SD (GC content) = (1 / N) (SW / N)^{1/2} GC skew = (G - C) / (G + C) SD(GC skew) = (2 / S) (CG / S)^{1/2}

where W = A + T, S = C + G, N = W + S and SD is the standard deviation. These formulae were applied to the GC content in the chromosomes with N (the window size) equal to 2000 giving 0.05% as the standard deviation of GC content between the windows. The GC skew statistics were determined for the plastid, mitochondrion and the nuclear chromosomes with Lobry's formulae to seek possible origins of DNA replication. Since the raw skew plots may be difficult to interpret, Lobry suggested that a cumulative skew plot would be easier to understand. Since almost nothing is known about DNA replication in *P. falciparum* the cumulative GC skew plots were computed. The location of each base within the genome was determined and the mean, standard deviation, skewness and kurtosis of their distribution were also computed. These latter statistics were computed with from the standard formulae with a programme specially written for the task.

Chargaff proposed a third parity rule: oligonucleotides should be present in equal numbers to their complements within the same strand. This rule was also tested on the *P. falciparum* genome with a variety of oligonucleotides.

Cursory examination of any of the nuclear chromosomes shows that blocks of adenosine and thymidine nucleotides are a common finding. Chargaff coined the name 'isostich' (*iso* = identical, *stichion* = element) for these sequences. The length of an isostich is the number of bases within it. Dechering *et al* (108) earlier investigated the relationship between the number of isostichs and their length and proposed a formula:

 $p_k = p_A^k (1 - p_A)^2 + p_T^k (1 - p_T)^2$

where p_k is the expected proportion of isostichs of length k within the sequence and p_A and p_T are the proportions of A and T respectively in the sequence. Here p_A and p_T are both approximately 0.4. These predicted values were also determined here for each chromosome.

The actual number of adenosine and thymidine isostichs of varying lengths were determined and plotted. The plots of the isostich number-length values suggested a log-linear relationship with a break at ~7-10 units (Figure 2.14). To identify more accurately the point where the total squared error was minimal - the knot - two regression lines were fitted. Isostich lengths for the first regression were varied from 4-10 bases and from 5 to the maximum length in the second. In all chromosomes the knot was located at 8 +/- 1 bases. The isostichs were then divided into short (length \leq 8) and long (length >8) groups. Separate regressions were fitted to each (Table 2.7 and 2.8).

2.3 Results

The existence of isochores in *Plasmodium* has not been investigated previously. The distribution of the GC content here was bimodal (Figure 2.1) with the bulk of the sequences having a GC content between 12 - 28% and a small fraction possessing

higher values (>30%). The mean nuclear chromosome GC content is 19.6%. Given that the difference between the high GC and low GC regions is approximately 20% Lobry's test's confirms that there are at least two types of intrachromosomal region – one with low and the other with high GC content.

The nuclear chromosomes are organized into three distinct regions - here termed isochores: both ends of the chromosomes have a GC content of 30 - 35% and the body has a GC content of 10 - 30%. In the body of some of the chromosomes the GC content rises above the usual background level (Figure 2.2). In all cases this was associated with the presence of an internal group of *var* (variant antigen) genes.

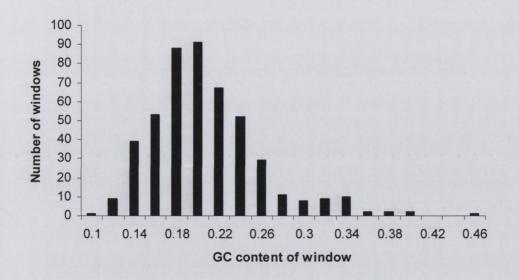


FIG. 2.1. GC content in Chromosome 2. Window size was 2000 bases. Two peaks are visible: one centered on a GC content = 0.2 and second much smaller one around GC = 0.34.

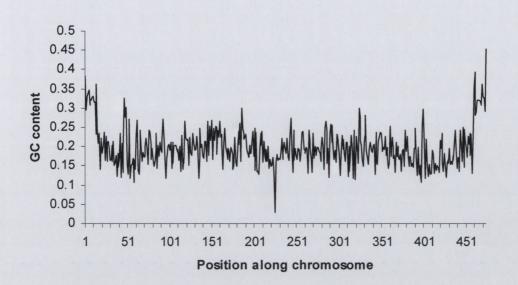


FIG. 2.2 GC content along Chromosome 2. Window size was 2000 bases. At either end of the chromosome are the GC rich isochores. The centromere near the middle is visible.

Close to the centre of the chromosomes and lying within a gene poor region, troughs in the GC content are present (Table 2.2). These troughs coincide with sequences believed to be the centromeres. One of these troughs can be seen in Figure 2.3.

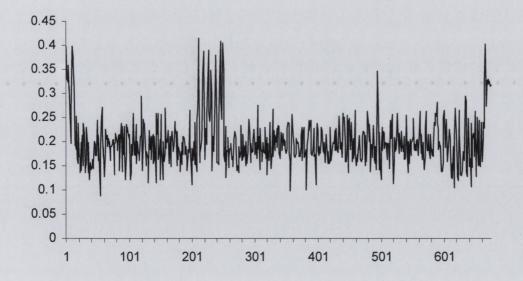


FIG. 2.3. GC content in Chromosome 7. Window size was 2000 bases. The GC spike around window 220 is associated with the presence of an internal *var* gene cluster. The putative centromere is located near window 400.

Chr	Position
1	460 200
2	449 000
3	595 700
4	650 200
5	456 400
6	479 700
7	865 700
8	301 400
9	1 243 400
10	937 100
11	830 800
12	1 283 600
13	1 169 500
14	1 072 400

TABLE 2.2 Approximate centromere centres. Abbreviation: chr - chromosome

The GC content of each window was highly correlated with the succeeding ones with the correlation extended as far as the next ten windows (20,000 bases). The regression is heteroscedastic (Figure 2.4) so conclusions drawn from it need to be examined carefully. If this effect is genuine it is purely a local one: 20 kilobases represents < 3.5% of even the smallest of the chromosomes (chromosome 1).

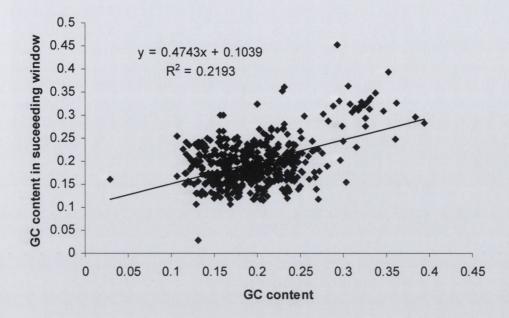
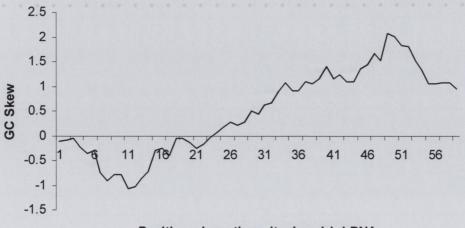


FIG. 2.4. GC content by window in Chromosome 2 plotted against the GC content of the succeeding window. Window size was 2000 bases. (F = 125.0, p < 10^{-23})

A change in the sign of the GC skew was evident (Figure 2.5) in the mitochondrion around base 2000. The mitochondrial cytochrome c oxidase subunit 1 (EC 1.9.3.1) is located between bases 2035 to 3471. It seems probable that the origin of replication for the mitochondrion lies to 5' to this gene but this awaits experimental verification.



Position along the mitochondrial DNA

FIG 2.5. GC skew and position in bases along the mitochondrial DNA. The window size was 100 bases.

A change in the cumulative GC skew occurs in the A inverted repeat (IR-A) of the plastid around base 1600 (Figure 2.6). The lsu rRNA large subunit ribosomal RNA is located between bases 2335 to 5116. The GC skew change found here is close to the experimentally known origin of DNA replication. Within the inverted repeat B (IR-B) no such change in the cumulative skew was found (Figure 2.7) consistent with the single known origin of DNA replication.

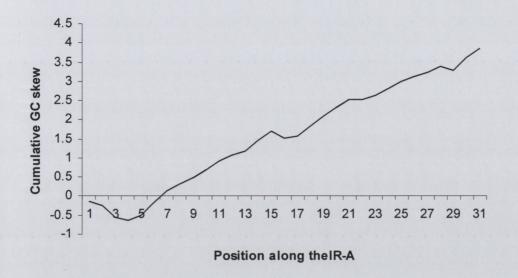


FIG. 2.6. GC skew along the inverted repeat A of the plastid. The window size was 500 bases. The ticks on the X axis are multiples of 500 bases.

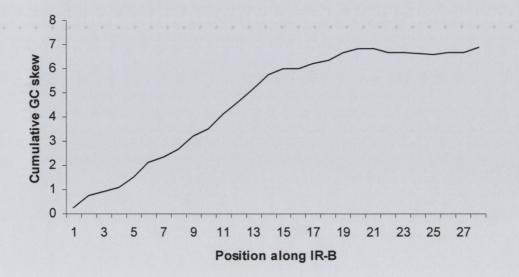


FIG. 2.7. Position along the inverted repeat B of the plastid. Window size was 500 bases. The ticks are units of 500 bases.

In chromosome 1, the GC skew curve changes sign around base 586 000 near which are located two pseudogenes and two hypothetical proteins (Figure 2.8). The cumulative GC skew changes sign in several chromosomes: around base 120 000 of chromosome 5 in which region there lie a number of hypothetical proteins: around base 1 950 000 of chromosome 11 which lies within the giant erythrocyte membrane associated protein antigen 332 (Pf332): around base 2 200 000 in chromosome 12 within a group of *rifin* and *stevor* genes: around base 2 600 000 of chromosome 13 close to a number of hypothetical proteins. The *stevor* and *rifin* genes are two closely related groups of genes that like the *var* genes are variable, antigenic and surface exposed. Whether this apparent association of antigenic variability and GC skew is of biological importance remains to be investigated.

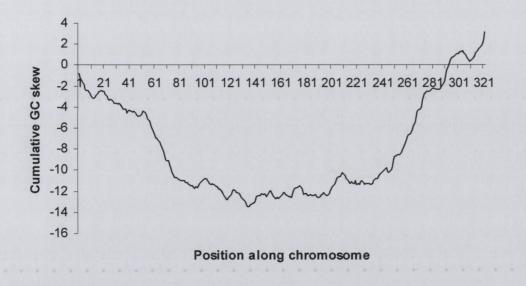


FIG. 2.8. Cumulative GC skew along Chromosome 1.

Chargaff's second rule can be seen to hold for the nuclear chromosomes (Table 2.3) As expected this rule does not hold for either the plastid or the mitochondrion: the organellar rule (A + G = 0.5) holds reasonably well. The relative excess of purines in the plastid may reflect its greater gene density.

With the exception of chromosome 13, base distribution within the chromosomes is typical of the eukaryotes (314). The mean was approximately constant at half the length with other statistics also being approximately constant. The values for

adensosine are shown in Table 2.4. The typical mean-skewness and variance-kurtosis plots can be seen Figures 2.9 and 2.10

	Α	Т	G	С
Mitochondrion	32.4	36.0	15.9	15.7
Plastid	37.6	32.4	17.9	15.1
Chr 1	40.6	38.9	10.4	10.1
Chr 2	40.1	40.2	9.8	9.9
Chr 3	39.9	40.2	10.0	9.9
Chr 4	39.2	40.1	10.0	10.7
Chr 5	40.9	39.8	9.8	9.5
Chr 6	39.9	40.4	9.7	10.0
Chr 7	39.5	40.5	9.8	10.1
Chr 8	40.1	40.2	9.8	9.9
Chr 9	40.6	40.4	9.4	9.6
Chr10	39.9	40.4	9.9	9.8
Chr 11	40.7	40.4	9.6	9.4
Chr 12	40.6	40.1	9.7	9.6
Chr 13	40.6	40.2	9.7	9.5
Chr 14	40.7	40.9	9.3	9.2

TABLE 2.3. Chargaff's second rule in *P. falciparum*. The percentages may not equal100 because of rounding.

The plot of the means against the respective skewness and the variance and kurtosis showed the expected negative linear correlation. In both plots Chromosome 13 was identifiable as an outlier. This chromosome was known to be incomplete at the time of publication since PlasmoDB contained an unmapped contig marked as chr13_2: this is may be the cause of its anomalous behavior seen here.

Chr	Mean A	Var A	Skew A	Kurt A
1	0.501	0.282	0.002	1.821
2	0.505	0.285	-0.028	1.815

30.4990.2860.0081.80540.5000.287-0.0021.79350.5000.2860.0101.80560.5020.2870.0041.79570.5020.289-0.0101.785849.40.2850.0311.818950.00.286-0.0031.8091049.80.2890.0141.7811150.00.2870.0031.7991250.20.287-0.0501.7471450.10.286-0.0051.810					
50.5000.2860.0101.80560.5020.2870.0041.79570.5020.289-0.0101.785849.40.2850.0311.818950.00.286-0.0031.8091049.80.2890.0141.7811150.00.2870.0031.7991250.20.288-0.0111.7991350.20.287-0.0501.747	3	0.499	0.286	0.008	1.805
60.5020.2870.0041.79570.5020.289-0.0101.785849.40.2850.0311.818950.00.286-0.0031.8091049.80.2890.0141.7811150.00.2870.0031.7991250.20.288-0.0111.7991350.20.287-0.0501.747	4	0.500	0.287	-0.002	1.793
70.5020.289-0.0101.785849.40.2850.0311.818950.00.286-0.0031.8091049.80.2890.0141.7811150.00.2870.0031.7991250.20.288-0.0111.7991350.20.287-0.0501.747	5	0.500	0.286	0.010	1.805
8 49.4 0.285 0.031 1.818 9 50.0 0.286 -0.003 1.809 1.809 10 49.8 0.289 0.014 1.781 11 50.0 0.287 0.003 1.799 12 50.2 0.288 -0.011 1.799 13 50.2 0.287 -0.050 1.747	6	0.502	0.287	0.004	1.795
9 50.0 0.286 -0.003 1.809 10 49.8 0.289 0.014 1.781 11 50.0 0.287 0.003 1.799 12 50.2 0.288 -0.011 1.799 13 50.2 0.287 -0.050 1.747	7	0.502	0.289	-0.010	1.785
10 49.8 0.289 0.014 1.781 11 50.0 0.287 0.003 1.799 12 50.2 0.288 -0.011 1.799 13 50.2 0.287 -0.050 1.747	8	49.4	0.285	0.031	1.818
11 50.0 0.287 0.003 1.799 12 50.2 0.288 -0.011 1.799 13 50.2 0.287 -0.050 1.747	9	50.0	0.286	-0.003	1.809
12 50.2 0.288 -0.011 1.799 13 50.2 0.287 -0.050 1.747	10	49.8	0.289	0.014	1.781
13 50.2 0.287 -0.050 1.747	11	50.0	0.287	0.003	1.799
	12	50.2	0.288	-0.011	1.799
14 50.1 0.286 -0.005 1.810	13	50.2	0.287	-0.050	1.747
	14	50.1	0.286	-0.005	1.810

TABLE 2.4. Mean, variance, skewness and kurtosis of the adenosine bases. Abbreviations: chr – chromosome; var – variance; skew – skewness; kurt – kurtosis.

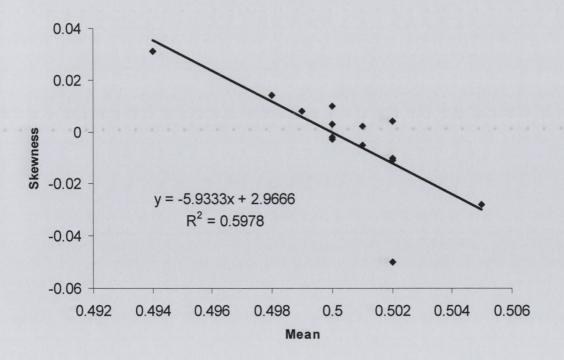


FIG. 2.9. Mean adenosine position and skewness for nuclear chromosomes. The outlier is chromosome 13.

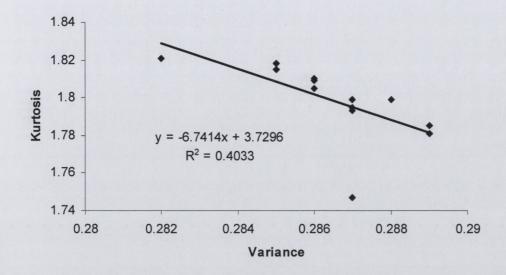


FIG. 2.10. Adenosine variance and kurtosis of the nuclear chromosomes. Chromosome 13 is an outlier.

Chargaff's third parity rule was found to hold in all the chromosomes and over six orders of magnitude. This is illustrated in Figures 2.11, 2.12, 2.13 and Table 2.4.

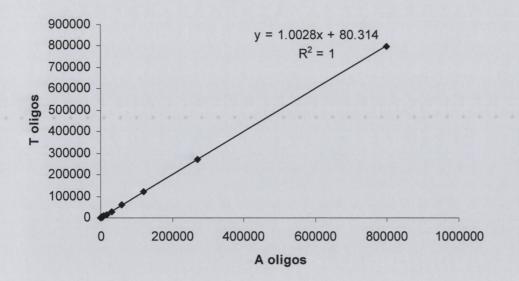


FIG. 2.11 Number of A and T isostichs in Chromosome 14. The number of each tract was plotted against its complementary tract.

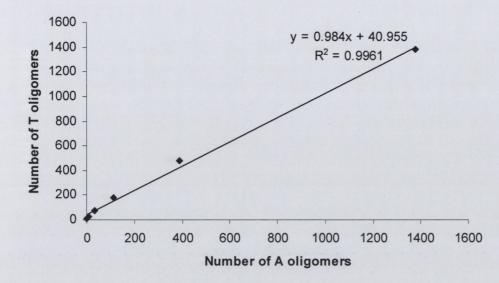


FIG. 2.12 Number of A and T isostichs in the mitochondrion. The number of each tract was plotted against its complementary tract.

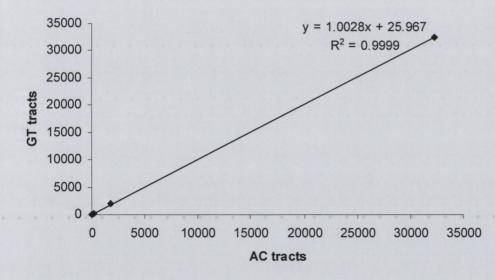


FIG. 2.13 AC and GT tracts in Chromosome 2. The number of each tract was plotted against its complementary tract.

Tract length	AC	GT
1	32 288	32 394
2	1 750	1 983
3	171	161
4	31	34

37

5	10	7
6	2	4
7	2	3
8	2	1
9	2	1

TABLE 2.5 Number of AC and GT tracts in Chromosome 2.

The observed and expected isostich numbers in chromosome are shown in Table 2.6. The model consistently underestimated the number actually found and the discrepancy increases as the isostich length increases. This discrepancy was also noted by the authors (108).

Figure 2.14 shows the presence of the knot in the data. This knot is to be found in all the nuclear chromosomes.

The slope and constant from the short and long regressions are shown in Tables 2.7 and 2.8. Two observations can be made immediately: the constant term in both the regression is approximately constant while the slope varies between the chromosomes.

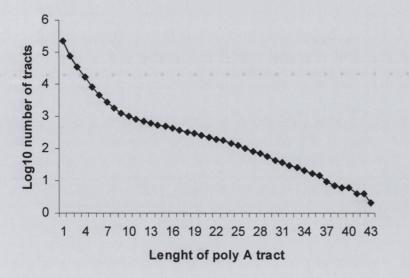


FIG. 2.14. Length of the adenosine isostichs (poly A tract) against log_{10} (number of isostichs) in Chromosome 2.

Tract length	Observed	Expected
1	227 762	185 268.0
2	76 160	36 489.8
3	33 628	14 616.7
4	16 241	5 855.0
5	8 254	2 345.4
6	4 4 3 0	939.5
7	2 669	376.3
8	1 785	150.7
9	1 272	60.4
10	992	24.2
11	827	9.7
12	704	3.9
13	604	1.6
14	533	0.6

TABLE 2.6 Number by length of adenosine isostichs in Chromosome 2.

Chr	Intercept	Coefficient
1	5.752	0.808
2	5.903	0.798
3	5.960	0.806
4	6.000	0.797
5	6.086	0.816
6	6.074	0.804
7	6.063	0.803
8	6.052	0.801
9	6.130	0.812
10	6.169	0.810
11	6.255	0.809
12	6.301	0.809
13	6.394	0.815

14	6.473	0.815

TABLE 2.7. Coefficient and intercept from regression of observed and expected short adenosine isostichs. Abbreviation: chr – chromosome.

The variation between the regression slopes suggested that a relationship between the length of the chromosome and the regression slopes might exist. A strong linear relationship between these slopes and the log of the length of the chromosomes was subsequently found (Figures 2.15 and 2.16).

Chr	Intercept	Coefficient
1	3.871	0.205
2	4.023	0.200
3	4.075	0.200
4	4.120	0.198
5	4.116	0.189
6	4.062	0.182
7	4.002	0.178
8	4.142	0.190
9	4.293	0.202
10	4.261	0.193
11	4.392	0.204
12	4.394	0.194
13	4.424	0.192
14	4.575	0.203

TABLE 2.8. Coefficient and intercept from regression of observed and expected long adenosine isostichs. Abbreviation: chr – chromosome.

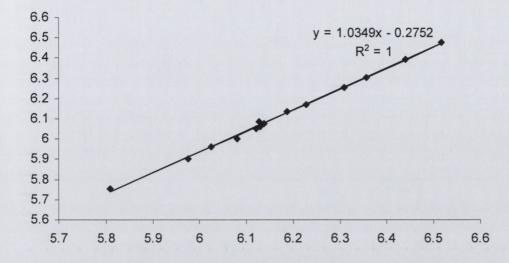


FIG. 2.15. Log_{10} chromosome length versus intercept value for short adenosine isostichs. (p< 10^{-6})

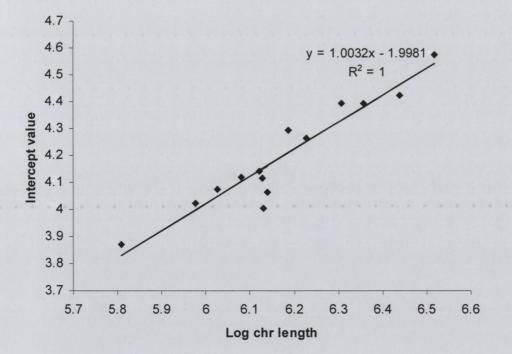


FIG. 2.16. Log_{10} chromosome length versus intercept value for long A tracts. (p < 10⁻⁶). The poor fit here for chromosome 5, 6 and 7 may due to chromosome misassembly. The value of R² is due to a rounding error in Microsoft's Excel program. The regression is highly significant.

2.4 Discussion

The study of whole genomes is an area as yet in its infancy and there is no systematic body of knowledge on the properties a typical genome should or should not posses. Most of these methods for studying these have been *ad hoc* with results scattered throughout the literature. To my current knowledge this is the first attempt to apply in a comprehensive fashion many of the individual methods that have been developed. Unfortunately since the subject currently lacks a theoretical framework the results may appear somewhat disjointed. This is a reflection on the state of the art in this area and it is in stark contrast the detailed theory that has developed around molecular evolution – a closely related topic. As results from the study of other whole genomes become available and as the theory develops more fully it is likely that it will be possible to integrate these findings here. With these caveats in mind an attempt will be made to integrate the findings here into current biological thinking.

Isochores were originally discovered by ultracentrifugation of vertebrate DNA (91). Since then they have been found in several groups of vertebrates and have been associated with the GC3 content of genes (320) certain types of gene (354) and the evolution of the vertebrates (33) and may have a role in gene expression. In vertebrates nucleosome formation potential correlates negatively with both GC content.

With the availability of entire vertebrate genome sequences how best to locate these isochores within the sequence data has become a debated question. Vertebrate isochores found by ultacentifugation fall into clearly defined peaks but when sought by windowing methods are never as well defined. In these genome with short windows (10 - 20 kilobases) isochores were not detectable while with longer windows they were (359). Theoretical problems of overlapping windows and long range correlations within the data have been recognized and addressed and alternative (85) methods have been proposed (272, 325).

P. falciparum genes and the surrounding non coding sequences typically have a GC content of $\sim 25\%$ and $\sim 10\%$ respectively. In humans the GC content of genes, while variable, lies around 50%. The majority of the *P. falciparum* genes lie within GC poor regions. The contrast here between the correlation of the GC content of vertebrate ischores and gene richness and the pattern found in *P. falciparum* is curious but may perhaps reflect the lower GC content of the *P. falciparum* genes.

The existence of a correlation between successive GC windows supports the hypothesis that DNA replication efficiency has been optimized by ensuring that the base composition changes relatively slowly over the length of the chromosome. If the effect is real it seems likely that this may have influenced both gene evolution and regulation. This hypothesis needs testing on additional data sets and should be presently regarded as merely speculative.

At the ends of the chromosomes lies a highly variable 21 base pair unit (366) – Rep20 – whose sequence is: TAA GAC CTA CAT TAG TTA TCT. Rep20 may span up to 22 kilobases (kbp) and is likely to be involved in telomere formation (48), chromosome organization during mitosis (346) and it confers an unusual three dimensional structure to the ends of the chromosomes.(358) Rep20 is relatively GC rich (6 GC bases per unit) and may be responsible for up to half the length of these GC rich isochores. This organization of the chromosome ends is conserved in *Plasmodium* (134, 366) but the biology is not yet understood.

The trough positions agree with those given by the published genome and similar sequences have been found in both *P. yoelli* and *P. vivax* (Huestis, unpublished work). While to date no centromeres have been confirmed experimentally no other candidates have yet been identified for this role.

Chargaff and his collegues (413) discovered that the base composition of single strands of DNA possessed similar relationships to those of double stranded DNA described earlier: to wit that A = T and G = C. The basis for the first rule was elucidated in the structure of DNA but that of the second remains elusive. Lobry found that the AT and GC skews changed sign near the origin or termination of DNA replication in three species of bacteria which he examined. This was particularly evident when only the GC content of the third codon positions (GC3) was considered. Later an improvement to the original method suggested - the use of a cumulative GC skew which makes the location of the origin more obvious. This has been confirmed in several papers (232, 304, 306). Studies of these skews have been done in the kinetoplastidia but the results have proved more difficult to interpret than in bacteria (344).

The plastid is known to have two origins of replication (455, 544) one of which lies in the region containing the tRNA and rRNAs: the second is as yet not well defined. The mitochondrial DNA is organized as whole polydispersed linear concatemers of the basic 6 kb unit mixed with a small proportion of circular molecules (374). While it is

thought that the circular forms may be the actively replicating units, this is far from certain (543). In the larger nuclear chromosomes multiple origins of replication are normally present and the use of the cumulative GC plots to locate these regions has not been as helpful as in bacteria. The molecular biology of DNA replication of DNA in *P. falciparum* was last reviewed in 1996 (537). The parasite's DNA polymerases and other major proteins and their various interactions resemble those of other eukaryotes. Nothing was known then of the sites on which these enzymes begin their work and since then there has been no progress in identifying these - undoubtedly because of the considerable difficulties such experiments would entail. In the absence of experimental evidence, while it is tempting to suggest that these locations on the nuclear chromosomes where the cumulative GC skew changes sign may be related to the origin of DNA replication, it seems wiser to decline to speculate until at least one such origin has been identified experimentally. The change is striking and it seems likely that these locations are related in some as yet unknown way to the biology of the parasite.

Chargaff's second rule has been shown that this rule holds for four of the twelve types of genome (316). The third parity rule was first tested and verified by Prahbu (373) in *E. coli*. Qi and Cuticchia (382) examined the number of complementary oligonucleotides (up to length 30) in 20 prokaryotic, 6 archebacterial and 8 eukaryotic genomes and their results agreed with those of Prahbu. Forsdyke (140) has suggested that the third parity rule is merely a consequence of Chargaff's second parity rule but this assertion is true only if the DNA sequence can be regarded as random, an assumption he failed to mention.

To show this consider a long random sequence of only As and Ts which both occur at frequency 0.5. The likelihood of an A occurring is (0.5) of AA occurring is $(0.5)^2$ and so forth. The expected number of A only oligomers is $(0.5)^n$ L where L is the number of bases in the sequence. Because the Ts occur with equal frequency the expected numbers of T oligomers of length n is equal to the complementary A oligomers of equal length. This argument can be extended to cover any oligomer composed of more than two types of base where bases occur with frequencies equal to their complementary bases. If the sequence is not random this argument fails to hold. The assumption of randomness is biologically implausible particularly in prokaryotic genomes which are ~90% coding. Forsdyke has since agreed that this suggestion was untenable (Forsdyke, personal communication, 2006).

Rejection of this assumption implies that a previously unrecognized form of selection pressure is acting to ensure that these two parity rules are obeyed. In spite of this rule clearly being a significant influence of genome composition and organization, the basis for this rule is not presently understood. Given its existence in many genomes it presently seems more likely that this rule has a structural basis similar to that of Chargaff's first rule than a primarily biological one.

That some previously unrecognized selective force appears to be acting on these sequences is supported by the findings here. The mean base position is eukaryotes is 0.5 times the length of the chromosome. The variance is approximately a quadratic function of the mean position and the skewness and kurtosis are linear functions of the mean and variance respectively. The complementary means (A and T; C and G) move in parallel unlike other genome types. These rules are known to be true for all eukaryotes examined to date and are true also in P. falciparum. That these relationships are true for all eukaryotes argues strongly that the location of the bases within genomes is not entirely random. That this third parity rule also holds suggests that they may be also organized into complementary blocks within the same sequence. That the adenosine and thymidine isostichs - relatively simple complementary sequences - both appear to obey a mathematical rule is consistent with this finding. If one of these isostichs obeyed such a rule and the other did not, then the third parity rule would not hold. The discrepancy between the observed and the expected numbers found here is not surprising. The model used to derive this formula assumed that the bases were distributed randomly along the sequence - a biologically implausible assumption.

This pattern of two log-linear regression lines with a knot at 8 bases appears to be common to all genomes with sufficiently long isostichs (315). A single log-linear regression line can be fitted to those with isostichs less than 8 bases. The constant in these regressions is always a linear function of the log of the genome length while the slope depends on the composition of the genome. Since this appears to be a biological law the models proposed for the formation of these sequences will be briefly reviewed.

The first paper proposing a mechanism for isostich formation was that of Streisinger *et al* (484) who suggested that DNA polymerase 'slips' during replication generating frameshift mutations: isostichs are assumed to be the result of this mechanism. Weiser

45

et al (534) measured the rates of increase and decrease and found that both were proportional to isostich length as this model predicted.

The relationship between isostich length and number within the genome has been examined before. Genetic theory based on the infinite allele model and without selection suggests a logarithmic increase in length with time (78): this model seems somewhat unrealistic. The results of Dechering *et al* have already been discussed. Dokholyan *et al* (117) suggested fitting a power law curve to the observed distribution of dimeric polymers in the human, mouse, *C. elegans* and *S. cerevisiae* genomes. Xu *et al* (555) looked at the mutation rate in two generations in 337 human families and found that the mutation rate could be explained by two types of mutations: one that tended to lengthen the tract and an opposing one that tended to shorten it. The model they examined proposed a constant rate of expansion of the tract and a contraction rate that increased with the length of the tract. While only one parameter had to be determined from the data the resulting fit was good.

One difficulty with this last model is that experimental measurement of tract expansion and contraction suggests a linear process (105). Kruglyak *et al* (249) proposed a model based on a stepwise mutation process. This model required estimation of three parameters and was of variable quality. Subsequent investigations raised questions about the usefulness of this model (66). In a survey of 27 eukaryotes, Zhou *et al* (561) found that this pattern was common to all organisms that they examined. Xu *et al* (555) found that the beta function with a single set of parameters gave an excellent fit to their data which covered only a range of five steps.

All the proposed models have difficulties. In the Streisinger's model if the insertion rate (I) exceeds the deletion rate (D), the entire sequence will eventually become a single isostich: conversely if D > I, the isostich will disappear. Dechering *et al* (108) did not consider bases other than thyimidine or adenosine. Both Dechering *et al* (108) and Zhou *et al* (561) assumed that the bases were distributed randomly: this seems implausible. The model of Dokholyan *et al* (117) was applied only to a small range of values and even over the chosen range the model fitted poorly. The knot in the data is also difficult to explain with any of these models.

Dechering *et al* (108) also found that the majority of the longer thymidine isostichs and AT tracts were found in regions that had been annotated as non coding. The authors also determined the number of isostichs in six additional genomes - *Homo sapiens*, *Caenorhabaditis elegans*, *Escherichia coli*, *Mycobacterium tuberculosis*, Saccharomyces cerevisiae and Arabidopsis thaliana - and found that the distributions in these genomes were similar to that of *P. falciparum*. They proposed that isostichs arose principally from replication slippage with some contribution from unequal crossing over. This second model seems more plausible than the others proposed to date but quantitative predictions have yet to be made from this hypothesis.

Some of the possible biological reasons for the existence of the knot in the regression slopes will now be discussed. Adenosine isostichs have unusual physicochemical properties that have not have gone unnoticed in evolution: they are unusually well hydrated in solution (60), bend DNA (203), form triple helices (18) and in blocks of 7 or more appear to delocalise electrons (332). A number of proteins exist that recognize these sequences specifically: datin in *Saccharomyces cerevisiae* and α -protein in mammals (548). These sequences tend to lie outside of nucleosome cores and enhance nucleosome formation (143, 378, 423, 489), act as promoters or enhancers of gene transcription - possibly because of their effects on nucleosome formation (6, 46, 125, 206, 219, 563) and their mutation rate (10⁻³ to 10⁻⁴ per generation) has been utilized to control gene expression (226, 301, 479, 534, 542).

Similar relationships exists for the poly AT tracts but the knot in the poly AT tracts appears at higher number of repeats - 10 to 12 – which again has been associated with a change in their physicochemical properties (187, 301).

From this brief overview it seems likely that the knot in the frequency data may reflect the biological properties of the length of the sequence. These properties may arise from changes in the physicochemical properties of these sequences. This leads me to propose the following semi quantitative hypothesis for the formation of isostichs in genomes.

There are two distinct classes of isostich: short and long. These are separated by their length with the separation length being 8 +/- 1 bases. Within each group DNA polymerase may slip generating a longer or shorter isostich. The rate at which this occurs depends on the physicochemical properties of the isostich. The newly generated isostich is then subject to the usual selection pressures including those against frameshifts within coding regions. Recombination may be important but this is more likely to be the case for the longer isostichs and its quantitative contribution to the distribution has yet to be determined. This hypothesis needs to be tested

experimentally. Potentially one such test would be to compare the rates of expansion and contraction of isostichs with lengths less than and greater than 8 bases.

2.5 Summary

All the 16 chromosomes are fairly typical of their type. The only unusual feature is the presence of the isochore structures at ends of the nuclear chromosomes and lying within four of them. These GC rich regions are always associated with the *var* (erythrocyte membrane protein 1) and the rifin / stevor families of proteins. The reason for the association between the GC isochore and the *var* genes is not yet clear: the GC and GC3 content of these genes is ~25% - typical of most *P. falciparum* genes so presumably this has something to do with gene control.

The position of the GC rich regions either near the center or approximately symmetrically at either end of the chromosome makes sense in view of the rules governing base distribution within a eukaryotic chromosome. Were these GC rich isochores not so placed within the chromosome then these rules would be violated. The hypothetical centromers also lie close to the center of the chromosomes but are distinct from these regions.

The putative origins of replication in the plastid and mitochondrion concur with the findings in prokaryotes. The relative uncertainty of the location of the orgin(s) of DNA replication in the nuclear chromosomes is also typical. It seems likely that there may be multiple origins of DNA replication but this is not yet known.

The third parity rule suggests that there may be additional selective forces acting on the genome that have not been previously recognized. The number and distribution of the isostichs within the genome is as *Dechering et al* have shown incompatible with a random distribution, a finding that is supported by the distribution of bases within other genomes. These patterns are at least in part likely to be due to the physicochemical properties of DNA. Additional work in these areas seems indicated.

Chapter 3: Annotation methods

3.1 Introduction

In this chapter the annotation methods used by the Malaria Genome Project (MGP) will be reviewed. This will be followed by a description of the methods used in this project. A list of the currently known problems is also given.

3.2 The MGP annotation methods

The main programme used in for gene prediction by the MGP – GlimmerM - was trained on a set of experimentally known genes. Two programmes were then used to verify the results: PHAT (70) and Gene Finder (129). Both Glimmer M and PHAT (77) were developed specifically for the annotation of *P. falciparum*. GlimmerM is based on an earlier program - Glimmer - that was developed for bacterial genomes (420). An additional programme was used at the Sanger Center earlier (Hexamer) but appears to have been phased out. The rule used (the 'best of three rule') was that if two programmes agreed on a gene then the gene was considered to be real and no further checks were made. The idea appears to have been that if two programmes agree then the error rate will be significantly lower.

Both GlimmerM and PHAT are closely related to Hidden Markov models (HMMs) which have been used for a number of years in genome annotation: examples include Genscan (62) and Genie (394). The other programs that also used - GeneFinder and Hexamer - do not yet appear to have been published but seem to have been based on HMMs.

3.2.1 GlimmerM

Glimmer is not a HMM but instead uses a number of similar ideas. The program has two parts: the first (Build-imm) reads the training set and generates a set of probabilities. The second part – Glimmer - scans for open reading frames longer than some input value (typically 500 bases) and then with the values obtained by the Build-imm program assigns probabilities to all the six reading frames. If these values exceed

some chosen threshold value the sequence is examined further. If there is an overlap, all six frames in the two genes are rescored and compared with the higher one being chosen. A note is made of the overlap and the proposed gene is reviewed manually.

Build-imm is based on an interpolated Markov model (IMM). Its main algorithm is a linear discriminant model with adjustments made for rare frequencies. Like a HMM, the program estimates a series of values which are used to determine the likelihood of a sequence being exonic or not. Build-imm reads all six frames of the training set and estimates the frequency value for all the Markov orders. If a string is particularly common than a predefined threshold in the training set the corresponding weight is set to 1. For very infrequent sequences an estimate is made from the lower Markov orders and the estimate is compared with the observed with the chi square test. If the observed values are significantly different from the expected they are given a higher weight. Conversely if they are not significantly different the higher orders are given a lower weighting.

The algorithm used by Glimmer is a dynamic programming one. The program reads in the sequence and continues until it finds a stop codon. Once a stop codon is found it examines the 5'region for possible coding sequences. If there are too many stop codons close together the program discards the sequence and moves on. 'Genes' of less than 200 bases in length were discarded. Where two coding regions were available the longer was chosen.

119 genes were used to train the GlimmerM. Introns were required to have at least 70% AT content. Values for the intron donor (5'GT) and acceptor (3'AG) sites were determined with known introns sites and then optimized against a randomly generated set of intron-like data. All putative intron splices sites in the sequence were scored. If the intron score suggests that there is a reasonable possibility of an AG site the program moves 5' to seek a suitable GT site. All the gene models were stored for evaluation later. Once all the reasonable gene models had been identified then each was scored by the IMM. Gene models whose scores are close to one another are both marked as possible genes for evaluation by human annotators.

3.2.2 PHAT

PHAT in contrast is based on a HMM with three main output states. exons, introns and intergenic regions. Exons are classified into four types - single, initial, internal and terminal. An exon state is has three parts - a pair of exon boundary state sites which flank a coding region. The five exon boundary states are translation start, donor, acceptor and translation stop. Introns are classified as phase 0, 1 or 2 according to the number of bases of the final codon predicted in the 5' exon.

The authors modified the basic HMM algorithm and termed this modification a 'generalized HMM.' In a standard HMM prediction of the nature of the state (the output) are made one base at a time. In PHAT the length of the exon was predicted with a gamma distribution whose parameters were estimated from the training set. In contrast the introns and intergenic regions were constrained to follow geometric distributions. These restrictions gave a considerable reduction in the running time and memory requirements of the program.

The Viterbi algorithm (281) is a dynamic programming technique that recursively calculates the most likely output state (here exon, intron, intergenic) depending on the input and was designed to work with one state at a time. PHAT's authors reduced the memory requirement of the program by recognizing that an exon state cannot jump directly into another exon state. In biological terms exons are separated by either intergenic regions or introns. This assumption may not apply in prokaryotes limiting PHAT's use there.

The initial and stop codons were required to be ATG and TAA, TAG, or TGA. Data from three bases upstream and ten bases downstream of the proposed GT site and 20 bases upstream and three bases downstream of the proposed AG site were used evaluate the predicted splice sites. Several of the estimated trinucleotide frequencies were zero in the training set so 1 was added to all values before the probabilities were calculated. This is an adjustment of the observed data that is not uncommonly used elsewhere. Adjustments for codon phases and that the entire length of the gene had to be divisible by three were also added in. In the original test set of 44 confirmed gene structures PHAT's published sensitivity and specificity exceeded 95%.

3.2.3 Weakness in GlimmerM and Phat

Both GlimmerM and PHAT programs suffer from elementary errors in design. A major weakness which they have in common with all annotation programs is that they require the genome sequence to be entirely error free – a situation that is not currently achievable. This is not easily correctable.

Build-imm assumes that (i) all genes started with ATG and ended with TAA, TAG, TGA, (ii) no genes have in frame stop codon and (iii) that the exons are in consistent reading frames with one another. It was also stated that introns were required to be at least 50 bases long (and less that 1500) but in the published annotation this was not in fact found to be the case which suggests there may be a problem in the code.

It was also assumed that the genes used to train the program were homogeneous. The included *var* genes were known at the time to be very atypical. HMMgene - another HMM based gene prediction program - was found to have an accuracy of 62 - 70% and which was noticeably affected by choice of training set (245). These problems did not arise with Glimmer in bacterial genomes where its predictive value is considered to be high. In the vast majority of known prokaryotic genes there are no introns and codon use is known to be homogenous. For example most *P. falciparum* introns are < 400 bases long: in the *var* genes there is a single intron of ~700-800 bases. As a possible strategy for improvement it may have been better to optimize the program to recognize the *var* genes separately in one run, locate as many as possible of these and then to re train the program on a set without *var* genes.

Alternative splicing is a common in eukaryotes: in *Arabidopsis*, humans and mice at least 10% of genes are alternatively spliced (217, 488) and in *Drosophilia* the figure may be as high as 40% (482). While alternative splicing is known to occur in *P. falciparum* (457) no examples of alternative splicing were reported presumably because the programs chose only the single highest scoring gene model.

The intron splice site prediction system was known to have error rate in identifying GT and AG sites when it was originally tested. The shortest intron known to date is a forty base *Drosophilia* intron: was the length of shortest intron found here was 67 bases: the proposed 50 base cut off seems reasonable.

The strategy of using all the training set is a known mistake. Standard practice is to separate the training set into two parts: one for training and one for testing. The two are need not be equal in size: the training set is normally larger. The reason for this

division is that to use the entire training set tends to overestimate the predictive value of the program since the parameters are optimized for that particular set of data. Since 119 genes represents < 2% of the genome, it is far from certain that this sample was representative of the entire genome. This question cannot strictly be answered until the genome is completed but the testing set acts in its place. The usual procedure is to train the program on the training set and then to test it on the test data. If the predictions are not adequate, the training set is re selected and the program re run until the best possible training set is available. This provides some indication of the likely error rate in the annotation.

This method also has its risks so the results have to be monitored carefully if it is to be used. Random reconstitution of the training set is usually not an optimal strategy. Random resampling is optimal only if the entire sample can be treated as homogenous. Some genes in *P. falciparum* are known to have introns: *ab initio* it is not possible to know if codon use differs significantly between these genes and the intron free ones. For this reason some form of stratified sampling of the training data to estimate genomic codon use would seem a more conservative approach.

Possible minor improvements in Glimmer (and GlimmerM) might include a different choice of statistical test. While the chi square is a reasonable choice here, this test is moderately sensitive to sample size and like virtually all statistical tests is more reliable with larger amounts of data. A maximum likelihood test included here may improve its predictive rate.

The authors of PHAT, in the interests of saving computer memory and running time, decided that the exons, introns and intergenic regions followed certain known distribution laws. The justification for this *a priori* assumption was not given: the biological logic seems suspect. From inspection of its predictions it can be seen that PHAT sometimes joins false 'exons' with false 'introns' to get long chains of artefacts; it ma join 2 genes into one or mistake untranslated regions for long introns. These appear to be a consequence of this 'optimisation.'

The authors used a data set of only 44 genes to train the program while 119 genes were available three years earlier to train GlimmerM. The reason for the use of this small data set was not given. As before the homogeneity and choice of the sample used are questionable. No attempt was made to split the data set into training and testing sets but here there may be some justification for this given the limited size of the training set. Predictions on the training set were said to be 95% correct a figure

53

that is somewhat better than one would expect on a *de novo* data. This suggests that the choice of genes in the training set may perhaps not have been entirely fortuitous. While it was known that an error rate existed for PHAT but no systematic examination of the predictions seems to have been done.

Problems with this computation approach should have been foreseen. Burset and Guigo (64) earlier had carried out a series of tests on nine gene prediction programmes and found that the prediction accuracy on the experimentally confirmed genes in their data set was variable but lay in the range 60 - 70%. Exon prediction was poor with less than 50% being correctly identified. Low GC content and the presence of introns in the genes further reduced the correct prediction rate. Long DNA sequences were given considerably worse annotations than shorter ones. Sequence errors including insertions or deletions are common problems in large sequencing projects and artificially introducing frameshifts or sequence errors into the test data set reduced the programmes predictive ability significantly. A second study has since confirmed these findings (179). These problems are illustrated and discussed further in the Appendix.

To illustrate the problems with the 'best of three' rule, let p be the probability of a programmes correctly predicting a gene and assume the gene predictions be independent. Put q = 1 - p. From the binomial theorem we have

 $(p + q) = p^3 + q^3 + 3p^2q + 3pq^2 + q^3$

Here p^3 is the probability of all three programs predicting the gene correctly and $3p^2q$ the probability of two programmes correctly predicting the gene. Adding these we obtain the probability (S) of a random gene being annotated correctly with the 'best of three' rule:

$$S = p^3 + 3p^2q$$

From Burset and Guigo's findings, we would expect p in general to lie between 0.6 and 0.7. This figure is chosen here rather than the claimed 95% rate (in PHAT) because of the concerns expressed here about the training and evaluation methods. With p = 0.5, S = 0.5. In words if the mean probability of a correct prediction of the programs is 0.5, then using all three programmes with the 'best of three' rule will not

improve the accuracy of gene prediction. With p = 0.6, S = 0.648. This is an improvement of 4.8% on the use of a single program but one in three genes would still be expected to be incorrect. If p = 0.7, S = 0.784, predicting an overall error rate of 21.6% in the annotation. In short while there is some improvement in the predictive value using this 'best of three' method, the improvement for typical programmes is limited.

When this practice was questioned, it transpired that the practice at both Sanger and Tigr was not to question the accuracy of the computer generated annotations but that error checking and corrections were considered something 'to be done in spare time' (Roos, personal communication, 2003: Berriman, personal communication 2003).

3.3 Annotation methods used here

The first pass of the genome was annotated in two steps which were then repeated a number of times until any obvious errors had been eliminated. Further genes and other information were then sought by examination of the literature, GenBank and other sources. All the rules listed here admit occasional exceptions but were adhered to as rigorously as possible. Heustis has described the origins of some of these in his Ph.D thesis (Monash University, submitted). Gene predictions using these rules have ben validated in a several *Plasmodium* species (134, 210, 499).

A first pass was made with GlimmerM which in spite all its faults is reasonably accurate at identifying potential coding regions. This process was later repeated with PHAT and GeneFinder both of which identified a small number coding regions that had been missed by GlimmerM. This step was performed by Huestis alone.

All such potential coding areas were then examined manually by both Huestis and myself according to a set of rules that had been developed earlier. (134, 210) All functional genes were assumed to start with ATG and to end with a canonical stop codon. For each predicted start ATG the 5' region was examined for potential coding regions. If no stop codon was found nearby, possible introns were considered until a potential start codon in an acceptable context was located. Very short initial exons were avoided if an alternative was available. Initial exons of three and four bases are known to exist in *P. falciparum* (101, 237), a factor that was borne in mind. An ATG was considered as a potential start codon if a purine was present in the -3 position. In the case of ATG ATG the second was chosen unless the first ATG had a purine in the

-3 position. Five prime of a genuine ATG site, there were other signs of a coding to non coding character were always present: long T tracts and multiple in frame stop codons are very common findings in these regions. ATGs lying within $(AT)_n$ – TATATGTATA - were considered to be non coding.

The region 3' of the stop codon was similarly examined for potential introns. Genuine termination codons were followed by other in frame stops codons and an A rich region. Genes predicted to end in either G TAA or G TAG were examined closely as potential donor splice sites. Where a gene was predicted to end close to a start of a second gene, the genes were both re-examined to rule out an unidentified intron between the coding regions.

Repetions of trinucleotides were assigned to coding regions with GAA, GAT and repeated AATs were considered diagnostic of a coding region. Dinucleotide repeats were considered diagnostic of non coding regions. Repeats do not normally cross exon-intron boundaries. Runs of $(RY)_n$ with n > 5 were assigned to introns where possible. In a given region the coding frame is most likely to be the longest of the three reading frames. The frame of the last exon was especially difficult to fix correctly. The rule here was to choose the longest reading frame available but this rule produced genes that required revision on several occasions. Sequences where all the coding permutations of T and A were present (AAA, TAT, AAT, TTT) and no other bases then the sequences was considered unlikely to be a coding region.

All introns, excepting only those found in pseudogenes, were assumed (1) to begin with GT (2) to end with AG and (3) to have a length of at least forty bases. Introns that violate one of more of these conditions may exist but it was felt that these were likely to be rare. Base use within introns is asymmetrical with runs of As preferred at the 5' end, runs of Ts toward the 3' end and runs of AT when they occur within the body of the intron. If an intron does not have stop codons in all three frames, it must have an (RY)n,tract, a binding site or a A tract adjacent to the donor GT. Otherwise it was considered to be either a coding or untranslated region.

Preference was given to potential GT AA, GT A or GT G over GT T donor sites. GT C was considered highly unlikely to be genuine. All GT T sites were with one exception followed closely by a run of As. Preference was also given to AG GT splice sites. Where potential GTGT donor sites were found the first GT was considered to be the usual donor. GT $(AT)_n$ was given a low preference as a donor site because AT repeats are normally located within the body of the introns rather than at donor sites.

Where two potential GT sites that looked equally probable preference was given to the GT with the longest coding region. The remaining GT may represent an alternative splice site. If the 3' exon begins with AG, the the 5' exon cannot end with T ruling out the possibility of a T GT sites. Any short exon bounded by exons that are in frame with each other presents an alternative splicing possibility.

T AG was the preferred acceptor site while G AG was considered rare. Potential A AG acceptor sites were reviewed carefully. The nearby presence 5' of a run of As made these unlikely to be a genuine acceptor sites. C AG while less common than T AG was considered as a possible splice site. In addition TAG, CAG and AAG do not usually occur in the region 5' of a genuine slice site and GAG is rare in this region.

Five prime of the genuine acceptor sites a T/C tract was invariably found, varing in length from 4-5 bases (rare) to over 30. Long T tracts are uncommon in introns and their presence was taken to be suggestive of an intergenic region. The acceptor AG may be considerably downstream of the TC binding site. In this case, no AG can intervene between the splice site and the TC binding site. Lying 5' of the T/C tract, a potential lariat site (*vide infra*) could usually be found. Where two equally probable acceptor sites were found, preference was given to the one that maximized the potential coding frame. The remaining AG site may act was an alternative splice site.

Multicopy genes could be predicted reliably by alignment. An analogous method for a similar purpose has been used elsewhere (90). In general multicopy genes preserve their splice sites and frequently their codon number. Additional copies were located within the genome by BLASTing genes and exons against the contigs or later the chromosomes and examining the matches. Transitions outside the splicing sites were considered as 'matches' while transversions were re-examined.

Frame shifts in coding regions were identified by eye. This step was carried out by Huestis. If the gene with the suspected frameshift was a member of a multigene family this task was considerably simplified by alignment. Sequence errors in the non coding regions are expected to exist but cannot presently be identified. Typically these indels (insertion/deletions) were additional or missing As in a short poly A tract. Sequence errors in the non-coding regions are expected to exist but cannot presently be identified.

The technique used was to maximize the relatively rare G and C bases on the first coding position starting initially on the left of the suspected frame shift and then repeating this procedure from the right. Where the frame shifts position can be identified an 'N' was inserted as a place holder in the annotation unless the annotator was sure of the nature of the base. In a run of As, it is extremely likely that the missing (or additional) base is an A. In multicopy genes, alignment allowed the identification of some bases. Genuine pseudogenes had multiple stop codons, degenerate splice sites, short potential coding regions or combinations of these problems.

On completion of this manual pass further examination of the codons and codon base use was carried out. Two summary statistics - the codon chi square test (χ^2) and Wright's effective number of codons (N_e) were used to examine the codon use. These last two steps were carried out exclusively by myself.

The codon chi square statistic was described in Sharpe *et al* (440). This was based on an earlier paper which described a new data transform. Stenico *et al* (474) defined a measure of codon use - the relative synonymous codon use (RSCU) - which controls for the unequal representation of amino acids in the codons and subsequently used this measure in genome analysis (441). The RSCU is defined as

$RSCU_{ij} = x_{ij} / [(1 / n_i) \Sigma x_{ij}]$

where x_{ij} is the number of occurrences of the jth codon for the ith amino acid which is encoded by n_i codons. In simpler terms, the RSCU_{ij} is the observed number of occurrences of a codon divided by the expected number. The expected number here assumes that all the synonymous codons are used equally frequently in the gene. The codon chi square statistic is the sum of the square of the RSCUs less their expected value divided by the RSCU. In symbols

$\chi^2 = \Sigma (RSCU_{ij} - [1/n_i])^2 / RSCU_{ij}$

where the $RSCU_{ij}$ is the relative synonymous codon use of the jth codon encoding the ith amino acid and n_i is the number of codons for the ith amino acid. The sum is taken over the length of the gene excluding the stop codon. This statistic has a lower bound of zero but no upper bound. Its upper value is limited only by the finite length of the genes.

The effective number of codons (N_c) in a gene is a summary statistic created by Wright (551). Like the codon chi square this statistic has no theoretical maximum value but for practical purposes a maximum value of 61 is imposed upon it. It has a

lower bound of 20. This statistic varies between 20 (only one codon per amino acid used) and 61 (no bias in codon choice). On biological grounds neither of these extremes seems likely but given the general preference for unequal codon use in all known genomes the distribution can be expected to be biased towards the lower end of the distribution. The effective number of codons is a mean and is determined in two parts. Firstly the 'homozygosity' (F) of the amino acid is determined

$$F = (n \Sigma p_i^2 - 1) / (n - 1)$$

where n is the number of times the amino acid appears in the gene and p_i the proportion of the amino acid encoded by the ith codon. The effective number of codons is the mean of these 'homozygosities'.

$$N_{c} = 2 + (9 / F_{2}) + (1 / F_{3}) + (5 / F_{4}) + (3 / F_{6})$$

where F_n is the mean homozygosity of the codons with degeneracy n. The 2 enters the equation because there are normally 2 amino acids with only one codon (Met and Trp). Wright also suggested an approximation if an amino acid was missing from a gene. Assume that threonine an amino acid encoded by four codons was missing. Then the estimate of F_4 is

$$F_4 = (F_{ala} + F_{pro} + F_{val} + F_{thr}) / 4$$

This method of calculating the effective number of codons has been reexamined. Because not all the synonymous codons contribute equally to the N_c (150) an alternative method of calculation has been proposed making allowance for GC bias (149). This proposal has generated some discussion as it is known that a correlation exists between Wright's N_c and the GC3 and to a lesser extent the GC1 content (293). This discussion remains active at the time of writing (152).

The Chebyshev inequality which applies to any probability distribution is

 $\mathsf{P}(|X - \mu| \ge k\sigma) \le 1 / k^2$

where μ is the mean of the distribution and σ is the standard deviation applies to any distribution with a finite mean and variance. The use of the Chebyshev inequality was based on the hypothesis that the translation system has been optimized by evolution and consequently has a preferred distribution of base use within codons and for codon use in genes.

A number of refinements have since been made to this inequality. If the distribution is unimodal, the Vysochanskiï-Petunin inequality is true (522): for $k > (8/3)^{1/2}$ we have

 $P(|X - \mu| \ge k\sigma) \le 4/(9k^2)$

showing that > 95% of a unimodal distribution lies within three standard deviations of the mean. While the Vysochanskiï-Petunin inequality could have been used here, a unimodal distribution of base and codon use was considered to be too strong an assumption. Consequently their use was examined systematically only with the Chebyshev inequality.

Genes with either N_e or χ^2 values that lay four standard deviations from the mean were regarded as suspect and reexamined. Genes whose base use in any position lying four or more standard deviations from the mean were also re examined. Genes with $N_e = 61$ were automatically regarded as suspect and re-examined.

Almost invariably genes identified as needing re examination were found to have overlooked introns, incorrect splice sites or other serious problems. Once the problems were identified and corrected the gene was once again re examined. Rarely re annotated genes were reidentified as suspect necessitating further re examination.

Once the annotation process was considered to be completed the genes were identified by BLASTing them against GenBank. Genes were considered to be orthologous if the blast score was $< 10^{-10}$ and the match was over half the length of the gene. A number of the protein encoding genes could not be identified. If there was experimental evidence of their existence (transcription or proteonomic) they were labelled 'protein of unknown function.' Where - rarely - two identifications were possible both were given.

In a small number of cases the base and/or codon use suggested that the gene might simply be an open reading frame rather than a genuine protein and these were labelled 'open reading frame.' This label was not applied until after the sequence had been inspected by both annotators and a BLAST search had been done to look for possible homologues in the databases. Only if both annotators agreed that the sequence looked unusual but could still potentially be protein encoding gene and that no homologues could be found in the on line databases, was this term applied.

The remainder were labelled 'hypothetical protein.' Where a protein encoding pseudogene could not be identified it was labelled 'pseudogene.' If confirmatory evidence was available (experimental, transcriptional, proteonomic) this was included in the annotation. This last rule was not applied in cases of *var* (Emp1), *rifin* or *stevor* genes as assignation of the clones to a single gene was impossible in these large multigene families. If homologs were known in related species this were included as given - including 'putative dentin phosphoryn' (*P. yoelli*). Dentin phosphoryn is found only in teeth.

The annotation was supplemented with daily examination of the literature for reported genes. The GenBank database, the MGP annotation and other publications were searched for additional genes. Bozdech provided the locations of the oligonucleotides used in his paper (53).

Where the annotation remained equivocal or appeared anomalous comparisons were made with genes from other *Plasmodium* species - mostly from *P. yoelli* - and these genes were available for almost 50% of the identified *P. faciparum* genes. The gene from the second species was translated; the proteins sequences from both species were then aligned and then both were mapped back to the original codons for comparison. This procedure was used to identify frameshifts, resolve ambiguous splice sites and to identify new genes. Additionally the work required a new system of gene identification as the five skip system favored by the MGP was found to be inadequate. Details are of this system are given in the next chapter.

3.4 Known problems

That no known single annotation method is sufficient to identify all the protein encoding genes alone became abundantly clear during this work. The use of multiple data sources resulted in the discovery of many additional genes that would have otherwise been missed. The reexamination process identified a number of in frame stop codons that had been initially overlooked. Significant numbers of changes were also made to initiation, termination, intron donor and intron acceptor sites were required.

The resulting gene predictions made here and subsequent gene product identification were both deliberately conservative. The present version is incomplete partly as a result of this conservative approach and it is certain further additions and revisions will be needed before any annotation can be considered complete. This having been said this initial version is believed to be free from obvious error.

The exclusive use in this annotation of ATG as the start codon may not be correct. Examples are known in other eukaryotes of non ATG start codons including GTG in the protein NAT1 (491). Nonetheless this is uncommon and is believed that this would have caused have caused many errors here.

The choice of sequences surprisingly became an issue. Differences between the chromosome sequences found on the Sanger/Tigr sites and those on the PlasmoDB sites were found. While all the sites agreed on the sequences of chromosomes 1 - 4, on the PlasmoDB site chromosome 5 had an additional 6 bases added to the start of the sequence and an additional three bases had been added to the start of each sequence of all the remaining chromosomes compared with the Sanger/Tigr sites. The cause of these differences is not clear. The Sanger/Tigr sequences were used here as it was felt that these were to be more likely to be correct.

The transcription data assisted in the discovery of a number of additional genes. Interestingly some clear pseudogenes eg Pf12_1:2188535w - a *stevor* pseudogene and Pf12_1: 779926c a *var* fragment – have been documented as being transcribed. Transcription data for multicopy genes such as these needs to be interpreted carefully as false positives may occur. Alternatively it is possible that these mRNAs are real and are involved in some as yet unknown way with gene regulation. This latter hypothesis presently seems less likely than the former but this question cannot be resolved without experimental work.

Detectable sequence errors were found in all chromosomes. Both runs of single nucleotides and the overall high AT content are difficult to sequence accurately and this represents a serious problem to automated annotation. In Pf13_14:815831w (erythrocyte binding ligand/antigen 1) - a gene that has been cloned and whose sequence has been confirmed - has a run of 10 Ts in the first exon. The Sanger chromosome 13 sequence gave this sequence as having 12 Ts giving rise to what would appear to be a frameshift mutation.

62

Pf2:863065c rifin related Maurer's cleft 2-transmembrane protein alpha (PfMC-2TM 2.1 α) is a pseudogene. The ancestral gene is a member, along with the *rifin* and *stevor* genes, of a protein superfamily (422). A functional paralog of this pseudogene is found on chromosome 1. All member of this family have two exons.

Start of exon 1

Chr 1: atg ttt cat tat att tat aaa ata tat att ttt acc ata ata cta tgt gca tcg aat cta ttt aat aac

Chr 2: att ttt cat tat ttt tac aaa ata tat att ttt acc ata ata ctc tgt gca tcg aat cta ttt aat aac

Within exon 2

Chr2: a tac ata aat agt gat gat ata tta gaa aaa aat aaa tca att [a] atg aac

End of exon 2

Chr 1: aca ttt ttt caa aac aaa aag caa ata aca aaa taa* Chr 2: aaa ttt ttt caa aaa aaa aaa*taa*atg ata aaa taa*

Before the ATG start codon lies an AAT sequence preceded by several in frame stop codons in both chromosomes. The ATG has mutated in the pseuodgene to ATT. A single example of a frame shift lying within the second exon is shown above. A deletion mutation is found further 5' within the exon. An additional premature stop codon is also present. These are typical features of *P. falciparum* pseudogenes.

Two *stevors* annotated here (Pf7_7:467609c and Pf11_1:56792w) have in frame stop codons. In each case the genes have been reviewed by both annotators. In genuine pseudogenes in *P. falciparum* multiple stop codons and insertion/deletion mutations can be found. These two genes have only one internal stop codon and the remainder of the gene is clearly coding. There are several possibilities for this: the sequence may contain a genuine stop codon and the annotation is incorrect; there may be a sequence

error; or this may be a form of regulation. It seems most likely that these stop codons are simply sequence errors and that the genes are translated. Complicating this hypothesis is that read through stop codons have been shown to be present in *P*. *falciparum* - Pf 60.1 on Chr 13 (39). The existence of translated in frame stop codons is problematic for automated annotation.

Five genes annotated here do not have the canonical GT - AG splice sites. The 5' site of Pf2_1:883238wp (a fragment of an integral membrane pseudogene) has mutated to GA: the intron here was identified by alignment with a functional paralog. Pf5_1:885711w (a putative nucleotide binding protein) has a 5' site of GC. This may be correct as some genes (<1%) are known to use GC rather than GT. Alternatively this could be a miscalled base.

Pf6:841221c the telomerase reverse transcriptase (TERT) gene has an acceptor site of AA instead of the canonical AG. The corresponding splice site was identified in the orthologous gene of *P. yoelli* and it seems that this non canonical site in *P. falciparum* is likely to be a sequencing misread. Pf14_1:1839984w (open reading frame) has a 3' splice site of GTG. The intron looks very typical of *P. falciparum* and alternative 3' sites are not present. Codon use in the surrounding gene is unusual: there no transcriptional or other data currently available to assist the annotator. Presently this gene should be regarded as highly suspect and experimental work is needed to clarify the status of this putative gene.

The genes Pf12:2188535w (*stevor* related integral membrane protein pseudogene) previously annotated by Stanford as PFL2580w (hypothetical protein) illustrates another difficulty encountered during the annotation process. C4 and Pf12 were the names of the contigs these sequences were located on.

Exon 1

Intron

Exon 2

c4: aag aat ttt caa agc aac ggt tac att tca cca c Pf12: aga cta ttt caa agt aac agt tac atg tca caa c

The upper row of the exon alignment is a *stevor* from chromosome 4 and the lower is the first exon from Pf12:2188535w. After an adenosine was inserted into the fourth codon of the ancestral gene, the reading frame was altered and the intron of the resulting pseudogene now has aquired a 5' AT site. An automated system will chose the GT dinucleotide lying within the intron, changing the reading frame again. The alignment with the stevors is extensive and shows the ancestry of the gene but it is possible to translate this new gene as Stanford did. This gene may be both a functional gene (Stanford) and a 'pseudogene (here).' Distinguishing these possibilities requires experimental work.

The presence of 'T' three bases before the initial ATG (position -3) is uncommon as it is normally associated with poor translation of a gene. The gene Pf8_8:350523w (a putative outer arm dynein light chain) has a TAA sequence before the initial ATG. This was identified and reexamined: an identical pattern is present in *P. yoelli*.

Py: taa ntg anc aga --- gta gtt aag gaa gtt acc ...
Pf: taa atg agt aga acg atg gca aaa gaa gct acc ...

The upper sequence is from *P. yoelli* and the lower from *P. falciparum*. The 'n' here in the *P. yoelli* is not a typo but is it is as given by the sequencing centres. The biological significance of this -3 TAA sequence is not known but since it appears to be conserved this may perhaps be a mechanism of translational regulation. Furthermore the second ATG within the sequence may act as an alterative start site – yet another question that cannot be resolved without experiment.

Alternative splicing is known to occur in *P. falciparum* (238, 328, 457, 515) but no alternatively spliced genes were reported by the MGP. While the present draft of the annotation predicts 46 alternatively spliced genes, this is almost certainly an

underestimate of the real number since alternative splicing is difficult to predict and all existing methods are inadequate. Antisense RNA (80) may exist in *P. falciparum* but again this could not be identified here.

tRNA and rRNA genes constitute the bulk of the non coding genes. Fourty tRNAs have been identified here. While this is less than the 44 found in *E. coli* it is an increase of 5 over the published version. Others may remain to be identified. These genes, with five exceptions - tRNA-Pro (chromosomes 12), tRNA-Ser, tRNA Arg, tRNA-Thr (chromosome 13) and tRNA-Pro (chromosome 14) - which are found alone, occur in groups of two or three without intervening genes even when encoded on opposite strands. Five rRNA genes all of which occur in the order 18S-5.8S-28S were found and are probably transcribed as a single unit as in other eukaryotes. The non protein encoding genes in the mitochondrion are all fragmented RNAs. One putative small nuclear RNA has been annotated here (Pf3_1:365243w). Experimental confirmation is needed here.

The BLAST programme used for gene identification was not free from difficulties either. While no false positives were found, in small number of instances genes known to exist from the literature were not identified. These problems were not reproducible and may perhaps reflect the BLAST servers' work load. Theoretical problems with Blast have also been reported (26). Along with the conservative approach these problems taken here may have resulted in the under identification of genes. 'Then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. Then they took him, and slew him at the passes of Jordan: and there fell at that time of the Ephraimites forty and two thousand.' Judges 12:6

Chapter 4: Gene identification system

4.1 Introduction

While hundreds of genomes have been sequenced and annotated this undertaking has not been without its difficulties. In the mouse genome 58% of the 5' ends have been shown experimentally to have errors (114): 20-25% of the genes have alternative splice sites that were not predicted and many new unpredicted exons were found. A second study reported similar findings (523). Problems have also been reported in the human (426), rice (30) and Aspergillus nidulans (451) genomes. After reviewing part of the human genome, Nelson (338) has likened the annotation as something that might be found in the satirical 'Mad Magazine.' In the less well-studied Cryptococcus genome, only 60% of known genes were predicted correctly (500). Bennetzen et al. (30) found an error rate of \sim 50% in the rice genome and described some of the published annotations as "fantasies." The Arabidopsis researchers have recognized this problem and a community-based project to improve this genome's annotation has been created (428). Laird et al (258) have identified in 60 papers examples of the misidentification of a single Arabidopsis gene (PR-1): this list is unlikely to be exhaustive. While attempts have been made (210) to maintain consistency with earlier papers (156) this practice has not always been observed (158). The presence of errors reported in the *Plasmodium falciparum* annotation (313) which have subsequently been confirmed (34) has lead to multiple copies of the same gene being listed as 'synonyms': a gene (GenBank accession number NC 004330) currently identified as PFI0010c is listed as having the following synonyms: PFI0015c, PFI0020w, PFI0025c, PFI0030c, PFI0035c, PFI0050c, PFI0055c, PFI0065w, PFI0070w, and PFI0075w. All of these synonyms have been given the principal gene ID of PFI0010c.

While the consequences of gene misidentification may not be as severe as the mispronunciation of shibboleth ('ear of grain' - Hebrew) was to the Ephraimites after being defeated by the Gileadites *circa* 1150 BC, it is still significant. Nebert and Wain wrote (337):

"Why is agreeing on one particular name for each gene important? As one genome after another becomes sequenced, it is imperative to consider the complexity of genes, genetic architecture, gene expression, gene-to-gene and gene-to-product interactions and evolutionary relatedness across species. To agree on a particular gene name not only makes one's own research easier, but will also be helpful to the present generation, as well as future generations, of graduate students and postdoctoral fellows who are about to enter genomics research."

With the large and increasing numbers of genes in the literature and the re-annotation projects underway, it is essential to have a simple and unique method of identifying genes which can be applied to any organism and that can be consistently applied between groups working independently of one another. This is a problem which has been studied for several decades (88).

Previously proposed solutions can be classified into two approaches: either an apparently arbitrary number or a simple enumeration scheme. GenBank uses the first of these methods. Each sequence listed therein is identified by a constant and unique alphanumerical identifier assigned by GenBank, ensuring that the sequence sought can be retrieved once the identifier is known to the end user. The problem with this approach for genome annotation is that it creates difficulties with consistency of the data that may be difficult to identify. If there is no central distributor of identifiers it is possible that a gene may receive multiple identifiers or that two or more genes will receive the same identifier. Recognition of these problems is difficult at least in part because of the non-intuitive identifiers used. The use of such identifiers is often recommended where database access is under the control of a single entity but this practice has clearly created difficulties given the distributed nature of gene identification.

The second approach is the use of sequential numbers starting at some arbitrary point in the genome. This method has been used by the Sanger Center and the Institute for Genomic Research for several their published genomes including that of P. *falciparum*. While this method has the significant advantage of being intuitive, it is highly prone to consistency errors. Consider a sequence of five genes where each has been allocated an identifier from 1 to 5. Gene 2 is subsequently found to be in error, is re-annotated and assigned a new identifier 6. At the same time a second group discover that gene 3 is incorrect and also assign the identifier 6 to this second reannotated gene. Unless both groups submit their findings to the same journal, it is unlikely that this inconsistency will be recognized for some time. Given the difficulties in genome annotation that currently exist continued use of these methods is likely to result in problems similar to those previously described.

A simple method of gene identification providing unique identifiers consistently between different groups of workers and which may be used for either primary annotation or in re-annotation work is described here. The method may be applied to protein encoding and non-protein encoding genes, promoters or other elements of biological interest. This method was developed jointly by Dr Huestis and myself.

4.2 Proposed system of gene identification

Like all such systems (297), a system of gene identification should comply with four principles:

1. Full address principle: the address must be unique and enable the user of the annotation to locate and to retrieve the gene.

2. Minimum content principle: additional information that should be provided to the user beyond the address. Where there has been revision of a previously annotated gene the revision history of the gene must be included. The function of the gene when known should also be included.

3. Compacting principle: unnecessary verbosity is to be avoided.

4. Formatting principle: the use of punctuation, abbreviations and other materials in the identification should be standardized and used consistently throughout.

A gene may be uniquely denoted by three things that are routinely available: the first base position, the strand, and the DNA sequence within which it lies. A gene lying on *Plasmodium falciparum* chromosome 1 on the Watson strand starting at base 100 is here designated as Pf1:100w. Genes lying on the Crick strand are designated with the suffix 'c'. A similar gene initiating at base 100 on the Crick strand would be designated as Pf1:100c. If the chromosome is not known, the contig identifier is used

- for example in *Plasmodium vivax* contig 1 the identifier would be Pv_contig1:100w. To reduce the potential number of possible identifiers where the chromosome is known then the chromosome identifier should be used. Where this is not known the contig where the gene has been most recently described should be used.

A species identifier is added as a prefix to make the system more portable. For P falciparum Pf1:100w was used here indicating a gene on chromosome 1 starting at base 100. While rules for the consistent use of species identifiers need to be developed, it seemed sensible to use of the first letter of the genus name and one or more letters of the species name.

Since it is possible for two or more groups sequencing genes to find the same gene but in different contigs a system of reconciliation is needed. Such a system has been used for centuries in manuscript analysis. Manuscripts that have been renamed are designated with the reserved word *olim* ('formally known as'- Latin). If a gene has been designated as PFA0005w and is later redesignated as Pf1:100w, this is contained in an annotator's note: Pf1:100w olim PFA0005w. In this fashion it is possible to ensure continuity of the gene identification throughout the literature. While backwards compatibility with earlier annotation systems is desirable, the multitude of systems currently in use has made this impossible. Earlier denotations can be incorporated so far as possible using *olim*.

Alternatively spliced genes sharing a common first base can also be accommodated. Where a gene (Pf1:100w) is known to have two or more alternatively splice forms these forms are denoted 'Pf1:100.1w', 'Pf1:100.2w' and so forth. Where two different laboratories discover different alternative splices, the gene identifier should make their relationship clearer than is presently the case. As additional splicing variants are discovered the numbers .1, .2 are to be used in the order of publication and not subsequently reused if the paper that described them originally is subsequently shown to be in error.

The sole remaining difficulty here occurs where two or more laboratories discover on the same contig two or more alternatively spliced genes and publish them simultaneously. While this does not presently seem to be a likely scenario, the relationship between these genes should be straightforward to identify and additional extensions of the rules can be developed: in this hypothetical case it is proposed that the longest gene be given the lowest available number.

To complete the gene identification the name of the proposed gene product is included. To use the overworked hypothetical example yet again this gene would now be Pf1:100w olim PFA0001w glucose dehydrogenase. Additional information may also be included here.

Pf1:100w olim PFA0001w glucose dehydrogenase Location=100..200|300..400

Here the exon locations have been included with the gene identifier. The inclusion of location, while not strictly necessary in order to generate an unambiguous gene identifier, provides additional data to prevent confusion between simultaneous gene identification arising in different laboratories.

During the *P. falciparum* re-annotation it was found that the presence of sequencing errors has required the insertion or deletion of a single base at various points in some coding regions. As in manuscript analysis, the use of annotator notes here has proved to be of additional value, allowing comparisons to be made directly between alternative possible gene models. The compilation of annotator notes for genome annotation also provides consistency between researchers. Such notes are unlikely to be of great use to scientists wishing to clone genes but are rather intended to be of use to those examining genome annotations. A limited vocabulary for this use is proposed in Table 4.1. By convention manuscript annotations are in Latin. This convention was used here with the exception of the term 'frameshift' - a usage unique to genomes.

anti	connected with		
caput	start		
intra	lying within another gene		
consortis	overlapping		
extra ordine	out of order		
finis	stop		
frameshift at [xxx]	frameshift at position [xxx]		
item infra	likewise below		

item supra	likewise above
Iterum	second instance
olim	formerly known as
sic	it appears thus in the original
pars altera	one of two parts of a gene
pars prior	the first of two parts of a gene
residuum vestigia	gene fragment
tertium	third instance
trans	on the opposite strand
vice [datum1] [datum2]	replace [datum1] with [datum2]
vide lacuna [xxx]	note gap at [xxx]

TABLE 4.1. Proposed vocabulary for annotation notes

If a gene is known to overlap a second gene (*consortis*) or to lie within a second gene (*intra*) this should also be included in the notes. If gene A overlaps gene B this is designated: gene A *consortis* gene B. Similarly if gene A lies within gene B this is indicated thus: gene A *intra* gene B.

The order in which these terms should be used has not been fixed but it seems sensible to place the terms following *olim* at the end of the notes. For example let gene B be a revision of gene A with a newly discovered frameshift at position 100. This would be then designated: gene B -1 frameshift at 100 *olim* gene A.

4.3 Discussion

The purpose of any system of identification is to ensure reproducibility of results. The method described here should make this possible. Additionally, unlike previous methods, the technique described here makes the process of re-annotation of the gene more transparent.

This method was proposed by the authors at the British Society for Parasitology meeting in 2004 and appears to have bee adopted at least as an interim measure by the Sanger Center in the work on the *Plasmodium vivax* genome.

Three criticisms of this method can be made: (1) that it generates too many gene identifiers for a single gene (2) that the start codon may not be correctly identified and (3) that as a genome is being assembled the identifier may become too cumbersome to use. Each of these points will now be addressed.

In 1945 Beadle and Tatum introduced the 'one gene, one enzyme hypothesis into biology. Since then it has been realized that genes have alternative translation initiation sites, alternative splicings and other modifications. In general genes should more properly be regarded as a collection of RNA entities rather than a single sequence within the genome. The difficultly for an annotator or experimentalist is to designate the sequence that is being described rather than the gene it forms part of. This is a problem of informatics rather than one of biology.

An admittedly arbitrary choice was made here to designate sequences with differing translation initiation sites from the single gene of which they form a part. This allows a consistent and unique identifier to be created for each sequence of the gene. If a gene is associated with sequences with several translation initiation sites then this relationship should be indicated in the accompanying annotation notes. The existence of number of identifiers for a single 'gene' merely reflects the underlying complexity of the biology.

The method described here relies on the identification of the start codon. If a gene is correctly identified then all groups will end up using the same identifier. Unfortunately current annotation or experimental methods may or may not identify the initiation base correctly.

Optimally an identification system will reflect a reproducible aspect of the biology but this is not an essential attribute of such a system because the creation of identifiers is a problem of informatics rather than one of biology. Currently existing systems bear little if any relation to the biology. One advantage of the method proposed here is that if a gene is found to have an alternative start site this change will be immediately apparent: with the existing systems such a change may be difficult to identify.

Genomes currently are assembled from a number of pieces of DNA. Genes may be annotated within these pieces before the genome is completely assembled. Keeping track of the identifiers of these genes during assembly is a challenge. The method proposed here allows for easy updating of the identifiers and previous identifiers can be recorded with the use of olim. However if the assembly process requires a number of builds the olim notes may become lengthy and cumbersome. While it is desirable to keep track of as many previous identifiers as possible a compromise seems to be to keep only the last two identifiers preventing the identifier system from becoming overly cumbersome but still allow previous versions to be traced in the literature.

Chapter 5: Database files

5.1 Introduction

This chapter an outline of the database files used in this project is provided. The files provided here were some of the materials used in the annotation of the genome and the preparation of this thesis. The only additional materials used were a large number of text files used largely as temporary storage for examination of the genes.

5.2 Description of the included files

The genes annotated in this project are available on the attached CD in two formats – a text file of comma separated values (*.csv files) and a set of Microsoft Access files (*.mdb files). The CSV files can be imported in to any current relational database. The *.mdb files support the open database connective interface allowing the data to be used even in the absence of a database. Additionally many databases recognize the *.mdb format and can import these files directly.

On the CD the data is arranged into two tables: one of gene parameters and a second listing the exons and introns and a number of associated parameters. In the gene table for each of these genes the following parameters are available: the identifier, chromosome, strand, annotation history, the start and stop coordinates, length of coding sequence and protein length, the start and stop codons, the number of introns within the gene, base composition at all three codon positions, codon use, relative synonymous codon use, amino acid use, coding sequence, translated amino acid sequence, predicted relative molecular weight, predicted pI, dinucleotide use in the 1:2, 2:3, 3:1 and 1:3 codon positions, the effective number of codons, and the codon chi square. The type of evidence for each gene is also provided – cloned, transcriptional only, proteonomic only or prediction only. When a protein is an enzyme the Enzyme Commission (EC) number is also provided.

The hydrophobicity of the protein was calculated with the Kyte Dolittle scale (Table 5.1). The relative synonymous codon use, the codon chi square and the effective number of codons were determined as described previously (chapter 3). The molecular weight was determined by summing the relative molecular weights of the

amino acids in the protein less the relative molecular weight of water (18) when the residues lay internal in the protein. The ionization of the side chains and the COOH and NH₂ terminals of the protein at varying pHs was calculated; the predicted value of the pI - the pH at which the number of ionized residues is minimal - was determined to the first decimal place. The type of evidence supporting the annotation was derived from examining the published literature.

	Mass	pK ₁	pK ₂	pK ₃	KD value
Alanine	89.09	2.35	9.87		1.8
Cysteine	121.16	1.92	10.7	8.18	-4.5
Aspartic acid	133.1	1.99	9.9	3.9	-3.5
Glutamic acid	147.13	2.1	9.47	4.07	-3.5
Phenylalanine	165.19	2.2	9.31		2.5
Glycine	75.07	2.35	9.78		-3.5
Histidine	155.16	1.8	9.33	6.04	-3.5
Isoleucine	131.17	2.32	9.76		-0.4
Lysine	146.19	2.16	9.06	10.54	-3.2
Leucine	131.17	2.33	9.74		4.5
Methionine	149.21	2.13	9.28		3.8
Asparagine	132.12	2.14	8.72		-3.9
Proline	115.13	1.95	10.64		1.9
Glutamine	146.15	2.17	9.13		2.8
Arginine	174.2	1.82	8.99	12.48	-1.6
Serine	105.09	2.19	9.21		-0.8
Threonine	119.12	2.09	9.1		-0.7
Valine	117.15	2.39	9.74		-0.9
Tryptophan	204.23	2.46	9.41		-1.3
Tyrosine	181.19	2.2	9.21	10.46	4.2

TABLE 5.1. Values used to determine the relative molecular weight, predicted pI and hydrophobicity of the proteins. Abbreviations: Mass – relative molecular weight; pK_1 – the pH of the COOH terminal; pK_2 – the pK of the NH₂ terminal; pK_3 – the pH of

the side chain where applicable; KD value – the hydrophobicity assigned to the amino acid in the Kyte Doolittle scale.

The exon-intron table contains additional data on all the intron containing genes. Each gene has its own identifier and strand. For each exon the following parameters are available: its number, length, phase, base composition in all three codon positions, dinucleotide use in the 1:2, 2:3, 3:1 and 1:3 codon positions, the start and end coordinates and the exon sequence. For each intron its number, the 5' and 3' splice sites, the phase, base composition, dinucleotide use, length and sequence are given. Additonally a set of HTML files are included on the CD. These were created at the request of the examiners – Professor K. H. Wolfe and Dr J. O. McInerney. These files were not used in the preparation of this thesis.

Chapter 6: Gene organization and structure

6.1 Summary

In this chapter an over view of the nuclear genome is first presented. This is then followed by a study of the intergenic distances and their relationship to relative gene orientation. Genes that are convergently transcribed have on average the least separation while those that are divergently transcribed have the greatest. It is also found that in four of the fourteen chromosomes potential clustering of genes in the same orientation is present. This finding is sensitive to the quality of the annotation and should be treated with caution. Other properties of the genes that are examined here include gene length, predicted pI and protein hydrophobicity. Gene length seems to follow a single distribution that is neither normal or log normal. The predicted pI of the proteins is distributed bimodally with the trough close to the expected pH of the cytoplasm. A linear relationship between the number of acidic and basic residues in the proteins was found. The predicted mean hydrophobicity of the proteins is unimodal unlike the pattern found in *E. coli*.

Finally the Chargaff differences are examined. Pyrimidine and purine content was found to be very highly correlated with the purines exceeding their complementary pyrimidines by an almost constant factor of 50%. While this purine excess might be expected by Sybalski's rule the almost constant excess was unexpected.

6.2 Nuclear genome overview

The current version of the annotation predicts 5100 genes in the chromosomes (Table 6.1) with an additional 125 genes lying on unmapped contigs. Of these contigs the chromosome of origin is known only for three. Twelve conserved non coding sequences lying 5' of the internal *var* genes were identified and have been labelled as putative promoters. This designation seems reasonable but may yet prove to be incorrect. Only one third of the genes - 1712 (33.8%) - have identifiable products. Of the remainder there are 579 genes encoding proteins for which there is either proteonomic or transcriptional evidence which have been designated 'proteins of

unknown function' and 2809 genes have been identified as 'hypothetical proteins' for which there is no independent evidence.

There are 118 pseudogenes - 2.2% of the gene total. Of these 6 are *rifin* pseudogenes, 25 are *stevor* pseudogenes, 4 *rifin/stevor* pseudogenes and 51 *var* pseudogenes. Interestingly all of these are non functional paralogues of *P. falciparum* genes that code for variable and immunogenic surface proteins. Of the remaining 32 pseudogenes a functional paralogue can be identified for only four: three are the remnants of other surface exposed genes and one is a chitinase. Both these pseudogenes and their functional paralogues tend to be located in or adjacent to the GC rich isochors at the end of the chromosomes. The number of *var*, *stevor* and *rifin* pseudogenes may be the consequences of an above average mutation rate. There is some evidence for an enhanced mutational rate in these relatively GC rich regions but this needs additional confirmation (520). Alternatively the representation here may reflect an unconscious annotation bias: pseudogenes of multicopy genes are considerably easier to find than pseudogenes whose functional paralogues have but a single copy.

Of the nuclear genes there are 364 cloned genes, 994 with proteomic data only and 88 with transcriptonome only data. This does not include the 66 *var* genes or the 175 *rifin* and *stevor* genes. While representatives of all these multicopy genes have been cloned before, it was impossible to assign the clone to specific genes in the 3D7 genome. It is likely that additional genes have been cloned that have not been included in these totals but the number missing seems likely to be small. The cloned genes were used to check the correctness of the annotation. Agreement was complete with one exception – that of normocyte binding protein 2 on chromosome 13. This protein is involved in invasion of the reticulocytes (388) and its gene has been cloned (389). The sequence in chromosome 13 agrees with the cloned gene for the first part of its length whereupon the similarity disappears. This part of the chromosome 13 sequence is non coding suggesting a missassembly at this location.

Overall mean gene density is 200 - 250 per megabase - higher than that of multicellular eukaryotes. 2725 (53.3% gene total) genes have introns and there are 8232 introns – an average of 3.02 introns per intron containing gene. 1250 of these genes have only a single intron, 478 have 2 introns, 302 have 3 introns and 695 have more than three. The greatest number (33) of introns within a single gene is found in Pf12_1:119402w - a putative dynein heavy chain gene.

	Proportion	Proteins	Non proteins
Mitochondrion	83.4	3	19
Plastid	96.6	28	31
Chr 1	51.8	149	3
Chr 2	56.2	228	1
Chr 3	59.0	246	3
Chr 4	59.2	246	11
Chr 5	60.9	317	10
Chr 6	60.8	315	3
Chr 7	63.0	290	15
Chr 8	60.5	298	7
Chr 9	57.0	373	*
Chr 10	42.7	408	*
Chr 11	59.2	478	7
Chr 12	43.0	530	7
Chr 13	62.7	677	10
Chr 14	60.3	768	6

TABLE 6.1. Gene content by chromosome. Proportion is the percentage of the chromosome that is gene encoding. The non protein counts include pseudogenes and putative promoters. Abbreviation: chr - chromosome

Including introns the proportion of the nuclear chromosomes encoding genes varies from 42.7% (Chr10) to 63% (Chr 7). The difference may be a genuine biological feature or indicate that there are more genes to be found on the lower density chromosomes - a question that cannot be resolved with the presently available information. *A priori* there is no known reason why gene densities should be the equal in all chromosomes. The *raison d'etre* for the differences in chromosome length within a genome and the variation in chromosome number between organisms are also obscure but all these differences almost certainly reflect some as yet undiscovered underlying principle.

6.3 Organisation and spacing of genes

The genes are equally divided between the two strands with 50.2% lying on the Watson strand and 49.2% on the Crick strand. The genes initially appear to be organised in a somewhat haphazard fashion – a biologically implausible scenario. Exploration of the functional organisation of the genes will require a considerable amount of experimental work. In contrast investigations of the intergenic distances and the organisation of gene orientation are immediately possible.

6.3.1 Methods

The intergenic distances and the strand the genes lie within were extracted from the database. Denoting genes lying on the Watson strand as "+" (Watson genes) and those on the Crick as "-" genes (Crick genes) and writing these out generates strings similar to '++-+-++'. Within these strings the symbols will form clusters or 'runs.' A run may consist of but a single symbol. A run ends where the symbol in the run changes. A test of randomness of such sequences has been described (490). Under the assumption of randomness the expected mean number of runs (k) is

 $E(k) = 1 + 2 / (1/n_1 + 1/n_2)$

and the variance is

 $\sigma_{\kappa}^{2} = (2\nu_{1} \nu_{2}) (2\nu_{1} \nu_{2} - \nu_{1} - \nu_{2}) / [(\nu_{1} + \nu_{2})^{2} (\nu_{1} + \nu_{2} - 1)]$

It has also been shown (161) that if the number of both types of symbol is greater than 10 the total number of runs is approximately normally distributed. A z score for each sequence of genes can be determined from the formulae given and the probabilities determined from the usual tables. Additional tables are available for values where either n_1 or n_2 or both are less than 10.

The intergenic sequence lengths were also examined. Considering the genes in pairs there are four possible cases: +/+ (a Watson gene followed by a second Watson gene), +/- (a Watson gene followed by a Crick gene), -/+ (a Crick gene followed by a

Watson gene) and -/- (a Crick gene followed by a second Crick gene). Adjacent genes with the same orientation are referred to genes in parallel: genes whose 5' ends lie adjacent are referred to as divergently transcribed genes (divergent genes); and genes whose 3' ends lie adjacent are referred to convergently transcribed genes (convergent genes). Because there might again be a difference between the chromosomes the intergenic distances were calculated chromosome by chromosome.

6.3.2 Results

The mean intergenic distance (Table 6.2) lay between 1367.5 and 1853.9 bases. The distribution while similar between chromosomes is markedly non normal - a fact reflected in the relatively large standard deviations. The number of runs lay within that expected of a random distribution in 10 of the chromosomes. In chromosome 4-6 and 14 the number of runs exceeds that expected by chance.

The mean intergenic distance (Table 6.3) for divergent genes lies between 2 and 2.6 kilobases, suggesting that on average the 5' control regions may lie within 1.0 - 1.5 kilobase pairs (kbp) of the start codon. Convergent genes generally lie between 600 - 1000 bases apart, suggesting that the 3' control elements – where and if they exist - lie on average within 300 - 600 bases of the termination codon. Genes in parallel are on average separated by 1.5 - 2.0 kbp - consistent with the sum of the estimates from the other sets of genes.

	Intergenic distance	Observed runs	Expected runs
Chr 1	1749.2 (1360.4)	68	74.9 (6.0)
Chr 2	1614.5 (1144.1)	111	114.7 (7.5)
Chr 3	1517.1 (1226.1)	139	125.3 (7.9)
Chr 4	1582.4 (1315.7)	150*	124.8 (7.8)
Chr 5	1635. 3 (1326.4)	147*	126.0 (7.9)
Chr 6	1633.8 (1528.0)	149*	126.0 (7.9)
Chr 7	1521.1 (1113.0)	107	114.3 (7.5)
Chr 8	1781.1 (1995.9)	123	113.4 (7.5)
Chr 9	1614.1 (1806.3)	129	124.4 (7.8)
Chr 10	1548.2 (1322.5)	136	124.4 (7.8)

Chr 11	1729.4 (1489.8)	133	126.0 (7.9)
Chr 12	1508.2 (1274.4)	136	126.0 (7.9)
Chr 13	1367.5 (923.2)	136	125.2 (7.8)
Chr 14	1853.9 (1658.9)	141*	121.8 (7.6)

TABLE 6.2 Mean intergenic distances (standard deviation) in base pairs, observed numbers of runs and expected numbers of runs (standard deviation). Observed runs with p < 0.05 are marked with an asterisk (*).

	+/+	+/-	-/+	-/-
Chr 1	2124.7 (1738.3)	1034.2 (1026.1)	2455.1 (1514.1)	1481.6 (686.0)
Chr 2	1766.4 (820.5)	920.2 (767.7)	2227.8 (1576.1)	1734.5 (899.8)
Chr 3	1497.4 (863.4)	672.1 (361.9)	2374.0 (1640.2)	1540.8 (849.8)
Chr 4	1636.6 (1034.8)	795.9 (532.5)	2420.3 (1529.0)	1972.8 (1583.7)
Chr 5	1993.4 (1804.9)	814.0 (541.9)	2123.8 (1220.8)	1547.2 (810.8)
Chr 6	1785.9 (965.5)	1005.6 (2138.6)	2179.8 (1258.1)	1599.9 (1037.9)
Chr 7	1622.6 (877.5)	892.4 (854.6)	2014.1 (1110.9)	1849.3 (1139.3)
Chr 8	1884.8 (1529.7)	848.3 (809.7)	2564.9 (2822.5)	1616.2 (860.5)
Chr 9	1915.4 (2262.9)	735.4 (547.1)	2281.4 (1369.5)	1587.5 (1026.7)
Chr 10	1404.3 (917.4)	864.6 (927.0)	2218.4 (1564.7)	1797.3 (1108.8)
Chr 11	1640.6 (901.2)	888.4 (1226.0)	2316.4 (1506.8)	1953.1 (1623.0)
Chr 12	1358.4 (785.7)	776.1 (486.8)	2283.1 (1547.1)	1697.9 (1098.7)
Chr 13	1394.6 (802.6)	871.6 (658.5)	2083.3 (1160.4)	1461.8 (722.8)
Chr 14	1707.9 (1078.3)	992.3 (2493.8)	2550.9 (1539.9)	1662.4 (1085.2)

TABLE 6.3. Mean intergenic distances (standard deviation) in base pairs organised by chromosome and gene order.

There were also a small number of genes with very small intergenic separations (Table 6.4). The shortest (17 bp) was between two genes in parallel (-/-) Pf14_1:1668935c (hypothetical protein) and Pf14_1:1672341c (protein of unknown function). There are three convergently transcribed gene pairs - Pf4_1:354376w and Pf4_1:358450c both hypothetical proteins separated by 28 bases; Pf5_1:167649w (ATP dependent helicase) and Pf5_1:169549c (hypothetical protein) separated by 47

bases; and a 60 base separation between Pf6:1288240w (DEAD/DEAH box ATP dependent RNA helicase) and Pf6:1290868c (hypothetical protein).

	+/+	+/-	-/+	-/-
Chr 1	284	165	431	296
Chr 2	456	196	193	133
Chr 3	167	129	533	79
Chr 4	268	28	513	460
Chr 5	115	47	132	138
Chr 6	407	60	598	514
Chr 7	250	92	504	399
Chr 8	154	160	582	416
Chr 9	367	181	645	315
Chr 10	365	95	20	48
Chr 11	410	88	391	221
Chr 12	357	81	198	258
Chr 13	96	92	349	338
Chr 14	137	17	113	225

TABLE 6.4. Mimimum integenic distances in base pairs.

6.3.3 Discussion

On the chromosomes where kilobases may lie between genes, coordination of expression is much more likely to be the result of *trans* acting factors so gene arrangement - at least in theory – may be more flexible rather than the operons found in bacteria where groups of functionally related genes tend to lie in the same orientation. To test this hypothesis the runs test was applied to the chromosomal genes. This test is sensitive to the accuracy of annotation. A single gene in either orientation at a critical location can alter the significance of this test. Given that it is suspected that there may remain a small number of errors and overlooked genes the results here should be regarded with considerable caution. In contrast to this test the

remaining statistical tests are fairly robust with respect to the quality of the annotation.

With these caveats in mind there appears to be a statistically significant grouping of genes by orientation on four of the chromosomes here but why this difference exists between the chromosomes is not presently clear. This association of genes may be worth reinvestigating once the quality of the annotation has improved.

The intergenic distances have large variances and are not normally distributed. While it is possible to statistically compare the intergenic distances, the variances are such that even a statistically significant result is presently unlikely to offer additional biological insight. Genes in parallel (+/+ or -/-) - unless they form part of an operon or lie close together - are unlikely to interfere significantly with each others transcription or other mechanism of gene control. In contrast convergently (+/-) or divergently (-/+)transcribed genes might - in theory - interfere with one another's transcription. Because of this potential interference theory suggests that divergently transcribed genes should lie further apart than genes in parallel. Convergently transcribed genes may lie closer together, further apart or equidistant to genes in parallel depending on the relative importance of the 3' region. With the additional assumption that on average the control region of the gene will lie more closely to the relevant gene than its neighbour we can estimate the length of sequence where the control elements lie both in the 5' and 3' directions. From the figures here it seems that the 5' control elements lie within 1-1.5 kiliobase pairs of the start codon. 3' control elements may or may not exist for all genes here. Where they do exist it seems likely that they on average exist within 300 bases of the termination codon. Both these estimates should be taken only as crude approximations as the variation between genes is considerable. Given the large variances that exist the minimum intergenic distance was also sought as this might give additional insight in the location of putative control elements. The convergently transcribed genes had in general the smallest intergenic separation generally 80 - 200 base pairs (bp). Assuming again that even in these extreme cases the separation is divided equally between the genes the 3 control elements – if they lie in this area – must lie within 40 -100 bp of the termination codon. Alternatively they may lie within the neighbouring gene.

Considering the minimum separations alone the divergent genes had on average the greatest minimum separation (300 - 600 bases); convergent genes the least (200 - 600 bases); and the genes in parallel lay between these values (300 - 500 bases). Theory

would suggest that a minimum distance exists between genes to prevent interference during transcription. These results are consistent with this hypothesis. These gene may either have short 3' untranslated sequences or the transcription extends into the neighbouring gene. Experimental examination here seems desirable.

The divergently transcribed pair Pf10_1:986592w and Pf10_1:988552c (both hypothetical proteins) and separated by only 20 bases represent an interesting case. It is possible that both sets of promoters lie within this short sequence or that one or either of the promoters lie within the neighbouring gene. Alternatively this could be an annotation error. The genes have been reviewed with this in mind but while this is always a possibility, this does not seem likely at present.

6.4 Protein length, predicted pl and hydrophobicty

The collection of proteins within the genome has other parameters of interest including the distributions of lengths, predicted pI and hydrophobicity: these were also examined.

6.4.1 Methods

The predicted pI values, the number of acidic and basic residues and the length of the proteins were extracted from the database. Since the principal determinants of the pI are the acidic and basic residues, the number of basic (lysine, histidine and arginine) residues were plotted against the number of acidic residues (glutamate and asparate). Since the proteins vary considerably in length the hydrophobicity per residue was determined. The total hydrophobicity and protein length were extracted from the database. All numerical calculations were carried out Microsoft Excel.

6.4.2 Results

The distribution of protein length is clearly non normal (Figure 6.1). The lengths vary over three orders of magnitude: the shortest Pf5_1:869990w (open reading frame) has 17 residues and the longest Pf6:1122802c (protein of unknown function) 10,287 residues. The mean length is 775.7 residues, the median length 473.0 and the standard deviation of the length 876.0. Over half the proteins (2593) have between 100 and 550

residues. Given the shape of the distribution it might seem possible to fit a gamma function but this was not found to be the case.

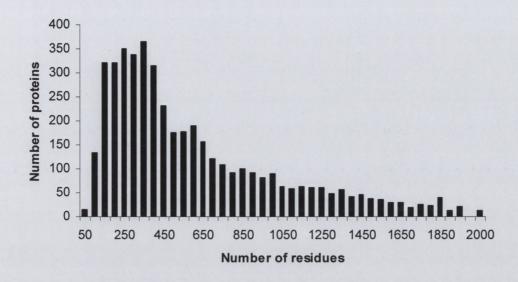
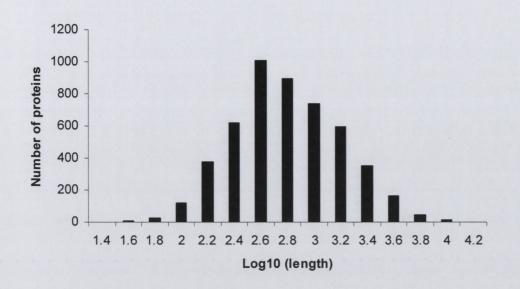
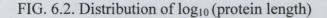


FIG. 6.1. Distribution of residues per protein. 388 proteins with more than 2000 residues have been excluded.

The visual appearance of the length plot (Figure 6.1) can be improved by first taking the \log_{10} of the protein length when it becomes more symmetrical (skewness 0.16) but remains non normal (kurtosis -0.27). These findings are incompatible with a log normal distribution.

The predicted pI shows a bimodal distribution (Figure 6.3) with the trough centered around 7.5.





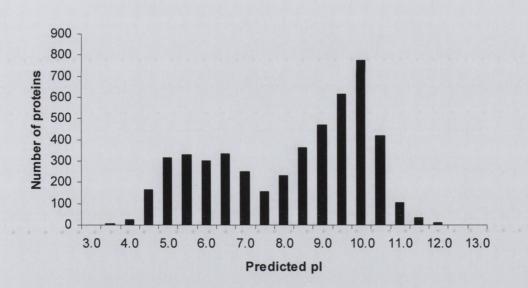


FIG. 6.3. Distribution of the predicted pI values.

With a single outlier there is a linear relationship between the acidic and basic residues (Figure 6.4). The basic residues exceed on average the acidic ones by $\sim 20\%$. This linear relationship may add to the buffering capacity of the proteins within the cell. The single outlier is Pf11_1:1947120w – the Maurer's cleft protein or erythrocyte membrane associated giant protein antigen 332 (Pf332).

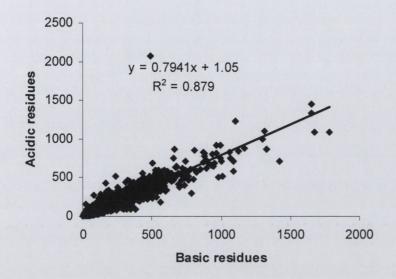
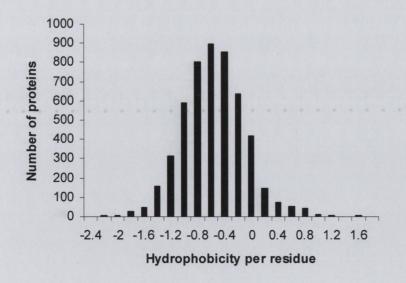
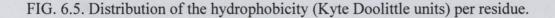


FIG. 6.4. The numbers of basic and acidic residues. $F = 35,085.3 \text{ p} < 10^{-10}$

The histogram of normalized hydrophobicity (Figure 6.5) shows a symmetrical unimodal distribution centered on the mean (-0.666 units per residue) and with a standard deviation of 0.467 units per residue.





6.4.3 Discussion

While no biological theory exists to give guidance as to the expected distribution of protein lengths in this genome, it seems reasonable to presume that protein lengths within a genome obey some as yet undiscovered law. This presumed law may be influenced by the size of the genome, the GC content, whether the organism is parasitic or free living, the domain it belongs to - eukaryote, prokaryote or archaea – as well as other factors. No known reason exists for any of these curves to fit a well known mathematic distribution such as the normal or log normal. This like other empirical findings will have to await the development of a satisfactory theory of genomes.

The cytoplasmic pH of the parasite is not known with certainty but is thought to be \sim 7.4 (Brey, University of Liverpool, personal communication): the trough values here are consistent with this. The trough presumably exist because of solubility concerns: a protein is at its least soluble when the solution is at the protein's pI. By avoiding the cytoplasmic pH the solubility of the proteins at translation is maximal.

A curious feature is the large peak in the higher pH range. This exists presumably because of the relatively large amounts of lysine encoded in the genome. How much of this high value peak is actually present after post translational modifications remains to be seen.

Since the process of metabolism generates acidic materials it makes biological sense that the basic buffering capacity exceeds that of the acidic. This is a difficult hypothesis to test experimentally as changing large numbers of acidic and basic residues would almost certainly disrupt the metabolism of the organism to the extent to render it non viable.

The reason for the linear relationship between these types of residues is not known. This relationship has not prevented the proteins from having a considerable range of pI values. This fixed proportionality between these hydrophilic residues is not accounted for in current biological theory. This finding suggests that there is some selective force acting here requiring the codons to covary in order to maintain this proportionality. This proposed selective force is likely to be at least in part a requirement for the overall shape of the protein. Replacing all or almost all the positively charged residues in a protein with negatively charged ones would seem likely to alter the protein's three dimensional shape and function.

On the acidic-basic residue plot one large protein - erythrocyte membrane associated giant protein antigen 332 (Pf332) - with a predicted molecular weight 688 kilodaltons (4) is an outlier. This protein of unknown function is found within the structures known as Maurer's clefts on the erythrocyte surface. Its composition is highly unusual: of its 6093 residues, 1718 are glutamatic acid. The predicted pI is 3.5 the third lowest in genome with only Pf14_1:1672341c (hypothetical protein) and pf13_23:1642938c (casein kinase 2 regulatory subunit) whose predicted pI of 3.1 and 3.4 respectively are lower. Pf322 has two exons of which the N terminal part of the first exon encodes a typical protein sequence. The vast bulk of the protein is composed of glutamic acid (E) rich repeats of which VSEEIPVEEKS and QLVPEEIKEE are typical. The function of these repeats is unknown. Since this protein is exposed to the immune system it seems reasonable to presume that these repeats interfere in some what with immune recognition. Experimental investigation seems desirable.

The unimodal distribution of hydrophobicity per residue contrasts with a similar examination of 999 proteins in *E. coli* that showed a bimodal distribution with the smaller second peak of greater hydrophobicity being composed of integral membrane proteins (280). The significance or otherwise of this difference is not known as this examination does not appear to be been performed for other organisms. Either of these patterns may be the norm; alternatively there may be a continuum of values between these extremes; or they may reflect a difference between eukaryotes and prokaryotes. Additional patterns may be found when other organisms are similarly examined. These questions can only be addressed by examining additional genomes.

6.5 Nucleotide content of the protein genes

In 1967 Szybalski discovered an asymmetry in the sense and antisense strands of the coliphage lambda (497). The protein encoding strand was invariably richer in purines than the complementary strand. Szybalski's rule has been since confirmed in other organisms. Like Chargaff's second rule, the *raison d'etre* for this rule is not yet known. Forsdyke has proposed the term 'Chargaff differences' for the A - T (weak) and G - C (strong) content of genes (141). Since no data has been published on these differences in *Plasmodium* this seemed to be worth investigating.

6.5.1 Methods

The base composition of the genes and their length was obtained from the database. The weak (A - T) and the strong (G - C) Chargaff differences for each gene were computed. These were then 'normalized' by dividing the differences by the length of the protein.

6.5.2 Results

In all the protein encoding genes the weak Chargaff difference (A - T) was positive, while in 375 genes the strong Chargaff difference was negative. In these latter genes no obvious pattern could be discerned. The weak and strong differences (Figure 6.6) are highly correlated (F = 7272.0 p < 10^{-10}) but after normalizing the differences by dividing by the length of the protein, this correlation disappears (R² = 0.001 p > 0.1). The correlation of Chargaff differences is an artifact: longer proteins have more bases and hence greater Chargaff differences.

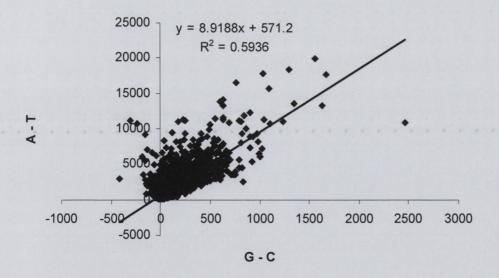


FIG. 6.6. Weak (A-T) and strong (G-C) Chargaff differences in the protein encoding genes.

The distribution of the normalized Chargaff differences (Figures 6.7 and 6.8) are markedly dissimilar and both distributions are non normal. The strong differences have a mean of 0.124 and a standard deviation of 0.093 (0.124 + - 0.093) while the

corresponding values for the weak differences are 1.793 +/- 0.144. Both distributions are platykurtic (kurtosis of 1.795 and -0.010 respectively) and skewed in opposite directions (skewness of 0.138 and -0.393 respectively).

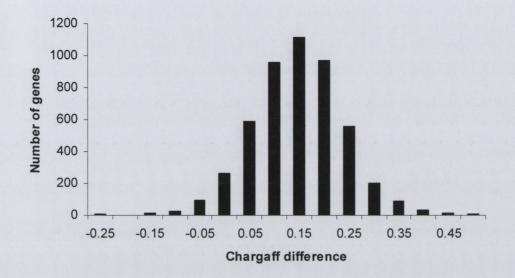


FIG. 6.7. Distribution of the normalised strong (G - C) Chargaff differences.

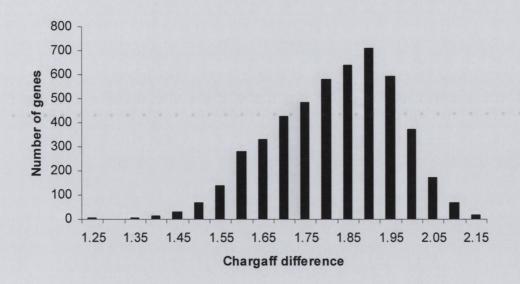
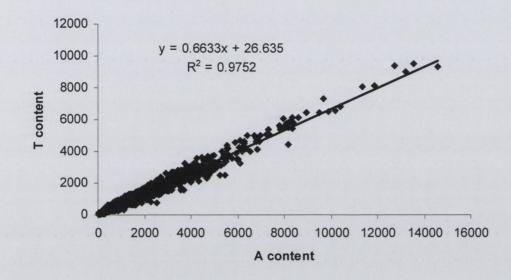
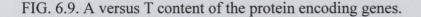


FIG. 6.8. Distribution of the normalised weak (A - T) Chargaff differences.

The distribution of the Chargaff differences was neither normal nor uniform and the symmetry - particularly of the strong differences - suggested that a relationship might

exist between the purine and pyrimidines in these genes. The number of adenosines was plotted against the number of thymidines (Figure 6.9) and the number of guanines against the number of cytosines (Figure 6.10). Correlations were found between these variables: for A vs T, the R² was 0.975 (F = 194,119.42, p < 10⁻¹⁰) and for G vs C, R² was 0.920 (F = 56,639.86, p < 10⁻¹⁰) confirming the impression given by the Chargaff differences.





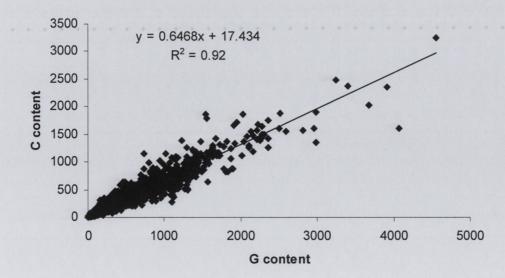


FIG. 6.10. G versus C content of the protein encoding genes

6.5.3 Discussion

Szybalski's rule was found to be true for all the genes in this annotation. Because of their relatively high purine content this rule had been found earlier to be of use in identifying potential protein encoding regions within the genome. As a consequence of Szybalski's rule and Chargaff's second rule within the non exonic regions of the genome that the pyrimidine content must exceed the purine content and this expectation has been confirmed for the introns (*vide infra*).

These findings here can be summarized as follows: within the protein encoding genes the purine content is 1.5 times that of the pyrimidines. This rule explains the symmetry found here in the Chargaff differences and is consistent with and extends Szybalski's rule. This rule has not been reported before. While Szybalski's rule suggests that this rule may be true in other genomes it needs to be investigated in other genomes before any conclusions should be drawn. In the *P. falciparum* genome, this relationship holds over four orders of magnitude. While speculation is possible about the origin of this rule it seems sensible to first examine test for its presence in other genomes as these may provide useful guidance to formulate meaningful hypotheses.

Analysis of codon use

7.1 Summary

In this chapter a number of investigations into codon use within *P. falciparum* are reported. Base use in *P. falciparum* by codon position was first investigated by Saul and Battistutta (532) who found in the 22 sequences they examined that (1) the A/T ratio in the coding sequences was 1.68 (2) the A+T content increased in the $1^{st}-2^{nd}-3^{rd}$ codon position (3) A or T were preferred in the third codon position and (4) the A + T content of coding sequences was 69.0% and that of the non coding sequences was 86.0%. They also found that while the overall dinucleotide use was close to random that the dinucleotide GC content was lower than expected. Given the somewhat larger data set available here these patterns seemed worth reinvestigating.

D'Onofrio and Bernardi have reported a 'universal compositional correlation' between the GC contents of codon position 1 and 2 and that of position 3 (97). Later D'Onofrio *et al* (1999) reported a correlation between the GC3 content and the hydrophobicity of the protein (98). Both of these correlations were investigated here.

While it is likely on biological grounds that the frequency of encoding of amino acids in the genome is positively correlated with the frequency of use in the translated proteins the precise nature of this relationship is not yet known in this organism. While currently data is lacking on most of the proteome the frequencies of the encoded amino acids and codons found here were documented.

Correspondence analysis is a popular method in genome analysis for seeking underlying relationships between codon use and the underlying biology. The first use in biology appears to have been by Hill in 1974 who examined 1333 coding sequences each of over 100 codons and identified four main trends: the first two relating to gene expression, the third to hydrophobicity and the fourth to location on the leading or lagging strand (198). This has since become a popular technique to identify trends in genomes (68, 147). The idea behind correspondence analysis is to use a distance function (or metric) to determine the distance between various genes, put these values into a matrix and to obtain the eigenvectors and eigenvalues of this matrix. The contribution of the eigenvector to the variation within the matrix is given by its associated eigenvalue. In theory if the variance within the matrix can be "explained" by the set of eigenvectors and if what the eigenvectors represent biologically can be identified then some insight into the rules governing genome composition may be gained.

Other authors have investigated variation of the codon chi square (χ^2) and the effective number of codons (N_c) within a genome. Wright proposed a formula for the expected value of N_c assuming that any codon bias found is only due to G+C content and proposed that genes lying at some distance off this curve were likely to be subject to translation selection. Some theoretical issues have been raised about this statistic in dos Reis *et al* (119) but these are beyond the scope of this work. Comparison between the theoretical and expected N_c values was investigated here.

Variation of the codon chi square and the effective number of codons with gene length has been investigated before. Sharpe *et al* (440) calculated the codon chi square statistic for genes in a number of different organisms and found similar values for all the organisms. The authors regressed the χ^2 value against the gene length and found the slope of the regression line was found to be negative. From biological principles the authors expected that the χ^2 - a measure of codon variability - would tend to increase with the length of the gene. Their expectation was correct: the χ^2 value does increase with the length of the gene. The error in this paper will be explained later. Variation of N_c with length has also been investigated (304): this paper has a similar error to that found in Sharpe *et al*.

Selection base use around the beginning and the end of the coding sequences has been reported before. Kozak has shown that the base immediately 3' of the start ATG (the +4 base) has a considerable influence on translation initiation (242). Similar investigations into the termination codons reveal that the base immediately following the stop codon (the +4 base) has an influence on the efficiency of the stop codon (276, 334). Sequences form 148 organisms show this pattern (496) and the +4 base has also been shown by chemical linkage to be part of the stop signal (368).

The peculiarities of the stop codons have been investigated previously (292, 483). It is known that around genuine stop codons there are frequently multiple 'off frame' stop codons – the so called ambush hypothesis (433) - and that the amino acid composition near the stop codon and the base immediately following the stop codon (the + 4 base) are biased (276). It is rare not to have within the last 10 amino acids both a bulky hydrophobic group (Y, F or W) and a positively charged residue (K, R or H). Choice

of termination codon has been investigated in *E. coli*, *B. subtilis* and *Saccharomyces cerevisiae* where it was found that highly expressed genes favored the use of TAA while genes with low protein expression favor TGA (437). Alterations in the sequence of an antibody expressed in *E. coli* 5' to the termination codon increased the expression of the protein 10 fold (383). Efficiency of the stop signal is controlled by the +4 base (367). In neither initiation nor termination is are the effects limited to the +4 base with the three bases lying 5' of the start site and up to six bases following the termination codon have been implicated. To date base use at these positions has not been investigated before in *P. falciparum* either computationally or experimentally.

7.2 Methods

The mean and standard deviation of the percentage base content of all three codon positions was extracted from the database. Two additional examinations on this data were performed. The GC3 content of a gene has often been used in investigations of various genomic parameters. Possible relationships between the base composition in the first and second codon positions and the GC3 content were explored by plotting these values against the GC3 content. Amino acids with A in the second codon position (A2) tend to be hydrophilic while those with T in the second position (T2) tend to be hydrophilic. The difference A2 - T2 was plotted against the hydrophobicity of the protein to investigate this potential relationship.

The encoded amino acid frequencies were extracted from the database and tabulated. They were then ordered by increasing frequency of occurrence within the genome and the least frequent (tryptophan) was given the value 1, the next least frequent (cysteine) the value 2 up to the most frequently occurring (asparagine) with the value 20. The ordinal number of the amino acid use was here referred to as the 'index.' The index was plotted against the frequency of amino acid use and the logarithm of this frequency. Likewise the codons were ordered by frequency of encoding, indeed from 1 to 61 - the stop codons were omitted from the analysis here.

As originally conceived correspondence analysis is defined only for integer valued variables. Much of the data this method has been applied to in the biological literature is not integer data and as such the method used therein is more properly described as an extension of correspondence analysis to non integer data. Also the basis for the choice of a chi square metric has yet to been justified on biological grounds. However

this metric has been used for to examine a number of genomes and shown to be capable of isolating factors of biological interest (305). For this reason and to ensure compatibility with previous work, this metric was used here.

To perform a correspondence analysis the data is placed in an N x M matrix, the chi square values for each value in this matrix is determined and principal components analysis is applied to the resultant matrix. The chi squared value in each position is the observed less the expected value squared and the divided by the expected value. In symbols, let x_{ij} be the observation at the intersection of the ith row and jth column and then

 $\chi_{ij}^2 = [x_{ij} - (x_{i.} x_{.j} / x_{..})]^2 / (x_{i.} x_{.j} / x_{..})$

where χ_{ij}^2 is the value of the new matrix at position ij, Σx_i is the sum of the ith row, $\Sigma x_{,j}$ is the sum of the jth column and $\Sigma x_{,i}$ is the grand total of all the x_{ij} values. Details of extracting eigenvectors and their corresponding eigenvalues can be found in many books (375, 435).

While the amino acid data may be used directly in correspondence analysis there are theoretical problems with codon use of the degeneracy in codon amino acid encoding. To overcome these Stenico et al (474) defined a new index - the relative synonymous codon use (RSCU) - which controls for the unequal representation of amino acids in the codons and subsequently used this measure in genome analysis (441). The formula for its determination has been described earlier in this work. When used to analyse codon data, it is usual to include the 59 synonymous sense codons, which may generate up to 58 axes. The alternative analysis of the 20 standard amino acids may produce up to 19 axes. The question arises as to the number of these axes that should be examined further. Several methods have been suggested but one popular method is the scree plot (76): this was used here. The remaining difficulty - which may be considerable - is the identification of the biological meaning of these new axes. Since this is a linear model the axes are usually identified by regressing the values derived from these axes against other variables of biological interest including the hydrophobicity or the GC3 content of the gene. The correspondence analysis was performed with Minitab 14 (Minitab).

The value of the codon chi square for each gene was determined as described in Chapter 3.1. The theoretical value of the N_c of a gene as proposed by Wright is

$$N_c = 2 + s + 29/[s^2 + (1 - s)^2]$$

where s is the corrected GC3 (GC3s). The GC3s is the sum of the number codons ending in C or G less the number of methionine and tryptophan codons divided by the total number of codons less the number of methionine and tryptophan codons. The correction is required because the methionine and tryptophan codons are non degenerate.

The plots of these values against the gene length showed marked heteroscedacity. Since this was present, the dependent variables (χ^2 , N_c) were plotted against a number of additional variables including the base content of the genes and various powers of the gene length.

Base use around the start and stop codons was examined in two ways. The first was to tabulate the occurrence of the three bases lying immediately 5' of the start codon (positions -3, -2 and -1) and the base immediately following it (+4). Similarly base use immediately after the stop codon (the +4 position) was recorded.

The second method used here was to obtain the mean base use in all three codon positions fifty bases after the start codon and fifty bases before the stop codon. The choice of 50 codons was somewhat arbitrary. This length was chosen after previous examinations of shorter lengths showed that the variations in base use stabilized within 20-30 codons either after the start codon or before the stop codon. The 20 odd additional bases were included here to provide a comparator with the remainder of the gene. The occurrence of positively charged residues (H, K or R) and bulky hydrophobic (F, W and Y) residues within the last 10 codons of the gene was also recorded.

7.3 Results

The overall AT content of the genes is 76.4% and the AT content of the third codon position (AT3) is 82.8% both of which reflect the genome's high AT content (Table 7.1). Codon composition here has both typical and atypical features. The G content in

position 1 exceeds that of position two which in turn is greater than that of position three: symbolically G1 > G2 > G3. The T content has the inverted pattern with T3 > T2 > T1. The mean purine content of the second and third codons positions is 58.4% and 50.0%. A is the most common base in the second codon position. The mean purine content of first position (68.3%) exceeds the mean pyrimidine content of the first position. While C is normally the least common base in all three positions, here C2 > G2.

	Α	С	G	Т
1	44.2 (6.0)	10.0 (2.6)	22.7 (5.6)	23.1 (4.8)
2	47.6 (8.2)	13.0 (4.1)	10.8 (3.6)	28.6 (5.3)
3	40.1 (5.0)	8.1 (2.6)	9.9 (2.8)	41.9 (5.4)

TABLE 7.1. Codon composition in *P. falciparum* by base and position. The figures are the mean (standard deviation) in percentages.

Base use in all three positions have distributions similar to that shown for T3 (Figure 7.1). None of the twelve are either normally or uniformly distributed – distributions that could compatible with random base use. Instead all have reasonably symmetrical distributions with a single peak. All the values lie within four standard deviations of the mean but not three.

Regression of the GC3 on C1, C2, G1 and G2 yields the following equation:

GC3 = 0.014 + 1.16 C1 + 0.234 G1 - 0.464 C2 + 0.659 G2

The constant in the equation is not significantly different from zero (p = 0.984) and can be ignored. All the remaining regression coefficients are each significant. The overall F value is 22,363.86 with $R^2 = 0.948$ and $p < 10^{-10}$. Plotting the A2-T2 value against the hydrophobicity of the protein suggested a linear correlation: regression analysis confirmed this.

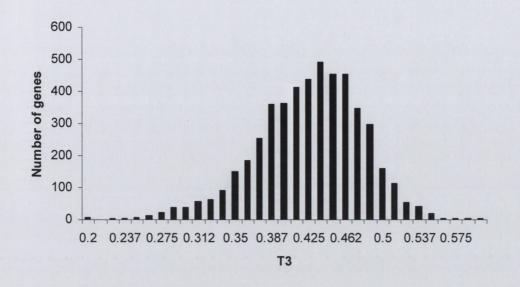


FIG. 7.1. Percentage T content in the third codon position. The skewness is -0.499 and the kurtosis is 0.630.

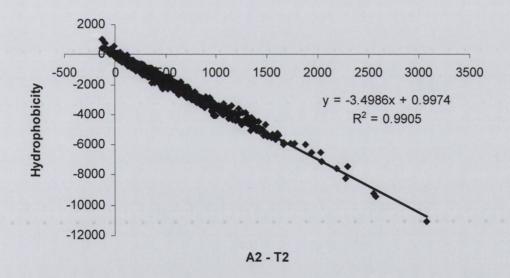


FIG. 7.2. Correlation of hydrophobicity and (A2 – T2). F = 529863.4, p < 10^{-20} . The GC3 content was also correlated with the hydrophobicity of the protein.

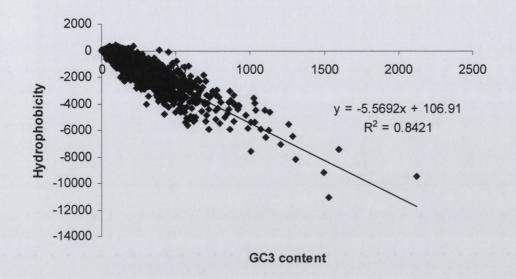


FIG. 7.3. GC3 and hydrophobicity. F = 26,328.3 and $p < 10^{-10}$.

With this correlation and the GC3-hydrophobicity correlation, it was natural to ask is there was a correlation between the A2 - T2 content and the GC3 content. Such a linear correlation also exists.

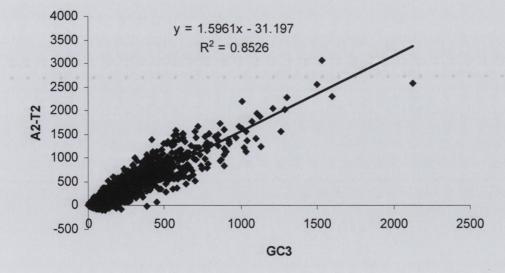


FIG. 7.4. Correlation of GC3 and (A2 – T2). F = 26328.3, $p < 10^{-10}$

Encoded amino acid use is very unequal with two (lysine and asparagine) constituting 26.3% of the total. In contrast the ten least frequently encoded amino acids make up 26.8%. Charged amino acid use differs between the basic residues. The basic residues are heavily biased in favour of lysine (11.8%) over the arginine (2.6%) and histidine (2.4%). The acidic residues are more equally encoded with aspartic (6.5%) and glutamic (7.0%).

Ala	1.9	GIn	2.8	Leu	7.6	Ser	6.4
Arg	2.6	Glu	7.0	Lys	11.8	Thr	4.1
Asn	14.5	Gly	2.8	Met	2.2	Trp	0.5
Asp	6.5	His	2.4	Phe	4.4	Tyr	5.7
Cys	1.8	lle	9.3	Pro	2.0	Val	3.8

TABLE 7.2. Encoded amino acid (percentages) within the *P. falciparum* genome.

The plot of the index against the frequency of amino acid content is well fitted by an exponential curve (Figure 7.5). The base of the exponent is 1.13701 [= exp(0.1284)]. Ignoring for simplicity any adjustments that would be required if the percentage of tryptophan were increased the observed value is only 36.4% of the predicted. The two least frequently used - CGG (0.027%) and CGC (0.041%) - and the two most frequently - AAA (9.6%) and AAT (12.5%) – codons while visually outliers are acceptable statistically. The remaining 57 codons differ in their frequency of use by a factor of 65 fold. On average the more AT rich the codon the higher its index of use.

On the semi log plot it is clearer that tryptophan is an outlier and that the remaining values are well fitted by a regression line. The expected value of tryptophan from the regression line is 1.37%.while the actual value is 0.5%.

Examination of codon use reveals a similar picture. Ranking the codons in ascending order in frequency of use and plotting this against the log_{10} of the frequency of use gives to a very good approximation a straight line.

The two least frequently used codons - CGG (0.027%) and CGC (0.041%) - and the two most frequently - AAA (9.6%) and AAT (12.5%) – while visually outliers are acceptable statistically. The remaining 57 codons differ in their frequency of use by a factor of 65 fold.

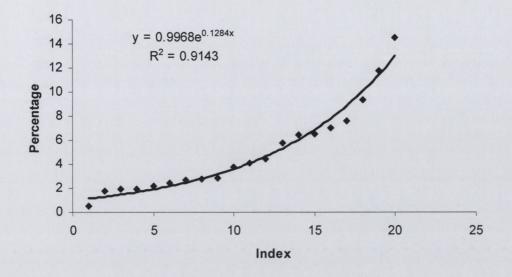
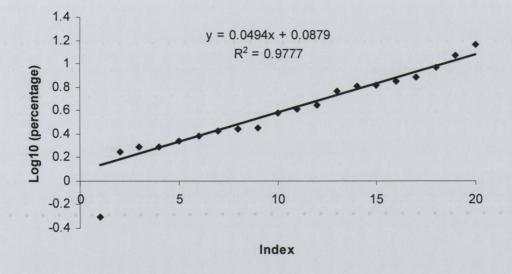
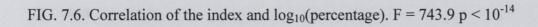


FIG. 7.5. Index and percentage of encoded amino acids.





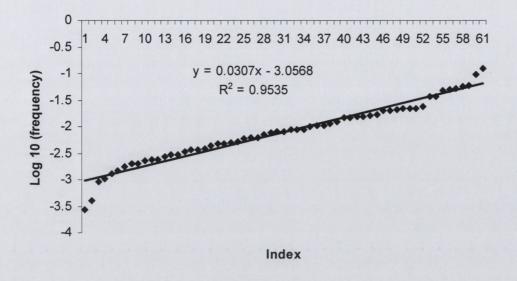
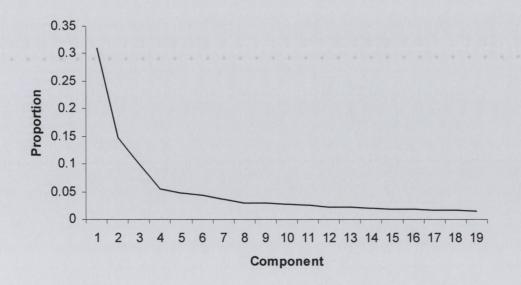
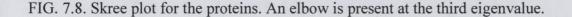


FIG. 7.7. Index of codon use and frequency of codon use. (F = 1209.1, $p < 10^{-40}$)

The eigenvalues of the protein correspondence analysis were computed and the scree plot drawn (Figure 7.8). The plot shows an elbow at the third eigenvalue with the first axis accounts for 31.0% of the variation between the proteins while the second and third account for 14.8% and 10.1%. The remaining 16 axes each account for less than 5%.





The contribution of the individual amino acids to the first two axes is shown in Figure 7.9 with the corresponding values in Table 7.3. The first two axes represent the GC content in the first two codon positions and the hydrophobicity respectively. The third correlates with the decreasing molecular weight of the amino acids. The plot is heteroscedastic so the non parametric Spearman rank correlation (470) was used (r = -0.530, p = 0.016).

Axis	s 1	Axis	s 2	Axis	s 3
Ala	0.675	Phe	0.300	Ala	0.173
Trp	0.519	Leu	0.185	Pro	0.156
Gly	0.444	lle	0.171	Gly	0.153
Pro	0.401	Tyr	0.166	Asn	0.143
Arg	0.213	Cys	0.127	Met	0.116
Val	0.212	Trp	0.078	Thr	0.077
Thr	0.149	His	0.024	Ser	0.074
Cys	0.114	Val	-0.006	His	0.053
Glu	0.074	Thr	-0.023	Cys	0.043
Gln	0.072	Ser	-0.027	Val	0.041
Leu	0.061	Lys	-0.028	Phe	0.001
Ser	0.035	Met	-0.032	Tyr	-0.004
Phe	-0.018	Ala	-0.045	Gln	-0.02
Asp	-0.031	Arg	-0.047	Trp	-0.025
Met	-0.040	Pro	-0.058	Leu	-0.033
Lys	-0.063	Gln	-0.067	lle	-0.036
lle	-0.084	Gly	-0.094	Arg	-0.045
His	-0.118	Asn	-0.118	Asp	-0.072
Tyr	-0.141	Asp	-0.167	Lys	-0.143
Asn	-0.287	Glu	-0.190	Glu	-0.229

TABLE 7.3. Residue values in the first three axes.

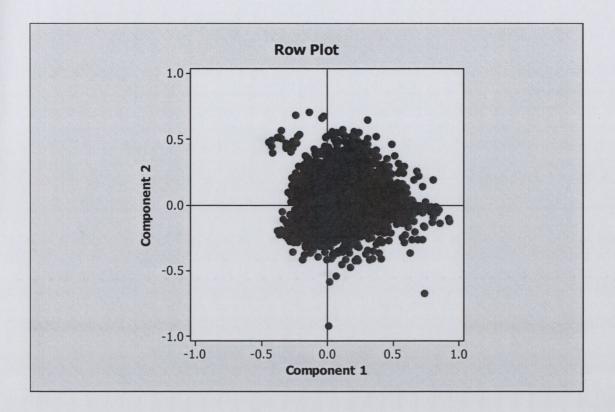


FIG. 7.9. Protein plot with respect to the first two axes.

The plot of the proteins along the first two axes identified by correspondent analysis is shown in Figure 7.9. The proteins form a single cluster close to the origin. While there are a few outliers there are no sizable sub groups.

The scree plot of the RSCU values had an elbow at the second eigenvalue. There was no single dominant trend with the largest eigenvalue was 0.613 and the second largest 0.510. The first 12 axes accounted for 50.5% of the intergenic variation in codon use. The codon plot with the first two axes is similar to the protein plots with no sizable subgroups identifiable (Figure 7.10). This was not unexpected given the small proportion of the total variation attributable to each axis. The first axis correlated with the A3, T3 and to a lesser extent with the C1 content. The second axis correlated with the C3 and G3 content. No correlation with gene orientation was found

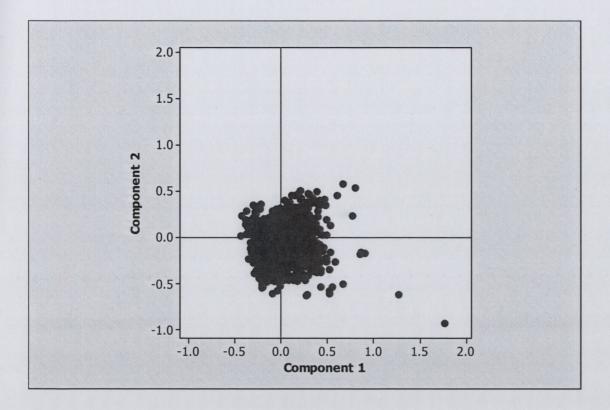


FIG. 7.10. Codon plot with respect to the first two axes.

On the Nc-GC3s plot (Figure 7.11) once two outliers are excluded - Pf8_6:70072w (open reading frame) and Pf11_1:129320c (early transcribed protein) - the genes cluster close to the theoretical line. The early transcribed proteins (*etramps*) including Pf11_1:129320c comprise small multicopy gene families. The codon use of Pf8_6:70072w is atypical of this family.

A highly significant negative correlation was found only against the log of the length of the gene. A plot of the length of the gene and the N_c value is shown in Figure 7.12. Marked heteroscedacity is evident. Inclusion of the log of the gene length in the regression significantly reduced the heteroscedasicity and changed the sign of coefficient of the length. The plot of the codon chi square and the gene length is shown in Figure 7.13, It too is clearly heteroscedastic. Regressing the codon chi square against both log_{10} (protein length) and protein length gave this equation:

 $\chi^2 = 129.777 + 0.005$ (protein length) -29.473 log₁₀ (protein length) (R² = 0.332, F = 1,225 and p < 10⁻¹⁰)

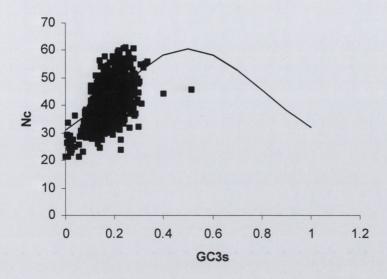
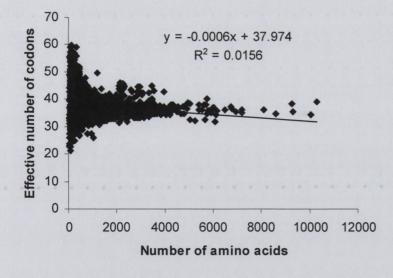
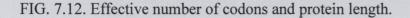


FIG. 7.11. Plot of the effective number of codons (N_c) and GC3s in *P. falciparum*. The solid line is the theoretical curve.





Regression of the N_c against both log_{10} (protein length) and protein length: N_c = 47.0 + 0.0008 (protein length) - 3.760 log₁₀(protein length) R² = 0.051, F = 130.97, and p < 10⁻⁶

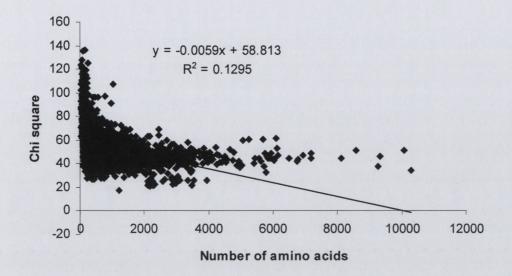


FIG. 7.13. Correlation of the codon chi square and protein length.

Base use immediately preceding the initial ATG shows the usual preference for purines with the -3 site (89.2%) being most biased (Table 7.4). The avoidance of T in these positions given its frequency in the non coding regions is noticeable.

	Α	С	G	T
-3	83.4	4.1	5.8	6.7
-2	64.6	7.7	7.7	20.0
-1	60.5	6.4	10.5	22.6

TABLE 7.4. Base use preceding the initial ATG codon.

-3/+4	Α	С	G	Т
Α	40.1	9.8	17.9	15.5
С	2.0	0.4	1.0	0.7
G	2.9	0.6	1.3	1.0
Т	3.1	0.7	1.7	1.3

TABLE 7.5. Percentages of genes grouped by -3/+4 bases

Subdivision of the bases by the -3 and +4 base is shown in table 7.5. The canonical - 3A/+4G occurs only in 17.9% of cases suggesting that suboptimal initiation sites are common and that alternative starts sites may be used. Genes with T at the -3 site have purine in the +4 position in 70.9% of cases. Only 148 (3.0%) of the genes have no purines in either the -3 or +4 sites.

Mean base use in the first codon position differs from the other two positions (Figure 7.14). The C content of the first position (C1) reaches a steady value (~10%) after the first six codons. The A1 content at the +4 position (48.6%) is slightly above the values found in the remainder (~45%). The T1 content is relatively low in the second codon (16.2%) but rises rapidly to the value found along the remainder of the protein (20-25%). The G1 content at the +4 position is elevated (28.2%) compared with the remainder of the protein (~20%).

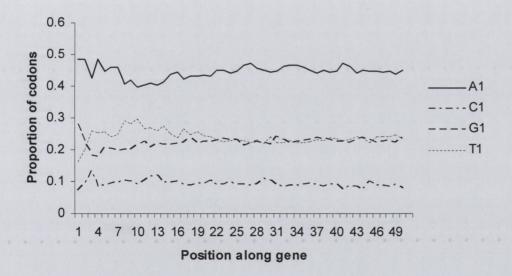


FIG. 7.14. Base use in the first codon position. Codons are numbered starting after the initial ATG

Base use in the second position follows a different pattern (Figure 7.15). The C and G use in this position (C2 and G2) again varied only slightly. The C2 content was greater on average than that of the G2 and this difference rises further down the protein. The A2 content fell from 48.4% to 39.7% over the first 11 bases. This fall in the A2 content was matched by a rise in the T2 content. These gradients reverse after the first 15-20 codons.

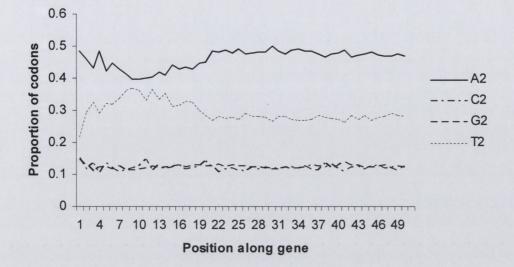


FIG. 7.15. Base use in the second codon position. Codons are numbered starting after the initial ATG.

Little variation is found in third codon position use (Figure 7.16). Over the first 50 codons (excluding the initial ATG) the proportion of A and T in the third position (A3 and T3 respectively) varies ~40-45% with C3 and G3 both stable ~10-15%%.

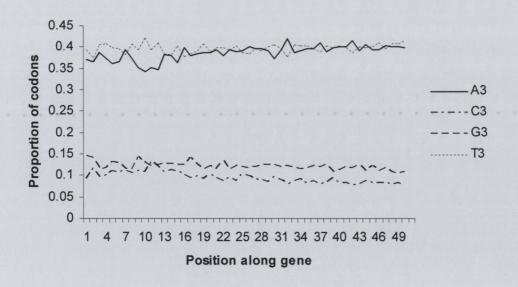


FIG. 7.16. Base use in the third codon position. Codons are numbered starting after the initial ATG.

Termination codon use is 70.0% TAA, 9.6% TAG and 20.4% TGA. This biased use of termination codons is typical of other organisms but the reasons for this pattern of choice of codons in any organism remains unknown. There was no association of termination codon choice with GC content or length of the gene. Expression level data is not available to test for an association there.

The frequency of the base following the termination codon differs somewhat between the termination codons (Table 7.6). The TAA codon is followed by A (56.2%), T (29.0%), G (8.3%) and C (6.5%) of the time. The corresponding figures for TAG are: A (51.5%), T (22.8%), G (17.3%) and C (8.3%); and for TGA: A (50.7%), T (33.8%), G (7.9%) and C (7.6%). The termination codon is followed by a purine in 60-70% of cases and the use of G after TAG is more than twice the frequency of its use after the other termination codons.

	Α	С	G	Т
TAA	39.4	4.6	5.8	20.3
TAG	4.9	0.8	1.7	2.2
TGA	10.3	1.6	1.6	6.9

TABLE 7.6. Stop codons by +4 base as percentages of all protein encoding genes.

The A1 content rises over last 5-6 residues while the C1 content declines (Figure 7.17). In this region of the genes T1 and G1 use was approximately equal up to the last 5-6 codons when the G1 content declines slightly. The approximate equality of T1 and G1 is similar to that found along the remainder of the gene after the first 20-25 codons. While the A1 content rises over the last 5-6 codons, the C1 content remains steady.

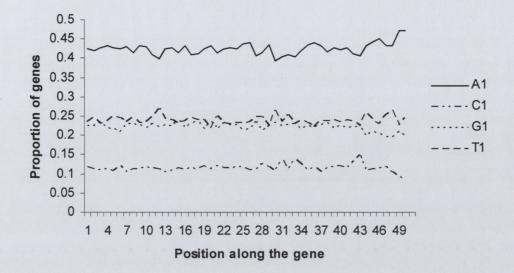


FIG. 7.17. Mean base use in the first codon position for the last 50 codons 5' of the termination codon.

Within 20 codons of the termination codon the T2 content declines slightly and the A2 content rises with a consequent increase in the mean hydrophilicity of this part of the protein (Figure 7.18). Like the bulk of the proteins the C2 content slightly exceeds the G2 here and both remain stable.

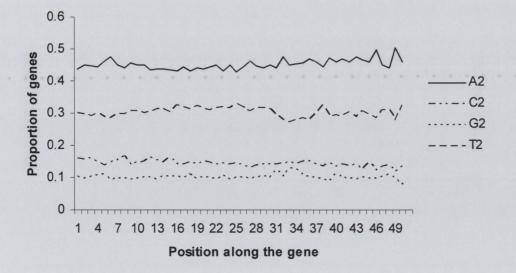


FIG. 7.18. Base use in the second codon position for the last 50 codons 5' of the termination codon.

In the last 20 codons on average the use of T in the third codon (Figure 7.19) position shows a linear decline with an anomalous rise in the penultimate codon. The A3 content rises sharply in the last two codons. These findings are consistent with those in other organisms that have reported evidence of selection over the last 20 codons. In the third position C3 and G3 content remain constant up to the last two codons when A3 is preferred. The T3 shows a linear decline for the last 20 codons with a curious rise in the penultimate translated codon (Figure 7.20). The A3 content rises in the last two codons from an average value of 43% to 47.1%.

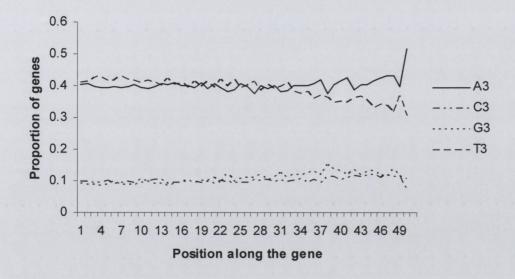


FIG. 7.19. Base use in the third codon position for the last 50 codons 5' of the termination codon.

The presence of a bulky (W, F or Y) and a positively charged residue (R, H or K) within the last 10 residues was also examined. 207 (4.2%) of the genes do not obey this rule. Relaxing this rule slightly and examining instead the last 15 amino acids the number of genes that do not obey this rule falls to 58 (1.2%): extending this rule to the last 20, the number failing falls to 26 (0.5%). Genes failing all versions of this rule were found on all the nuclear chromosomes and even on the plastid. No common factors were found between them.

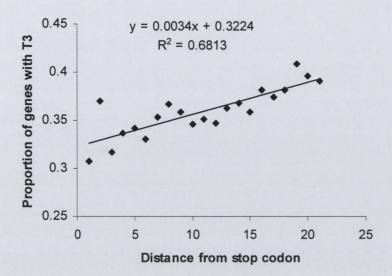


FIG. 7.20. T3 content and distance from the termination codon. The anomalous T3 content of the penultimate codon is visible. (F = 40.61 p < 0.0001)

7.4 Discussion

The G1 > G2 > G3 and T1 < T2 < T3 patterns found here are consistent with the earlier report as is the frequency of adenosine in the second codon position. Here the A1 content exceeds the G1 content. This is a pattern has been reported to be dependent on the overall GC content of the genome (288) and the findings here are consistent with this. It is presently an open question whether the patterns found here are found in all organisms.

While other organisms tend to have mean purine content in the second position of \sim 50% and below 50% in the third giving a mean codon pattern of RNY where R is a purine, Y a pyrimidine and N may be either, the pattern found here is RRN. Whether this is characteristic of the genus *Plasmodium* or simply of *P. falciparum* alone is presently unknown. Of these the *Plasmodium vivax* seems likely to be these first to be completed probably in 2006/7 at which point it may be possible to answer this question.

The distribution of the base content in each codon position is of some interest. All of these are clearly non normal. Because the values lies within four but not three standard deviations of the mean, the Vysochanskii-Petunin inequality indicates that these are not drawn from a unimodal distribution. This is consistent with the observed non normality of the distribution. One possible origin for the non unimodal distribution is its use of two hosts. Some genes are only expressed in the mosquito and some only in the vertebrate host: it is possible that within each set the third codon base use is distributed normally. The currently available data does not allow us to test this hypothesis.

The findings here are incompatible with the hypothesis that the bases in the any codon position are distributed randomly. While this might seem intuitively obvious for the first two codon positions, an important assumption of the neutral theory of evolution is that - to a good approximation - the bases in the third codon position can be regarded as random substitutions subject to a background mutation pressure. Given the number of codons examined here, under the law of large numbers, if this assumption was correct a normal distribution in the third codon positions would be expected.

This possibility that the third base composition should not be regarded as random is supported by the earlier findings of a correlation between the GC1 and GC2 content and the GC3, the correlation of the GC3 content with the hydrophobicity of the protein and the findings here. The hydrophobicity of a protein is determined largely by the average of the base content in the second codon position being negatively correlated with the adenosine content and positively with the thymidine content. The correlation here of the GC1 and GC2 content with the GC3 content were both positive and statistically significant. The correlation between the GC1 and GC2 content and the GC3 content is a consequence of the regression found here between the guanine and cytosine content in the first and second positions and that of the third. The regression equation suggests that 94.8% of the variation in the third codon GC content is potentially 'explainable' as a weighted average of GC content of the preceding two positions. If the GC3 content is in fact a weighted average of the GC content of the first two codon positions with some variation around this value permitted then it is likely that the base content of this position is in fact severely restrained. Selection pressure on the first two codon positions is considered to be much greater than that on the third. If this relationship is true then much of the selective pressure on the first and second codons would also apply to the third position.

These finding agrees with the dinucleotide analysis and suggests that freedom to mutate in the in the third codon position are somehow restrained by the GC content of the first two positions. The negative sign of the C2 coefficient suggests an avoidance

of C or G in the third position. The Karlin signatures make it clear that the sign relates to the G3 content rather than the C3. So while there is a tendency for CG rich codons to end in C or G, G is avoided in the third position when C is at the second.

Nonetheless it is possible that the situation here is unusual. The AT content of the genome is may be sufficiently high to impose unusual restraints on the permitted variation of the third base. These relationships need to be examined is a wide range of genomes before any firm conclusions can be drawn but the finding here do support the suggestion that third base content at least in this genome may not be as the neutral theory assumes.

While proteins encoded in the genome are not all translated equally, examination of the encoded frequencies sheds some light on the biology of the organism. Encoded amino acid content - while seeming at first glance random - obeys a quantitative rule. Amino acids are encoded with exponentially increasing frequency. This finding begs two questions: what is the origin of this relationship and why is tryptophan alone as an outlier?

There are no completely satisfactory answers. Starting with the second tryptophan is a large bulkly hydrophobic amino acid encoded by TGG. Phenylalanine is biochemically similar and is encoded by the codons TTT/C. In this genome TTT is the more commonly used codon. The TGG codon may have been replaced by a TTT codon wherever possible because of low GC content of the genome. If this hypothesis is correct when compared with orthologous genes in other *Plasmodium* species with higher GC content it would be expected that at least some of the TTT codons as in *P. falciparum* would be replaced by TGG in the orthologous genes. It may be sonn possible to test this hypothesis against the *P. vivax* genome whose GC content is \sim 60%.

To answer the first question examination of the encoded codon use may be helpful. This reveals a similar pattern. Ranking the codons in ascending order in frequency of use and plotting this against the logarithm of the frequency of use gives to a very good approximation a line. Clearly the 'reason' for the distribution of the encoded amino acids lies with the encoded codons. The two least commonly encoded amino acids – tryptophan (0.5%) and cysteine (1.8%) are also uncommon in other genomes. The next two least frequent amino acids - alanine (1.9%) and proline (2.0%) - have relatively GC rich codons - GCN and CCN respectively. It may be a combination of rarely used codons and the high AT content of the genome that has caused tryptophan

to become an outlier from the line. Unfortunately this does not provide much insight as to why this distribution of encoded codons exists. These plots are currently mysterious and need to be examined in other genomes as it is possible that these are features of this genome alone.

Turning now to the correspondence analysis there is considerably more comparative data in the literature. Starting with studies on correspondence analysis of proteins Palacios and Wernegreen examined amino acid and codon use in 479 proteins of obligate endosymbiont Buchnera strain APS (360). Buchnera is found in the specialized cells (bacteriocytes) of the body cavity of aphids and 583 proteins have been identified in its genome (444). The authors found that the first four of the 19 axes accounted for >50% of the variation in amino acid use. The first axis correlated with the GC content of the first two codon positions, the second with the hydrophobicity and the fourth with the cysteine content and the third could not be identified. The authors also examined the E. coli K12 genome and found that 49.2% of the variation there could be explained also by four axes. The first correlated positively with the hydrophobicity and negatively with the AT skew, axes 2 and 3 with the level of expression and axis 4 with the cysteine content. Correlation between the AT skew and hydrophobicity (21) is to be expected as was found here in the A2-T2 plot against the hydrophobicity. Similar examination of the bacterium *Thermotoga* maritime revealed four trends accounting for 48.4% of the variation. The first correlated with the hydrophobicity of the proteins, the second with the mean molecular weight, the third with the aromatic content and the fourth with the cysteine content (558). A study of 14 bacterial genome from the family *Bacillaceae* revealed three axes accounting for 90.3% of the variation (335). The first correlated with the GC content of the first two codon positions, the second with the optimal growth temperature and the third with the hydrophobity. In the protozoon Giardia lamblia three trends were found in examination of 75 proteins: the first axes correlated with the cysteine content and the mean molecular weight; the second with hydrophobiity and aromaticity and the third with the expression level (155). In Blochmannia floridanus, an endosymbiotic bacterium whose host is the ant Camponotus floridanus, found a single trend in the proteins that correlated both with GC rich amino acids and high expression. (21)

The findings here agree with the patterns found earlier. The first axis here correlated with the GC content of the first two codon positions while the second correlated with

the hydrophobicity of the protein. The third axis correlated with the molecular weight of the protein.

The studies listed here constitute the entire current literature on the application of correspondance analysis to proteins. A few common themes can be identified. Hydrophobicity and the GC content of the first two codon positions are clearly important. While expression data is lacking for many of the genomes this also may be important. Additionally molecular weight of the protein may have some significance. Precisely how these findings fit into biological theory is not presently clear but as recurrent themes they likely represent a set of important biological parameters that are not yet completely understood.

In contrast to the paucity of its application to proteins, correspondence analysis has been used to examine codon use in number of organisms including Lactococcus lactis (180), Campylobacter jejuni (175), Clostridium perfringens and Clostridium acetobutylicum (330), Helicobacter pylori (256), Borrelia burgdorferi and Treponema pallidum (255), Candida albicans and Saccharomyces cerevisiae (278), Bacillus subtilis (443), Streptococcus pneumoniae (294), Escherichia coli (167), Mycobacterium tuberculosis and Mycobacterium leprae (106), Mycobacterium smegmatis and its bacteriophage Bxz1 (418), Sinorhizobium meliloti (364), Rickettsia prowazekii (16), Caenorhabditis elegans (474), Entamoeba histolytica (404), Echinococcus spp (132), Dictyostelium discoideum (436), Schistosoma mansoni (331), three species of fish from the family Cyprinidae (405) and HIV (307). In general one to three axes have been found among the potential 58 and that at least one correlates with highly expressed proteins. This axis typically also correlates with G or C ending codons and it has been suggested that theses are 'optimal' codons with respect to translation. The other axes - if any - may or may not be found to be correlated with other properties.

Application of this methodology to the *P. falciparum* genome failed to identify any large scale trends in codon use. This is unlike most organisms that have been examined to date where one to three axes identified over 50% of the intergenic codon variation. While the reasons for this difference are not presently clear several possibilities spring to mind including the parasitic life style, the reduced genome and the high AT content being among them. Constructing a testable hypothesis to evaluate these possibilities is difficult.

The first axis here correlates with the A3 and T3 content and the second with the C3 and G3 content. If this organism behaves in similar fashion to the others examined the second axis should correlate with the expression levels. It is also possible that this correlation may not be related to translation. In a study of 80 species of bacteria Sharp *et al* found that translation pressures were present only in species with rapid growth (439). Additionally they found a strong correlation between the number of rRNA operons and the number of tRNAs. The *P. falciparum* data (5 rRNA and 40 tRNAs) fits this curve. Testing this hypothesis will need additional proteomic data.

Correspondence analysis of both the proteins and codons here revealed that while the trends in amino acid use found here seem to be similar to those in other organisms codon use may not be so. Whether or not this is a consequence of the relatively frequencies in codon use described earlier needs to be tested by examining codon use in the organisms where the trends in codon use are better understood.

The effective number of codons has been used to investigate translation efficiency before. Wright working with a set of *E. coli* gene proposed that genes with a low N_c value have very biased codon use and are expressed at high level: those with a high N_c value have low codon bias and low expression. Subsequent studies with this method and more sophisticated methodologies have tended to concur with this hypothesis (14, 15, 177) although Lafay *et al* have cast doubt on this interpretation (256).

The findings here tend to disagree with this hypothesis. The genes tend cluster at the low end of the GC3s axis and lie close to the theoretical line suggesting a lack of translational selection. This is consistent with the small eigenvalue found in the correspondence analysis for the second axis. While very little is presently known about the actual quantities of the proteins within the cell, genes that would be expected to be highly expressed (ribosomal proteins and translation elongation factors) were scattered throughout the N_e-GC3s plot again supporting the hypothesis of a general lack of translational selection. This hypothesis should be revisited when additional data on the protein quantities is available.

Considering now the variation of the codon chi square with the length of the gene this was investigated by Sharpe *et al* (440) in the case of the codon chi square and the effective number of genes by McInerney (305). It is clear from the plots in both papers that both regressions show significant heteroscedacity. The most probable explanation for this problem is that at least one variable of importance is missing from

the regression model. This seems biologically plausible: both the N_e and the codon chi square are measures of codon variability and it is unlikely that the sole source of codon variation is the length of the gene. Models that are underfitted - models that lack important dependent variables - produce biased and inconsistent results. As a consequence the regression coefficients and hypothesis testing procedures are likely to give misleading results (176). This appears to be the case here: Sharpe *et al* commented that the findings ran contrary to their expectations that codon variation would increase with gene length but were unable to explain this.

These values for *P. falciparum* were plotted against the gene length and the regression lines were similar to those found earlier in other organisms. The increase in the R^2 on the addition of a second variable – the log of the gene length – is what would be expected on the addition of a relevant variable to an under specified model. The change in the slope of the coefficient of the gene length in the regression brings the model more in line with biological expectations. In spite of the improvements in the models' fit the R^2 value remain low (0.332 and 0.051 for the chi square and the N_c respectively). It seems highly likely that there are additional parameters that have been omitted from the regression which are likely to influence the regression coefficients so caution is indicated in making inferences from these regressions.

Returning to the topic of the effect of codon choice on translation, it is appears that base use near the translation initiation site may be non random. In E. coli and Salmonella typhimurium synonymous substitution is reduced over the first 25-30 codons and gradients of both A and G exist for all the three codon positions: the percentage of A tends to decrease and that of G to increase (127). The authors proposed that this might reflect selection for 'optimal' codons at the start of the gene and avoidance of secondary RNA structures: this finding was later confirmed (205). A third paper found that the gradients in E. coli were more marked in genes with high expression levels and that these gradients were did not exist in Bacillus subtilis (151). A study of 22 208 human mRNA showed an over representation of G at the +4 site at 47% of the total (462). The authors also studied the R-3/G+4 (a purine at the -3 position and a G at +4) rule these genes and found it was adhered to in only 37.4%. Furthermore 12.5% of the genes lacked both the R-3 and the G+4. Among the histones the R-3/G+4 was adhered to only by H3: the H1, H2A, H2B and H4 genes lacked either R-3 or G+4. Alterations in transgenic mice of these sites lead to observable phenotypic changes.

The findings here suggest that *P. falciparum* also uses similar mechanism to regulate translation. Base use at the -3 site is heavily biased towards purines (89.9%). The preference for purines falls as the start ATG is approached. Thymidine constitutes \sim 40-45% of the non coding regions but thymidine residues are much less common at the -3 position than would be expected given this background frequency suggesting negative selection at this position. This negative selection pressure appears to act to a lesser extent on the -2 and -1 sites. This would be consistent with earlier work.

All the 326 genes in this annotation with a T at the -3 site have all been carefully reviewed. While some may be sequence errors, the majority appear to be genuine. The patterns found here are similar to those of Saul and Battistutta in their earlier examination of 22 sequences (424). In the initial publication of chromosome 3 in 1999 the highly expressed circumsporozoite surface protein (Pf3_1:217997c) was annotated as TTT ATG – rest of sequence. Saul and Battistutta had found this annotation in their earlier set of sequences and had suggested that this might be in error and that the correct ATG start codon instead lay several bases 5' with a superior initiation context. The circumzoite surface protein has a signal sequence which has made this difficult to resolve this problem experimentally. This protein has been reannotated here to have an adenosine residue at the -3 site as it is believed that this is the correct translation initiation site.

The +4 position (the base immediately after the ATG) has also been shown to be of some importance in translation initiation. Here a purine is at this position found in >60% of cases. The preferred base at the +4 site is adenosine with guanine being the second choice. In other organisms guanine is normally the preferred choice at the +4 position. The difference here presumably has arisen as a result of the high AT content of the genome.

While reviewing the annotation genes without purines either at the -3 or +4 sites were difficult to resolve. Preference has been given to ATGs with purines in the -3 position and TTT ATG sequences avoided wherever possible. These may not be correct: the apparently weaker site may also be used. Translation initiation is a complex process (514) and it is simplistic to assume that all the relevant information is contained within these sequences. Examination of 26,225 mRNA sequences revealed that suboptimal ATG codons are common near the start of genes (240) and that ACG and CUG can act as alternative start codons (432, 494). These cases are presently regarded as exceptional and the results here suggest that *P. falciparum* is similar to other

eukaryotes recognizing the initial ATG on the principally on the basis of the -321 sites with some contribution from the +4 base.

The trends in base use by codon position found here may not reflect the "average" messenger RNA used by the ribosome as all the genes have been given an equal weighing which is unlikely to be realistic. In the absence of experimental data to resolve this issue the patterns found here should be viewed with caution.

In the first codon position the tendency is for the mean purine content to fall over the first twenty or so codons after an initial peak. Over this same distance the thymidine content rises to a steady value. The cytosine content remains stable throughout. The initial peak in the mean purine content may be connected with translation initiation as discussed earlier. While changes in this position are almost always non synonymous there is no obvious association with a known biological property associated with these trends making it difficult to relate these findings to the known biology.

In the second codon position the thymidine content rises and the adenosine content falls over the 10-15 codons after which these trends reverse. The cytosine and guanine content remain stable throughout. Given that the hydrophobicity of a sequence is closely related to the A2-T2 value these trends are likely to reflect the presence of signal sequences at the N terminal end of the proteins.

The pattern found here in the third codon position is one of stability. Despite a small drop in the mean A3 content about the tenth codon the third codon content is stable with adenosine and thymidine making up 40-45% of the bases and cytosine and guanine 10-15%. This is in contrast to the earlier findings in *E. coli* where changes are obvious. The trends found in *E. coli* may be unique to that organism. Alternatively it may be that such trends are found only in highly or lowly expressed proteins. This data is not currently available for *P. falciparum* and this question may be worth revisiting once this data is available.

Translation termination like translation initiation is a complex process and was recently reviewed by Bertram *et al* (36). In eukaryotes translation is terminated by a heterodimer consisting of two proteins, release factors eRF1 and eRF3. The interaction of yeast eRF1 and eRF3 is mediated by the last 6 to 11 amino acids of eRF1. A stop codon located in the ribosomal A-site is recognized by the release factor complex, which binds the ribosome and triggers the release of the nascent peptide. The protein eRF1 recognizes all three canonical stop codons, triggers peptidyl-tRNA hydrolysis by the ribosome and catalyses the release of the peptide. The efficiency of

this process is greatly enhanced by the GTPase release factor eRF3 which in yeast eRF3 is not an essential gene. The structure of human eRF1 has been solved and resembles a tRNA molecule (468). The *P. falciparum* eRF1 is located on chromosome 2 (Pf2_1:492616w) and two other potential release factors are found on chromosomes 7 (Pf7_6:216652c) and 9 (Pf9_1:1292107c).

There are three canonical stop codons: TAA, TAG and TGA: many exceptions to this rule are known. In the *Mycoplasma* species *genitalium*, *capricolum*, *pneumoniae* and *gallisepticum* the TGA codon encodes tryptophan. The ciliates *Paramecium* and *Tetrahymena* use TAA and TAG to encode glutamine and *Euplotes* encodes cysteine with TGA. The stop codons may also code for rare amino acids: selenocysteine is encoded by TGA (269) - an amino acid which may be found in *P. falciparum* (326) - and pyrrolysine which is encoded by TAG (250) - which to date is not known to occur in *P. falciparum*. In addition to encoding amino acids read through of these 'stop' codons may be physiologically important. CFA/II strains of enterotoxigenic *E coli* express three types of surface-associated hair-like fimbriae - CS1, CS2 and CS3 – and expression of the genes required for biosynthesis and assembly of CS3 pilli requires the suppression of a UAG codon (222).

There is no experimental data available on translation termination in *P. falciparum* and the material presented here is the first summary of the context and use of termination codons in any *Plasmodium* species. The high AT content of *P. falciparum* makes 'ambushes' - which may be important in other organisms (383) - almost redundant because in most proteins there is only one possible reading frame for most of their length.

The frequency of use of the stop codons - 70% TAA, 20.4% TGA and 9.6% TAG – is that typically found in other organisms. The most common base immediately following the stop codon is adenosine with the combination of TAA followed by A is found in almost 40% of the genes and seems to be the optimal stop signal. This is consistent with findings elsewhere. Intuitively one might perhaps have expected the more G rich stop codons (TAG or TGA) to be associated with the genes richer in GC or the longer and metabolically more costly genes to be terminated with an optimal stop codon (TAA). Neither of these hypotheses was confirmed here and the reason for choice of stop codon remains unclear. It may be that stop codon use is conserved between *Plasmodium* species and that despite the presumed rise in AT content that has occurred in this genome the stop codons have been preserved. When genes from

the other *Plasmodium* species become available it will be possible to test this hypothesis.

There is a gradual decline in the mean hydrophobicity of the protein over the last 20 codons. Both a bulky residue (ϕ) and a positively charged residue (+) are normally (~95% proteins) found within the last 10 residues. The importance of this ϕ_+ rule is not known but is a common if not universal finding and as such is likely to be important in the termination process.

The mean base use analysis approaching the termination codon is subject to the same caveats mentioned earlier in relation to the initiation codons. In the first codon position adenosine is the most common base (40 - 45%), guanine and thymidine are about equal in frequency (20 - 25%) and cytosine is the least common (10 - 15%). These frequencies appear to be fairly stable over the last 50 codons with the possible exception in a rise in the use of adenosine towards the end of the gene.

In the second codon position base use is fairly constant (adenosine ~40%, thymidine ~30%, cytosine ~15% and guanine ~10%). There is a small rise and correspondingly small decrease in the adenosine and thymdine content respectively in the last 10-20 codons paralleling the increase in mean hydrophobicity seen in this region. As noted earlier on average the C2 content exceeds that of the G2 content. At the start of the genes these two values are approximately equal but towards the end of the gene cytosine has clearly become more common. This is the first time such a pattern has been described and the reasons for its occurrence are not known. Unlike the case of the A2-T2 difference there are no obvious biological correlates.

In the third codon position while the cytosine and guanine contents remain stable $(\sim 10\%)$ the adenosine and thymidine contents diverge over the last ~ 15 codons. The A3 content rises in a linear fashion over this distance and the T3 falls in a similar fashion. The T3 content in the penultimate codon appears to be anomalous high judging by this linear regression. The pattern here is quite different to the mean third base content near the start of the genes which was largely constant. If it is assumed that the presence of a trend in third base use is evidence of selection which is plausible on biological grounds then it appears that selection may be acting on these codons.

This hypothesis is consistent with the change in mean hydrophobicity and the almost constant presence of a large hydrophobic and positively charged residue in this region. This hypothesis is also plausible on biological grounds as this part of the protein might interact with the translation mechanism acting as an 'early warning signal' that a stop codon is being approached. The hypothesis seems testable as it predicts that mutations in the last 10-15 codons should influence the termination rate. Any changes in translation could be monitored by the use of a protein such as florescent green protein with the terminal segment of a *P. falciparum* protein attached at the C terminal end.

Many of the patterns found here cannot be presently be explained by biological theory. This inability to integrate this data into existing theory merely reflects our current lack of understanding. The presence or absence of these patterns may be worthwhile investigating in the human genome also as if these are specific to P. *falciparum* they may represent a biochemical difference that is biochemically exploitable.

Chapter 8: Intron organization and structure

8.1 Summary

In this chapter a number of characteristics of the introns of *P. falciparum* are described.

Intron containing genes are found in all chromosomes and they tend to be longer than the genes with introns. Neither exon nor intron lengths follow an easily identifiable pattern. Nor is there is a correlation between gene length and intron number. Introns are more AT rich than the surrounding exons. Base use within the intron appears to be subject to a set of selective forces influencing the location of bases within the intron. It has been proposed that the splicing mechanism may be the origin of at least some of these patterns. The GC3 content of the bounding exons are not correlated suggesting that different selective forces may act on the 5' and 3' exons.

Both the phase of the exons and the introns and the surrounding bases themselves are non randomly distributed. This may be due to the splicing apparatus rather than historical accident. A sequence $(AT)_{3-4}$ lying close to the 3' end of the introns that may act as the intron lariat site has been identified.

8.2 Overview

Previous to this there have been no large scale surveys of intron composition and organization in this organism. The current version of this annotation has identified 8227 introns in 2166 (42.5% total) genes – an average of 3.8 introns per gene. The number of introns per chromosome are listed in Table 8.1.

The number of introns per kilobase of chromosome seems to be approximately constant (Figure 8.1). While it would seem probable on biological grounds that some relationship might exist between the length of the chromosome and the number of introns therein, the strength of the linear relationship found here was unexpected. Given the variability in the number of introns per gene, it is possible that this relationship may contribute to the location of the genes encoded in the genome. If this intron-chromosome length is a biological law then in the long run only genes whose number of introns fit this law may be incorporated to a given chromosome unless the

genes are extensively modified. This hypothesis has a parallel in Chargaff's second rule: coding sequences tend to have an excess of A and G and introns tend to have an excess of C and T. This to some extent limits the type of gene that can be accommodated in a chromosome - particularly in the shorter ones. A similar rule may hold here concerning the number of introns. If this is so it is likely to relate to some structural property of the genome. This finding needs testing in other genomes as it may simply be an artifact unique to *P. falciparum*.

Chr 1	218
Chr 2	337
Chr 3	374
Chr 4	369
Chr 5	454
Chr 6	516
Chr 7	384
Chr 8	472
Chr 9	602
Chr 10	636
Chr 11	818
Chr 12	782
Chr 13	1001
Chr 14	1269

TABLE 8.1. Number of introns per chromosome.

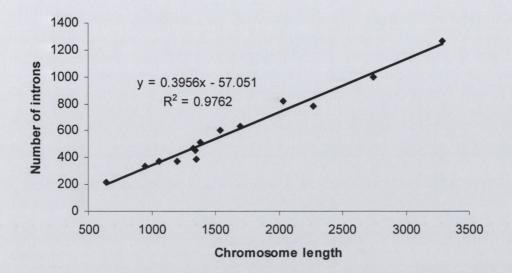


FIG. 8.1. Number of introns per kilobase of chromosome.

8.3 Methods

The exon and intron lengths, intron number, intron base composition, exon GC3 content, intron and exon phase, and the intron sequences themselves were obtained from the database. The distribution of the intron length modulo 3 was examined. The lengths of the first introns were compared with the length of the other introns with the t test with the assumption of unequal variances.

Base composition around the splice sites was also studied. The GT and AG sites were removed and the mean composition of the introns were determined in the 20 bases preceding the 5' splice sites and following the 3' AG and the 18 bases within lying adjacent to the GT and AG sites.

The existence of a correlation between either the GC or GC3 content of the 5' exons, the 3' exons and the GC content of the introns was tested by linear regression. Clustering of bases within the intron was studied with two statistical tests that have not been previously applied to the analysis of introns. The first was a test for distribution of points along a line (236).

Let there be n points along a line at positions x_i with $0 \le x_i \le 1$ and let $d_i = x_i$ for $i \le n$. Put $d_n = 1 - x_n$ and $Y = \Sigma$ i d_i and Y' = 1 - Y / n. For $n \ge 25$, Y'is approximately normally distributed with mean (n - 1)/2n and variance $(n - 1) / 12 n^2$. The condition on n is satisfied by the A and T bases in all the introns. This test seems presently to lack a name: it will be referred to here as the 'Y test'. Within each intron the mean position by percentage length of the A and T bases in each intron was also determined and then the grand mean of these values was estimated.

The Y test was also applied to the location of introns within the gene. The last base coordinate of the 5' exon was used as the location of the intron within the gene. To determine the expectation of the Y' statistic and its standard deviation for the number of introns (n) between 1 and 25, a boot strap method was used. For each n, 10,000 randomly generated 'genes' with the correct number of 'introns' were examined and the mean and standard deviation computed from these values.

The second method used was the Erdos Renyi test. In 1950 Rényi and Erdös determined the number of runs of digits in a random sequence and this was later was used as the basis of the probability estimates used in the BLAST algorithm (11). The proof of the Erdos-Renyi theorem while trivial is not commonly found in biology text books so will be given here in full.

Consider a sequence of characters of length l in which all the members of the set of characters have an equal chance (p) of appearing. We seek an estimate of the maximum expected length of run of identical digits. The probability of a run of length n of a given digit is

$$P(n) = l * p^{r}$$

The probability of there being a maximum length run is 1 as there will always be a maximum length of the run of a character even if it is of zero length within the sequence. So we have

$$1 = 1 * p^{n'}$$

where n' is the maximum run length. Solving this for n' we have

$$n' = -\log(1) / \log(p)$$

For a set of digits 0 to 9 inclusive and using the base 10 for the logarithms we have

n' = log(l)

In words, the maximum expected length of a run of random digits in a sequence of length l equals log (l).

The maximum length of the A and T runs within the introns was determined and assuming a random distribution, the expected lengths were determined by the Erdos-Renyi theorem. The observed and expected values were then compared with the two sample t test with the assumption of unequal variances.

With the assumption that dinucleotide pairs are random distributed throughout the intron the expected number of AG dinucleotides and their location with in the intron was determined. If the location of these dinucleotides within the intron is random - here a uniform distribution - then on average when there are n expected dinucleotides then the one closest to the end of the intron will be located at 1/(n + 1) times the length of the intron from the 3' end. The mean and variance of these expected distances was computed and these figures were then compared with the observed mean and variance with the t test again assuming unequal variances.

While the preceding tests are elementary and provide some useful information more sophisticated approaches may yield more biologically useful information. One such method is the Gibbs sampler - a Bayesian technique – to seek the lariat site within the introns. To date no lariat site has been identified in *P. falciparum* (199). Since a lariat on biological grounds seems very likely to have a common pattern shared by all the introns, this seemed suitable for identification by a Gibbs sampler. The Gibbs sampler is an example of the popular Metropolis Monte Carlo algorithm and has found application in areas other than biology: Gelfand *et a*l have used it in the analysis of real estate data (162).

The basic Gibbs sampler algorithm (264) assumes we are given a set of N sequences with a common pattern of width W. The sampler maintains two data structures. The first is a set of positions a_k ($1 \le k \le N$) the position of the pattern within each of the N sequences. The second is a set of variables $q_{i,4}$ ($1 \le i \le W$)that represent the frequencies of the bases in the positions within the pattern. In addition a second set of background frequencies $p_1, ..., p_4$ with which the bases occur outside the pattern. The algorithm identifies the pattern by maximizing its likelihood compared with the background.

The algorithm chooses a random starting position within the N sequences and then iterates through the following two steps.

- 1. Predictive update step: A sequence z is chosen either at random or in a specified order. The pattern description $q_{i,j}$ and background frequencies p_j are calculated from the current positions a_k excluding the chosen one.
- 2. Sampling step: Every possible sequence of width W within z is considered as a possible instance of the pattern. A probability Q_x of generating sequence x according to the current pattern probabilities are calculated as are the probabilities P_x of generating these segments with the background probabilities p_x . A weight A_x is assigned to each segment x. Once each segment is weighted a one is chosen with probability $A_x/\Sigma A_x$ and this new segment becomes the new a_k .

The idea behind this algorithm is that the more accurate the pattern description in step 1 the more likely it is to be chosen in step 2. Once the correct pattern is found convergence is rapid.

There are two problems with the basic algorithm. The first is the possibility that the pattern selected is suboptimal. This is overcome by comparing the pattern after a number of runs with a segment a fixed number of bases to its left and right, computing the probability of this being the correct pattern and then using these as weights selecting one of these three possibilities at random. The second problem is that the window length initially chosen may not be optimal. This is overcome using the incomplete data log probability ratio G which formula is given below. The rationalle behind the choice of this function is given in the original reference.

To determine the $q_{i,j}$ from the current set of position a_k we consider the ith position of the current pattern in the current sequence z. There are N – 1 bases in this position when we exclude the current sequence. Let $c_{i,j}$ be the base count in this position. Bayesian analysis suggest that that the raw counts should be supplemented with pseudocounts to yield the pattern probabilities and the authors proposed the following formula

 $q_{i,j} = (c_{i,j} + b_j) / (N - 1 + B)$

where B is the sum of the b_j . The p_j are determined with the same formula with the counts being taken over the bases outside the pattern. The choice of the b_j is fairly arbitrary but the authors suggested putting b_j equal to the frequency of the j^{th} base in the whole data set multiplied by the square root of N.

Following normalization the A_x give the probability that the pattern in sequence z belongs at position x. The algorithm finds the pattern by selecting a set of a_k that maximizes the product of these ratios. Maximization of the logarithms of these ratios (F) is functionally equivalent to this and less prone to rounding errors and is given by

 $F = \Sigma \Sigma c_{i,i} \log (q_{i,i}/p_i)$

where the first sum is taken over the pattern length and the second over the number of bases (4).

The formula for G which the authors found empirically to be their best choice to optimize the window size is

 $G = F - \Sigma (\log L_i + \Sigma Y_{i,j} \log Y_{i,j})$

L is the number of possible position for the pattern in the ith sequence, and Yi,j is the normalized weight of position j and is equal to Q_j/P_j divided by the sum of these weights within the ith sequence. The first sum is taken over the N sequences and the second over the length L_i

Larence *et al* found that their algorithm runs in O(N) time and identified variable length motifs reliably in proteins. Thijs *et al* ⁽⁵⁰²⁾ found that this algorithm was unreliable for analysis of DNA and discovered that it could be improved by incorporating a Markov model to determine the background frequencies. Empirically they found a third or fourth order model reliably enabled them to identify regulatory regions in plants.

While the original algorithm – a zero order Markov model - worked well for proteins, preliminary investigations here concurred with the findings of Thijs *et al* that a third order Markov model was more suitable to identify motifs in natural DNA sequences. Lariat sites are generally 6-8 bases long and have an invariant A in the penultimate 3' position which is covalently bound to the 5' GT during the splicing process. Since

every intron has a lariat site, this motif within the intron was sought using the full set of introns as well as within smaller subsets. In addition initial motif lengths were varied to ensure that local optima were not biasing the results.

In 1986 Schnieder *et al* (431) presented a new idea in biology – a method of measuring biological information directly from sets of sequences apparently based on the Shannon's information theory. In Schnieder's method, sequences were aligned and counts made over the four types bases by position and this frequency data was then converted to "information" using a formula similar to that used by Shannon in his analysis of information over a communication channel. Given a set of sequences of identical length count the frequency of the base in all positions, Schiender's statistic (S) is

 $S(x) = 2 + \Sigma p_i(x) \log_2 p_i(x)$

where $p_i(x)$ is the frequency of the ith base at position x within the sequence and the sum is taken over all four bases. Schnieder subsequently used this idea in several interesting and original ways (429, 430, 477). Unfortunately there appears to be a misunderstanding in the original paper that has been replicated in both his and others subsequent work.

In 1971 Tribus and McIrvine applied information theory to biology (508) and used the example of the reading of a ticker tape. They noted that if you can reliably expect to see (say) a mark at a particular point then on the arrival of this mark there is no gain of information consequent on receiving that part of the message and that if you start with a set of ticker tapes and then find patterns in them your knowledge increases – an implied Bayesian system. Once you recognize that some of the regular marks are (say) time signals recognizing any more of these afterwards gives you no new information. Schnieder *et al* seem to have confused the entropy of the message with the entropy of the mutilation of the message.

An example may make this clearer. Consider a typical start codon ATG in *E. coli*. While the three bases before this ATG site are biased with the -3 site is normally a purine, the remainder of the untranslated region will normally have all the bases in roughly equal quantities. To obtain a measure of the similar to Shannon, Schiender took the background frequencies in *E. coli* to be the equivalent to Shannon's random

noise and assigned them an information value of zero. This, in Schnieder's theory, is the value of the 'background noise' from which important biological sequences must be identified by the organism.

Because the ATG is fixed as a start site in Shannon's theory this is redundant and has an information content of zero. In Schnieder's theory this fixed ATG is the 'information.' Schnieder computed the difference between the background values and these redundant values and called the difference - confusingly - 'information.' Because the ATG is fixed this codon is found with probability 1 giving $\log_2 (1) = 0$. Schnieder assigned a value of 2 (= 2 - 0) to all the positions of the redundant ATG codon and called this the 'information' at this position.

This choice of terminology of this statistic by Schnieder *et al* was unfortunate as it appeared to imply that this new statistic was equivalent to those of Shannon's theory. This new statistic has been used in several papers since it was first described but this confusion in its meaning seems to have persisted. This statistic is a useful summary statistic and can be used to determine in a quantitative fashion biased base use around functionally important sites. Schnieder provided an estimate of the variance of the expected value of his statistic but for most purposes graphical representation is sufficient to identify the biased base use (430).

This statistic was applied to the exon-intron junctions to search for unusual base use. Exons with less than 10 codons (353 in total) were arbitrarily excluded. The majority of these (231) are first exons which are likely to have biased base use for additional reasons and it was felt that this might bias the results. The value of S was determined for the ten bases before the 5' and 3' sites.

8.4 Results

Exon length is very variable ranging form a single base to 32 691 bases. The shortest intron is 58 bases and the longest 1194. While the intron lengths are equally distributed with respect to mod 3, neither the intron nor exon lengths are geometrically distributed (Figures 8.2 and 8.3) as the authors of the Phat program had assumed.

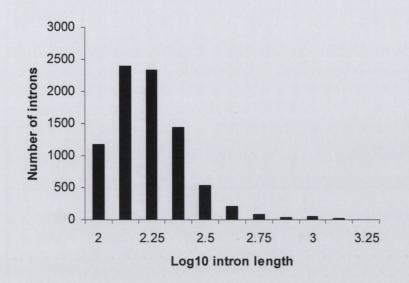


FIG. 8.2 Distribution of log₁₀ (intron lengths).

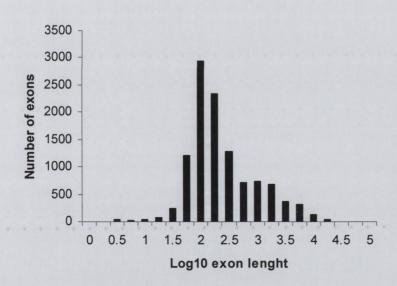


FIG. 8.3 Distribution of log₁₀ (exon lengths).

In contrast to the relationship found between the length of the chromosome and the number of introns therein, no functional relationship (Figure 8.4) was found between the number of introns and the length of the gene (p = 0.18)

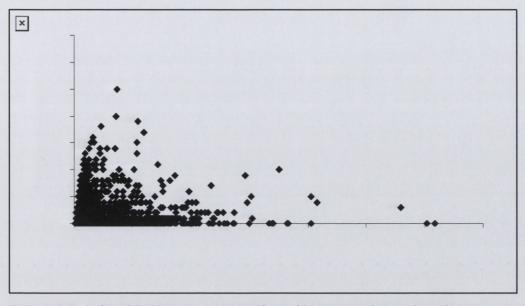


FIG. 8.4 Relationship between the number of introns and gene length.

Genes without introns are on the whole longer – mean length 801 residues - than those with introns – mean length 737 residues (t = 2.62, p = 0.009) and have a lower GC3 content (19.4% and 16.9% respectively: t = 18.0, p < 10⁻⁶). First introns are significantly longer than other introns. First intron lengths are 200.2 +/-138.2 bases while the other intron lengths are 146.9 +/- 62.0 bases. (t = 19.7, p < 10⁻¹⁰).

The introns are AT rich (Figure 8.5) with an AT content of 87.5 +/- 3.5% (mean +/standard deviation). In comparison the AT content of exons is 74.8 +/- 5.2%. Breaking down the intron content by base: A is 41.1 +/- 7.3%, T 46.1 +/- 7.2%, G 7.0 +/- 2.5% and C 5.8 +/-2.7%. All base content distributions are non normal. While not significantly skewed all the base content distributions have platykurtic tails (kurtosis < 0.4). As might be expected from Chargaff's second rule and Sybalski' rule the intron pyrimidine content (51.9 +/- 7.1%) exceeds the purine content (48.1 +/- 7.1%).

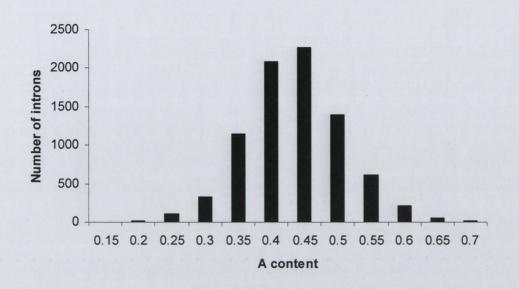
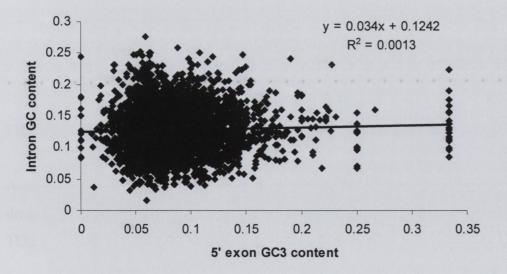
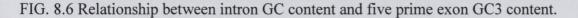


FIG. 8.5 Distribution of intron adenosine content (percentage)

Regression (Figures 8.6 and 8.7) of the GC3 content of the five prime and three prime exons against the GC content of the introns is markedly heteroscedastic: consequently it is difficult to draw meaningful inferences from this regression. Regression of the GC3 content of the five and three prime exons is also heteroscedastic (Figure 8.8).





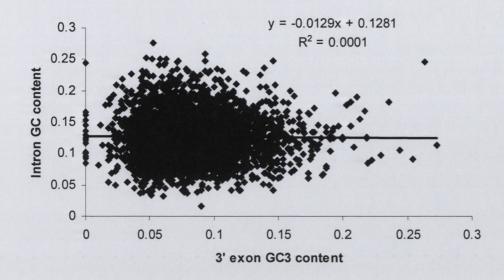


FIG. 8.7. Relationship between intronic and three prime exon GC3 content.

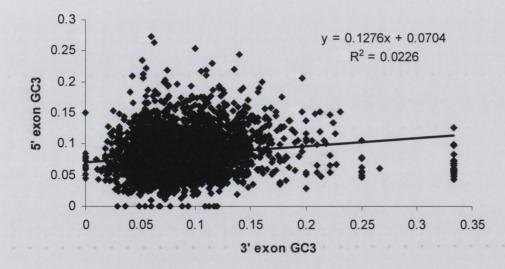


FIG. 8.8. Relationship between five and three prime exon GC3 content.

A plot of the mean AT content around the splice sites (Figure 8.9) shows a marked drop in the mean AT content within the exons immediately adjacent to the splice sites. This effect is lost within 3 bases of the splice sites.

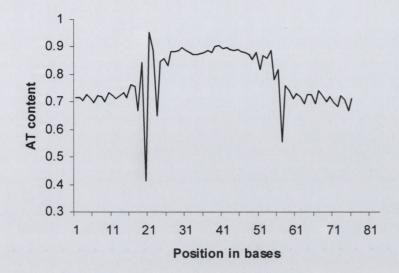


FIG. 8.9. Mean AT content surrounding the splice sites. The sharp troughs in the AT content occur within the exons adjacent to the splice sites.

Adjacent to the splice sites themselves (Table 8.2) base use is biased: a chi square test for association between intron phase and choice of base at the 5' splice site is highly significant - 229.5 ($p < 10^{-10}$) - suggesting that base and intron phase may be functionally linked. Of the phase 0 introns, the last base of the 5' exon is G in 52.0% of cases, A in 23.0, T in 19.5% and C in 5.5%. In contrast the overall third codon position is 81% AT. The difference is even more marked in the phase 1 introns with the ultimate base of the five prime exon being G in 62.2%, A in 25.3%, T in 9.0% and C in 3.6%. The pattern is continued in the phase 2 introns with the five prime base being G in 45.8%, A in 31.8%, T in 15.8% and C in 6.5%.

Α	С	G	Т
828	198	1875	702
659	93	1621	235
642	131	924	319
	828 659	828 198 659 93	828 198 1875 659 93 1621

TABLE 8.2. Bases lying immediately five prime of the GT splice site.

The base use lying immediately 3' of the AG splice site while biased (Table 8.3) is somewhat less so: the chi square for association being here 20.0 (p < 0.003). In the phase 0 introns the AG is followed by A in 40.6% of cases, G in 38.1%, T in 13.2%

and C in 8.1%. While these values for A and C are close to those in the general coding sequences the figures for G and T are not. The figures for the phase 1 and phase 2 introns are similar. When the intron is in phase 1, this base is A 41.4% of the time, G 35.9%, T 14.2% and C 8.5%. If the intron is in phase 2 this base is A in 42.9% of cases, G in 32.4%, T in 15.7% and C in 8.9%.

	A	С	G	Т
Phase 0	1464	293	1371	475
Phase 1	1079	222	937	370
Phase 2	865	180	653	317

TABLE 8.3. Bases lying immediately three prime of the AG splice site.

Among the symmetrical exons – those with the same intron phase before and after the exon - the phases are not distributed equally with the 0:0 phase exons accounting for 45.8% of the total (Table 8.4). A chi square test for equal distribution gives a value of 296.1 ($p < 10^{-62}$) confirming this impression.

0	1	2
3767	702	574
2246	746	542
1458	507	573
	3767 2246	3767 702 2246 746

TABLE 8.4. Exons by phase distribution. The 5' phase is given by the row and the 3' phase by the column

Of the plots of intron composition the A and T content are highly correlated (Figure 8.10) but the remaining plots are heteroscedastic and consequently difficult to interpret (Figure 8.11).

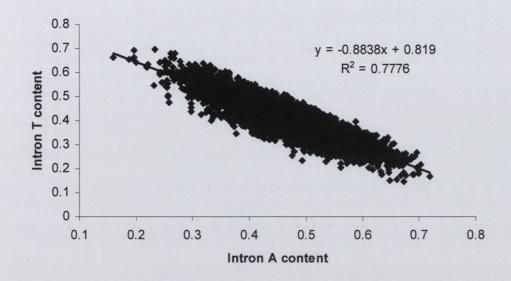


FIG. 8.10. Relationship between intronic A and T content.

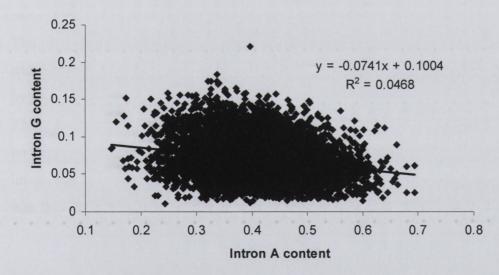


FIG. 8.11. Relationship between intron adenosine and guanine content.

Eighty six introns were found to have an exceptionally high C content including genes that might be expected to be expressed at high level: 60S ribosomal protein L7, 60S ribosomal protein L8, 60S ribosomal protein L18, translation elongation factor P, eukaryotic translation initiation factor 2 alpha subunit, glutaredoxin as well as a cyclophilin, heat shock protein hsp40 and others. Most of these have a run of 5 or more Cs. Aside from their relatively high C content these introns appear typical in other aspects. With their association with genes that can reasonably be expected to be expressed at high level, it seems plausible that this C rich element may enhance either translation or transcription in some as yet unknown fashion.

The mean of the mean A position was 41.8% of the intron length and that of the T means was 58.3%. Application of the Y test to the A and T bases within the introns shows that their positions are non random (t > 4.0 and p < 0.001 in all cases). Only 309 introns have 25 or more C bases: the bases are also non randomly organized (all t > 4.4, p < 0.001). There are 315 introns with more than 24 G bases and these also are non randomly organized (all t > 3.2, P < 0.001).

The observed A and T runs are longer (Table 8.5) than expected by chance (t = 29.50 and 42.28 respectively, $p < 10^{-20}$). The C runs are also longer than would be expected by chance (t = 2.14, p = 0.016) and are usually found toward the 3' end of the intron. These may simply be part of the pyrimidine tract. The G tracts are shorter than would be expected (t = 15.83, p < 10⁻²⁰). These tend to be found towards the 5' half of the intron but no correlation between their presence and that of the complementary C base was found suggesting that guanine runs of 3 or more may have a functional role.

	Expt A	Obs A	Expt T	Obs T	Expt C	Obs C	Expt G	Obs G
Mean	5.82	8.29	6.73	10.57	1.75	1.77	1.72	1.57
Std dev	1.40	7.29	1.55	7.90	0.34	0.89	0.33	0.77

TABLE 8.5. Means and standard deviations of the expected and observed run lengths.

Of all the intron containing genes 34 introns (1.6% of the total intron containing genes) have run lengths that lie outside three standard deviations of the expected means. This finding is compatible with an underlying unimodal distribution controlling the length of runs within the introns but the precise nature of this distribution has yet to be determined. Those lying outside these limits may have particular biological significance or may simply have occurred by chance. Clarification of this point will require additional work.

The expected distance of the AG dinucleotide from the 3' end of the intron was 27.7 +/- 20.25 bases. The observed distance was 60.63 +/- 88.87 bases. The t value for the difference is 57.4 (p < 10^{-20}). This is to be expected on biological grounds: to enhance recognition of the 3' splice site, AG dinucleotides are avoided near to it.

One motif was returned consistently by the Gibbs sampler: a sequence of alternating As and Ts of length 6-8 bases with an A in the penultimate position. This sequence is usually found within 50 bases of the 3' end of the intron. While it is not found in 403 (4.9%) of the introns inspection of these intron reveals motifs similar to this consensus sequence. While confirmation by experiment is lacking it seems likely that this motif may be the presently unknown lariat site of *P. falciparum*.

At the 5' site the increase in Schinder's information (S) in the base before the conserved GT (positions 1 and 2) is evident (Figure 8.12). Curiously the penultimate base of the 5' exon is even more significant which is in favour of the the idea that introns are located where they are for functional reasons rather than as suggested by either the intron early intron late theories. The dinucleotide following the GT site is biased and is usually AA. This suggests the optimal 5' splice site is AG – GT –AA and there is relatively little contribution from the adjacent sequence outside this hexanucleotide.

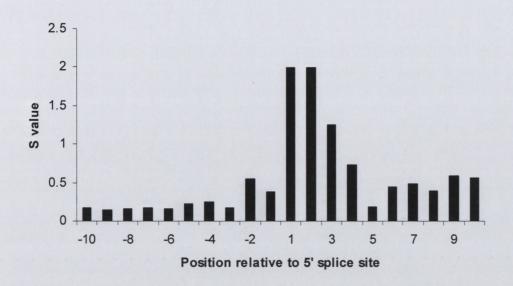
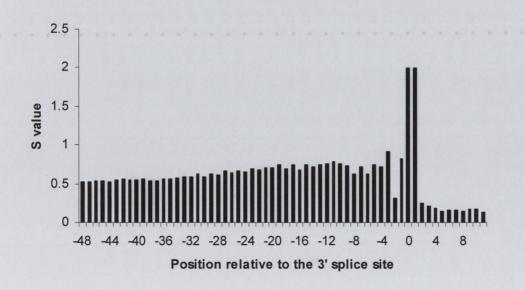
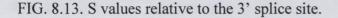


FIG. 8.12. S values relative to the 5' splice site.

The situation at the 3' splice site is somewhat different (Figure 8.13). The intron sequence for some considerable distance 5' to the AG site is biased. This is due to the presence of the pyrimidine tract 5' of the 3' AG site. The S value of the penultimate nucleotide is curiously low suggesting this has little or no contribution to make. The 3' exon S values are much lower than the adjacent intron values suggesting little contribution to splicing recognition.





8.5 Discussion

The difference in length between the intron containing and intron free genes appears to be novel and the biologic significance – if any - is presently unclear. This appears to be the first time such a finding has been reported. In other genomes first introns tend to be longer than other introns and it has been suggested that this may be due to the presence of regulatory sequences with these introns (112). While the existence of such sequences have yet to be described in *P. falciparum* this hypothesis seems plausible and further work may confirm this.

The reason for the presence of introns within genes is not known. If we assume that exons have some 'built in' limitation in length then a correlation would be expected to exist between the number of introns and the length of the gene. The lack of correlation found here suggests that there is no such limitation and that the introns instead posses some other function. While it could be argued that the greater length of the intron less genes also support this argument, there may exist unknown differences between these genes and those with introns which would render this point moot.

The introns have a higher pyrimidine and AT content than the exons. Base substitutions within introns are frequently treated as being random events. If we assume as a null hypothesis that the A and T bases occur randomly within the introns then we would expect the maximum length of these runs to obey the Erdos-Renyi theorem and that base content distribution would be normally distributed - this latter hypothesis following from the law of large numbers.

The lack of normality in the intron base content suggests that some selective force has acted on the introns. The linear correlation between the A and T content of the introns suggests a functional relationship between these - a hypothesis supported by the locations of these two complementary nucleotides within the introns. Inspection of the introns reveals that the 5' end of the introns tend to be A rich and the 3' end T rich with the long T runs mostly confined to the region 5' of the AG splice site. These runs which are longer than would be expected by chance is consistent with the known biology of 3' splice sites. The biased location of the A and T bases was statistically confirmed with the Y test. Given the linear correlation of the A and T content and their location within the introns it seems plausible that these features may play a role in splicing: possibly by forming a loop by complementary base pairing. This hypothesis is also consistent with the lack of a normal distribution in base content.

The lack of such a relationship between the relatively few C and G bases suggests potential complementary pairing is not functionally important.

Guanine runs are found in 784 of the introns – a group that includes ribosomal gene translation factors, a histone H2a variant and 43 of the 66 *var* introns. At least some of these genes are expected to be present at high level and it is possible that the G runs within the intron may either be a marker or have some role in this process.

It seems likely that a selective force may also act on the GC composition of the exons. If the GC3 content is synonymous then it would be expected that the GC3 content of the 5' and 3' exons would be similar. If this was the case a linear correlation with a slope of \sim 1.0 between the GC3 content of the exon pair would be expected. This was not found to be the case. There does not appear to be an easily identifiable pattern between these values suggesting that the forces on the GC3 content are more complex. Similarly if the GC content of the introns were linked to either that of the 5' or 3' exon some form of pattern might be expected. Again the plots (Fig 8.8) suggest that if a relationship exists between these values it is likely to be complex.

In addition to the location of bases within the intron, the location of dinucleotides may be informative. Except in the case of the 3' splice site integrating this information into biological theory is difficult. An AG dinucleotide is required for intron splicing and to enhance recognition of this site by the splicing system selection against a second AG lying close to the 3' intron splice site would be expected. The results here are consistent with this hypothesis.

How the exons and introns are distinguished by the splicing system is not entirely clear for any organism. The contrast in the average AT content between the exons and introns (Fig 8.9) suggests that these bases may play some role in intron recognition. Since this contrast is visible to the unaided eye it seems likely that there exists some as yet unknown molecular analogue.

Base use immediately before and after the splice sites is biased. For the 5' sites a purine is preferred over a pyrimidine. G is more common than A despite of their relative frequencies in the genome: C is less common that T but its use relative to T is much closer to their frequency in the coding sequences than those of A and G. These patterns are found in all three phases. The avoidance of T in the first intron phase in particular makes biological sense particularly in an AT rich genome. All the stop codons begin with T and a single point mutation could convert the gene into a

pseudogene or at least significantly reduce its expression. The reasons for the remaining biases are not so obvious.

Following the AG site again there is a preference for a purine: in this case A is preferred over G. Again the use of G over A in this position is more common than would be anticipated from their frequencies in the coding sequences. T is again more common than C in this position.

Exon phase is non random. Two patterns are noticeable: (1) at least one end of the exon is in phase 0 and (2) considered by 5' phase symmetrical exons are the most common. Symmetrical exons are exons where the 5' and 3' phases are identical. In the case of the phase 2 exons the 2:2 exons are the second most common type by a single exon. The length in nucleotides of the symmetrical exons is always a multiple of three. These exons could be spliced out without distorting the remaining coding frame. Whether this occurs *in vivo* is not known. The reason for these patterns is not known. Much ink has been spilled over whether or not the patterns in exon and intron phase or preference for base use adjacent to the splice site supports either the 'intron early' or 'intron late' theories (164, 283). The relative merits of these positions are beyond the scope of this work.

More recently it has been proposed that these patterns of base use close to the splice sites may owe less to historical accident and more to selection pressures. The splicing process is complex and poorly understood. In humans at least 145 proteins are involved in this process (562) and it is known to be coupled with export of the mRNA to the cytosol (393). In *P. falciparum* the hexanucleotide at the 5' splice site and the pyrimidine tract at the 3' end of the intron are similar to those found in other organisms. The G lying immediately before the 5' GT appears to have a role to play but the penultimate 5' base appears to be even more significant, something that seems to have been overlooked in the earlier intron theories but is consistent with intron conservation for functional reasons. There is no experimental data from *P. falciparum* to guide our intuition as to the import of the other findings here.

Ruvinsky *et al* have suggested that the intron splicing mechanism works more efficiently with particular bases adjacent to the splices sites and that conservation of intron location is a consequence of this preference (414). Their results of their simulations were close to those observed in the data set. This hypothesis explains the biased base use and the conservation of intron position within the gene. If this hypothesis is correct then base use around the splice sites in addition to those

immediately adjacent may also be biased. This hypothesis is consistent with the findings from the Schnieder statistic findings.

Chapter 9: An application of the annotation

9.1 Background

One major obstacle in the development of materials that are selectively toxic to pathogens has been - and continues to be - our ignorance of the metabolism of important pathogens at a molecular level. For this reason - among others - many genome sequencing projects – including that of *P. falciparum* – have been undertaken. Before genome projects became possible, the nihilistic philosophy of *ignoramus et ignorabimus* - that the end of scientific investigation is not the discovery of truth but merely the seeking after truth without ever finding it – was prevalent. With the genome annotation now available it should be possible – at least in theory - to reconstruct much of *P. falciparum*'s metabolism.

Metabolic reconstruction requires not only the sequence and function of all the encoded proteins but also data on the timing and quantity of RNA expression – transcriptomics – and of the proteins - proteomics - and the degree and locations of their activity and interaction with other proteins. These form part of the growing disciple of systems biology (504) – a disciple that seeks to integrate the different levels of information to understand how organisms function. While not all of the relevant data is available these methods are being applied to *P. falciparum*: (556) non radioactive isotopes are being used to study the quantities of the protein during the different stages of the life cycle (38) and gene transcription is being studied in both the host and the parasite (557).

P. falciparum presents interesting challenges to metabolic reconstruction. Catalase is not present in the genome and it appears that the parasite makes use of the host's erythrocyte's catalase. Furthermore glucose uptake by the infected erythrocytes is increased 50-100 fold with almost all of the glucose subsequently being converted to lactic acid with the infected cells containing 10 times more NAD(H) than uninfected red cells while the NADP(H) content remains unchanged (411). The parasite posseses only one lactic acid dehydrogenase gene. A second gene is present whose product may be either a lactic acid or a malic acid dehydrogenase: deciding this point will require experimental investigation. Whether these genes alone account for the

increased turnover or if the erythrocyte's metabolic system is also modified remains unclear.

The *P. falciparum* annotation will be of assistance here in four ways. Firstly by providing the sequence data it will make the task of cloning the genes somewhat less difficult and provide assistance to those who only have partial sequences to work with. Secondly by linking the gene annotation to images of location of the proteins within the parasites and to other relevant data, a more comprehensive understanding of the parasite's metabolism may be constructed.

Thirdly the annotation will help with automated reconstruction of the metabolism. In the present annotation 954 proteins with putative enzymatic activity were identified and of these 594 could be assigned Enzyme Commission (EC) numbers uniquely identifing their activities. With the collaboration of Professor K. Tipton (Trinity College, Dublin) 353 of these enzymes - including all the known biochemical pathways in the parasite – have been mapped into metabolic pathways. This work is beyond the scope of the current thesis.

It also became apparent that a number of enzymes known from experimental work to exist within the genome were not given in the MGP's annotation. An example: phosphoglucose isomerase (EC 5.3.1.9) - the enzyme that converts glucose-6-phosphate to fructose-6-phophate – and now known to be on chromosome 14 - was not present in the MGP annotation.

While the MGP identified a single triose phosphate isomerase on chromosome 14 with a constant level of transcription, a second gene was found on chromosome 3 with a transcription profile that varies during the asexual life cycle. The reason for the differing transcriptional profiles is not clear but may well reflect some as yet unexplored aspect of carbohydrate metabolism in this organism.

Finally computational analysis of the genome may identify new metabolic pathways that are not easily found with other methods. This method will be illustrated with the role of calpain in the invasion of the erythrocytes by the parasite's merozoites.

9.2 Introduction

Merozoites are ovoid or elliposoidal cells (1.0 x 1.5 μ m) with a three layered membrane covered in filaments whose nature is presently unknown (23). At one end

is the nucleus with the tubular mitochondrion adherent to the cylindrical plastid (0.5 μ m x 0.15 μ m), at the other (the apex) a set of organelles – the micronemes, dense granules and rhoptries and between these lies the cytoplasm, rich in ribosomes. Deep to the plasma membrane of the apex lie three dense proteinaceous rings (polar rings) to which are anchored the apical border of the two inner membranes and a longitudinally running band of subpellicular microtubules.

The rhoptries, micronemes and dense granules are membranous vesicles involved in erythrocyte invasion. The two rhoptries are pear shaped, densely staining bodies 650 nm long and 300 nm wide at their widest diameter and their narrow ends (the rhoptry ducts) converge on the flat end of the apical prominence. The smaller (120 x 40 nm) micronemes cluster around the rhoptry ducts and probably discharge their secretions through the ducts during erythrocyte invasion. The dense granules are rounded bodies (80 nm diameter) lying within the apical cytoplasm with densely packed granular contents. The contents of these organelles is not completely known but they include the three rhoptry proteins, apical membrane protein antigen 1, Pf60.1, serine proteases, ring erythrocyte surface antigen, ring infected membrane antigen, erythrocyte binding antigen 175 and others (421).

Invasion of the erythrocyte by a newly released merozoite requires approximately 30 seconds. The merozoite adheres to the erythrocyte via the filaments, rotates itself until its apex is adherent to the erythrocyte membrane, discharges the contents of the apical organelles and burrows into the red cell. The rhoptries and the micronemes provide lipid to form the parasitophorous vacuole into which the parasite burrows within the erythrocyte while the dense granules are involved in the invasion mechanism itself (22). The process of invasion involves the formation of a tight junction between the merozoite and the erythrocyte membrane and the invasion route forms part of the parasitophorous tract (5).

The role of proteases in the invasion process is only partly understood but appears to involve at least two steps (72). Primary processing of the surface proteins including the removal of terminal peptides or segmentation of the proteins into multiple fragments activates the adhesive ligands for binding to host receptors. The second step - freeing the merozoite from its attachment to the erythrocyte - occurs during the penetration of the erythrocyte and leads to either the shedding of the proteins from the surface of the merozoite or first displacing them to the posterior part of the cell before

shedding them with this latter fate being preferred for those proteins released from the secretory vesicles.

While most of the proteases involved in this process remain unidentified two proteases - calpain and Sub2 – are known to be involved in this process (41, 513). Calpain (EC 3.4.22.17) is a Ca²⁺-dependent cysteine protease first isolated in 1978 whose pH optimum lies between 7.0 and 8.0 (218). There are at least 15 distinct calpain genes present in the human genome with several having a number of isoforms (up to 10) and along with the ATP-dependent proteasome, calpain appears to be responsible for the majority of non-lysosomal targeted proteolysis.

Calpain is a member of the papain superfamily (289) - a group of proteases that includes papain, calpain, streptopain, ubiquitin-specific peptidases and many families of viral cysteine endopeptidases. It is a protein of ancient origin with homologues found in vertebrates, insects, crustaceans, nematodes, fungi, higher plants, *Dictyostelium*, kinetoplastid protozoa, and bacteria. It evolved from a gene fusion event between an N-terminal cysteine protease and a C-terminal calmodulin-like protein, an event predating the eukaryote/prokaryote divergence (193).

The enzyme cleaves preferentially on the C-terminal side of tyrosine, methionine or arginine, preceded by leucine or valine - i.e. P1 = Y, M, or R; P2 = L or V according to the established nomenclature (427). Calpain occurs either as a heterodimer with a small regulatory subunit and a large catalytic subunit or as the catalytic subunit alone (425). It has been crystallised and its structure has been solved for several species (207, 485) and the active site consists of a conserved triad of cysteine, asparagine and histidine. The catalytic domain is divided into two subdomains (2a and 2b) with the cysteine residue lying in domain 2a and the histidine and asparagine in 2b. Calpain has a natural monomeric protein inhibitor, calpastatin (503) and in the presence of Ca²⁺, calpain undergoes a conformational change, dissociates from or cleaves the associated calpastatin and finally cleaves its own first domain to become fully active.

Calpain is involved in cell fusion reactions (402) and substrates of this enzyme appear to be recognised principally by the presence of PEST sequence(s) within the protein although exceptions are known (391, 447, 524). PEST sequences were first described in 1986 (400) and are short subsequences (usually 10 - 60 residues) within proteins that are bounded by but do not contain basic residues (H, K or R), and are enriched in proline (P), glutamate (E), serine (S), threonine (T) and aspartate (D) residues. An algorithm (the PEST-find score) has been described for assessing the significance of

such subsequences: a score of 5 or greater is regarded as significant. PEST sequences are found in ~10% of all cellular proteins in the organisms analysed to date and are typically found in highly regulated proteins. PEST +ve (PEST sequence containing) proteins typically have short half lives (0.5 to 2 hours) in intact cells compared with most other proteins (>24 hours). In PEST +ve proteins, removal or disruption of the PEST sequence increases the protein's half life to more "normal" values while insertion or creation of a new PEST sequence within a PEST -ve (PEST sequence free) protein decreases that protein's half life to a value typical of a PEST +ve protein. Two papers have described the effects of calpain inhibitors on P. falciparum. The first described the effect of calpain inhibitors on the invasion of erythrocytes (352). The authors found the inhibitors used were ~100 times as potent (IC₅₀ ~10⁻⁷ M) than the other protease inhibitors (chymostatin, leupeptin, pepstatin A and bestatin) examined. Erythrocytes normally contain only calpain 2 and it was not clear if the effect of these inhibitors was as a result of its action on the parasite's and/or the erythrocyte. This was clarified by Hanspal et al. who reinvestigated this effect in calpain 2 knock-out mice (189). While the mouse erythrocytes were shown to have no detectable calpain activity, they still supported the invasion and growth of P. falciparum in culture. Calpain inhibition again prevented re-invasion. A third paper has shown that removal of Ca²⁺ from the growth medium results in growth arrest in the late trophozoite stage and failure to invade erythrocytes (529) findings consistent with a role for calpain in the parasite life cycle.

Biological evidence for the involvement of calpain in *P. falciparum* has been supported by the discovery of a putative calpain gene (Pf13_31:2391896w) on chromosome 13 (553). With 2,047 residues the *P. falciparum* enzyme is unusually large - more than twice the size of other known calpains which are usually 600 - 800 residues long. The catalytic domain is the only part of the enzyme with similarity to the vertebrate enzymes and in the *P. falciparum* gene this domain is unusually distant from the N-terminus (residues 1002 - 1470): in other organisms the active site lies within 150 residues of the N-terminus. The 5' end of the *P. falciparum* gene contains a low-complexity region and it seems likely that a fusion event has occurred at the original 5' end of the calpain gene with a second as yet unidentified gene. If the activation of the *P. falciparum* calpain is similar to that of other organisms, this 5' domain would be removed during enzyme activation. This new element may be

responsible for the selective toxicity of the inhibitors or may play some regulatory role.

PEST sequences have been previously identified in *P. falciparum* by Mitchell and Bell (313). Examination of chromosome 2 identified 27 proteins (13%) with PEST sequences (PEST +ve) of which 18 were surface exposed proteins including the merozoite surface antigens 2 and 5 - known to be cleaved during erythrocyte invasion. The other PEST +ve proteins were either hypothetical proteins or proteins involved in DNA replication or intermediate metabolism. With the annotation of the genome now almost completed it seemed pertinent to examine this new and larger data set for the presence of PEST +ve proteins.

9.3 Methods

The PEST score was calculated with the standard algorithm. Sequences of 10 or more residues bounded by, but not containing basic residues (H, K or R), were first identified. The mole percent (MP) of this subsequence was then determined after subtracting one mole equivalent of P, E/D and S/T. The normalised hydrophobicity value is the value of the Kyte-Doolittle index (254) for that residue multiplied by 10 plus 45, giving values between 0 and 90.

Stellwagen (http://emb1.bcc.univie.ac.at/embnet/tools/bio/PESTfind/) has suggested that a value of 58 for tyrosine rather than 32 as originally given gives a more reliable PEST score: the former value was used here. The average hydrophobicity (Ho) of a subsequence was determined by summing the MP of each residue and its normalised hydrophobicity value. The PEST-find score is 0.55 (MP) - 0.5 (Ho). The original paper has a misprint with PEST-find = 0.5 (MP) - 0.5(Ho).

9.4 Results

Five hundred and twenty PEST +ve (PEST scores > 5.0) proteins were found here (Table 9.1). The entire set of these genes is available in the accompanying database. The PEST sequences varied in length from 12 to 311 (27.3 +/- 20.0) residues. Of these proteins 280 are hypothetical proteins or proteins of unknown function, 120 are surface expose proteins and 120 are involved in cell metabolism. Most of those to which a function can be currently ascribed are given in Table 9.1 Some of the surface

exposed proteins are involved in invasion of the erythrocyte while others are known to be involved in adhesion in other parts of the life cycle.

Meroz	Merozoite surface proteins				
Rhopter	ry proteins 1, 2 and 3				
Merozo	ite surface proteins 1, 2, 5 and 7				
Merozo	ite capping protein 1				
Erythro	cyte binding antigens 175 and 181				
Reticul	ocyte binding protein homolog				
Surfac	e proteins of other stages				
EMP1 ((49)				
EMP11	ike protein				
Serine r	repeat antigens (7)				
Rifin (9)				
Pf322					
Acidic	basic repeat antigen				
Falcipa	rum interspersed repeat antigen				
Mature	parasite infected erythrocyte surface antigen				
Ring in	fected erythrocyte surface antigen				
Maurer	's cleft protein				
Skeleto	n binding protein 1				
Secrete	d protein with thrombospondin domain				
Circum	sporozoite protein and thrombospondin related protein				
Circum	sporozoite related antigen				

Asparagine rich antigen	(2)
Replication related p	proteins
Mitotic control protein	
Origin replication comp	lex protein
Teleomerase	
DNA polymerase (7)	
Mismatch repair protein	
Minichromosome maint	enance protein
Structural maintenance	of chromosomes protein
Regulator of chromoson	ne condensation protein
Chromosome condensat	ion protein
Chromosome binding pr	rotein
CDC 2 kinase like prote	in
Transcriptional prote	eins
DNA helicase (10)	
Histone deacetylase	
Silent information regul	ator homologue
Transcription factor	
RNA polymerase 1 and	2
Basic transcription facto	or 3

Splicing proteins

Alternative splicing factor

Pre-mRNA splicing factor

Splicing factor 3b subunit 1

Spliceosome associated protein

U5 small nuclear ribonucleoprotein specific protein

U1 small nuclear ribonucleoprotein

Cleavage and polyadenylation specificity factor protein

Small nuclear riboprotein auxiliary factor

RNA helicases (3)

RNA binding protein

Translation proteins

Translation initiation factor 2

Elongation factor G

Elongation factor 3 related protein

Ribosome releasing factor

Ribosomal proteins (4)

Co-chaperone grpE

Cyclophilin (9)

Heat shock protein (16)

Glutathione pathway

Glyoxalase 1	
Gamma glutamylcysteine synthetase	
Oxidoreductase	

TABLE 9.1. Partial listing of PEST +ve proteins. Where there is more than one paralogs, the number of paralog is given in parentheses after the protein. Abbreviation: EMP1 – erythrocyte membrane protein 1.

The findings here extend the number of PEST +ve proteins involved in cell adhesion. PEST +ve proteins are also found within the erythrocyte. Spectrin, band 4.1 and ankyrin proteins known to be bound by *P. falciparum* merozoites during the invasion process and are all PEST +ve (138). Band 3, another PEST +ve erythrocyte protein, is eliminated from the site of contact with the merozoite (115): all of these erythrocyte proteins are putative calpain substrates.

The presence of PEST sequences in a subset of *var* gene products, the serine rich antigens and the rifins is suggestive of differential processing or cellular turnover and may shed some light on the reason for the large number of these genes in the genome. The presence of PEST +ve membrane proteins suggest that this system may be involved in cell adhesion processes in *P. falciparum* other than the merozoite. Also curious is the absence of PEST sequences in other genes known to be linked to adherence – cytoadherence linked asexual genes (CLAGs) and sub-telomeric variable open reading frames (STEVORs).

PEST +ve proteins are common in DNA-binding proteins and are part of the control of intermediate metabolism (83). As the parasite progresses from the trophozoite to the schizont stage, there is a thirty-fold increase in *P. falciparum*'s level of transcription of calpain. The erythrocyte normally maintains a submicromolar intracellular Ca²⁺concentration which rises thirty-fold as the parasite matures. PEST sequences are present in nine proteins known to be involved in DNA replication. Given the effects on Ca²⁺ removal on trophozoite to schizont progression, it is possible that the lack of Ca²⁺ may inhibit calpain activation and that this in turn may be responsible for the effects seen in the calcium depletion experiments (159, 300, 493, 528).

The presence of PEST sequences in proteins involved in the tightly controlled central processes of transcription, splicing and translation was expected. In contrast the presence of PEST sequences in the proteins of the glutatione pathway was surprising. Naively one might expect these proteins to be relatively long lived as they are involved in the redox balance of the cell. These proteins are either unusual in being PEST +ve and relatively long lived or else *P. falciparum* may be actively managing its redox balance during its life cycle - questions cannot be settled without experimental work.

A number of other PEST +ve genes involved in metabolism have not been listed here. Most are phosphatases or kinases whose targets are not yet known. As these proteins are likely to be regulators themselves, and important to the parasite, their levels would also be expected to be tightly controlled.

9.5 Discussion

This is not the first attempt to reconstruct metabolic pathways from a genomic annoation. Large scale attempts are being made presently in other organisms - *Escherichia coli* and *Saccharomyces cerevisiae* - where considerably greater detail is available concerning the protein-protein interactions, metabolic flux, transcription and translation (287). These studies suggest that organisms differ not only in gene content but also in the interactions of proteins and the flow of metabolites along the pathways – physiological differences that may be exploitable. This strategy is already being investigated in the large pharmaceutical companies seeking to understand how their agents work and seeking new targets (146).

The *P. falciparum* calpain gene differs significantly from those found to date in vertebrates which may explain at least in part the demonstrated selective toxicity of the inhibitors. The lack of a calpastatin analogue in the genome is curious. While it is possible that its orthologue has yet to be identified within the genome, it is also possible that the large N terminal domain may be acting as it own inhibitor - a question that cannot be answered without experimental work.

PEST sequences have also been shown to be involved in the degradation of proteins within the proteasomes (266). The proteasome is a barrel-shaped 20S particle that typically comprises about 1% of cell proteins. This particle consists of four stacked rings that enclose a central chamber where proteolysis occurs. The proteolytic process

in this particle requires the presence of ATP. This mechanism accounts for 80-90% of protein degradation in a typical mammalian cell. Components of this system are found in *P. falciparum* including the 26s proteasomal regulatory subunit (Pf2_1:235237w). Proteins are targeted for degradation by this system by the covalent addition of ubiquitin. The enzymes needed for this pathway are present in the genome eg ubiquitin conjugating enzyme E2 (Pf6:261027c) suggesting that this pathway is functional. The significance of its role in *P. falciparum* is not known and may much less significant than in other organisms. This parasite leaves a number of proteins when the merozoites escape from the plasma membrane of the schizont. This 'residue' may include degraded proteins.

In the role of PEST sequences in targeting proteins for degradation by the proteasome has been investigated previously. Some proteins - neural proliferation and differentiation control protein-1 (471), Notch1 receptor (302), cyclin D1 ((260) and cytoplasmic polyadenylation element binding protein (395) – appear to use the PEST sequence as a marker targeting the protein for degradation by the proteasome while others - retinoic acid receptor gamma and retinoid X receptor alpha (49), erythroid kruppel-like factor (384) – PEST sequences do not appear to be involved in targeting these to the proteasome. At the present there do not appear to be any useful rules to decide if a PEST sequence is involved in targeting a protein to the proteasome and this will have to be determined in additional proteins in a number of organisms before any such rule can be formulated. Given the effects seen by microscopy it seems most likely that the killing mechanism of the calpain inhibitors is by prevention the invasion of the erythrocyte rather than a block of proteolysis in the proteasome. This hypothesis can be tested by examining the effects of a number of calpain inhibitors with varying effects on proteasome inhibition.

A number of calpain inhibitors are currently available and the majority of these are small peptides that can be freeze-dried and stored at room temperature – properties important for any drug. Several have been used in phase 2 human trials for treatment of myocardial infarction, stroke and cancer. By themselves they may be worth exploring as alternative antimalarial treatments. If however the PEST hypothesis applies to the proteins involved in DNA, RNA and protein synthesis – something that awaits experimental confirmation – given the rate at which *P. falciparum* develops resistance then investigation of calpain inhibition with agents acting elsewhere on these pathways may be interesting.

Chapter 10: Summary and future work

10.1 Introduction

Malaria is and seems likely to remain a serious problem for many countries for the foreseeable future in spite of considerable efforts to control it. That this seemingly Herculean task may be possible is supported by the history of this disease. Much of the disease burden of malaria was lifted beginning in Roman times with the draining of marshes adjacent to human habitation, provision of piped water and control of the sewage disposal. Aside from the control or even elimination of malaria these public works also contributed to the control of other diseases. These lessons were lost for many years and were only gradually rediscovered. Once the causative organism and its vector were identified the task of malaria control was simplified. The eradication of malaria in Italy provides an illustration. In the 1930's malaria was present in much of Italy (184) and a malaria eradication programme was begun in 1947 and within one year transmission had been interrupted. (312) In 1970 the WHO declared Italy malaria free.

These methods used in Italy and elsewhere may not be as applicable to the countries where malaria remains endemic. Many of these countries have large rural or nomadic populations living in close proximity to bodies of water suitable for mosquito development. Additionally there is a growing need to discover new therapeutic materials with the problem of increasingly widespread resistance to the commonly used drugs to treat those who now have malaria. An effective vaccine would be an enormous contribution to the control of this disease. With such an agent eradication of this disease might then be possible much as the cowpox contributed to the eradication of smallpox. For these reasons the malaria genome project was begun.

While the first genome – that of *P. falciparum* strain 3D7 – was reported in 2002 (158, 186, 216), several other *Plasmodium* genomes including those of *P. berghei*, *P. vivax*, *P. reichenowi*, *P. chabaudi* and a second strain of *P. falciparum* (HB3) are now being sequenced. The only other reported *Plasmodium* genome to date – that of *Plasmodium yoelii* (71) – remains only partly assembled. Once these projects are completed it is hoped that the insights gained from them may help guide the development of new strategies to control this disease.

To obtain optimise the benefit from this work an understanding of the organisation of the genomes and reliable methods of annotation will be needed: it is hoped that some of the material presented here will be of use in this work. The material presented here will now be reviewed and suggestions will be made as to its future application.

1.2 Organisation of the genome

The genome of *P. falciparum* is organised into 14 nuclear chromosomes and one each for the plastid and mitochondrion. While it was initially thought that there were only 8-10 nuclear chromosomes (452) it now seems likely that all species within the genus have 14 nuclear chromosomes (223, 299). It is not yet known if the other *Plasmodium* species posses the isochores found in *P. falciparum* as the erythrocyte membrane protein 1 (*var*) genes seem to be unique to *P. falciparum*: in this genome these genes are always associated with a relatively elevated GC content. The putative centromeres of the other genomes have not yet been identified: given the difference that exists in genome composition (*P. falciparum* ~ 80% AT; *P. vivax* ~ 40% AT) comparison of the centromere composition and structure may be of interest as even between closely related species these structures may be quite diverse. (24)

The putative origin of replication within the mitochondrion and plastid may also be worth investigating in the other *Plasmodium* species. The organisation of these smaller chromosomes tends to be conserved (387) so it seems likely that the origin of replication may also be conserved between the species. Knowledge of the biology of these organelles may be of considerable use as it is known that some of the agents used to treat malaria target these. (100)

Chargaff's second rule holds for all eukaryotes examined to date (316) and the A+G ~ 50% rule holds for the plastid and mitochondrial genomes: it is anticipated that these will also hold for the other *Plasmodium* genomes. It is also expected that these other *Plasmodium* genomes will also obey the laws governing genome base position. (314). Similarly the other *Plasmodium* genomes are likely to obey the isostich rules found here: (1) the log of the number isostichs is dependent in a linear fashion on the length of the isostich (2) the intercept in this regression is itself a linear function of the log of the length of the genome and (3) a knot occurs in the plot between isostich lengths 8-10. These predictions are based on the finding that the first two of these appear to be universal laws of genome construction applying to RNA and DNA genomes of both

the single and double stranded variety. (315) The third rule appears to apply only in double stranded DNA genomes as these are the only genomes apparently sufficiently long to host large numbers of isostichs with lengths greater than 10. Why this should be so it not known at the present time.

10.2 Annotation methods

The presence of significant errors in the published genome was not overlooked. Hill had noticed that the proteins used in the vaccine (RTS,S) he had been working with for some time in clinical trials in The Gambia (44) had not been included in the annotation (Hill A. V. 2003, personal communication). While other experts had noticed that additional genes known to be present were also missing, this was perhaps the most obvious example of this problem.

While Huestis was able to annotate correctly relatively small sets of genes (211, 499) his methods began to break down when applied to larger sets of genes (211): the accepted but then unpublished version of the revised annotation of chromosome 2 contained incorrect splice sites out of a total of 200. These errors were identified and then corrected with the methods described here. Examination of Dr. Huestis' unpublished gene annotations revealed a similar error rate which prompted the whole scale adoption of the methods used here. This in turn lead to the creation of a new curation project to complete the published *P. falciparum* genome (35).

The annotation methods described here have subsequently been used to annotate parts of both the *Plasmodium chabaudi* (134) and the *P. falciparum* genomes (210) and it is thought that these could be used to annotate any *Plasmodium* genome. Dr. Huestis (unpublished work) has been using the rules described here to annotate both the *P. yoelii* and *P. vivax* genomes. Laboratory confirmation of these predictions is in progress.

The Chebyshev and Vysochanskiï-Petunin inequalities as applied to codons and base use are novel methodologies for annotation and need to be tested on other similarly large data sets. While in theory these inequalities should identify any sequence annotated as a gene that is behaving in an anomalous fashion in any genome, this hypothesis remains to be confirmed.

The approach used here with a combination of computer predictions, manual review, statistical quality control and the use of multiple independent data sources seems to be

similar to other recent recommendations (103, 178, 271). Comparisons between multiple genomes is likely to improve the quality of the genome annotation still further and seems likely to resolve - at least in part - the problem of identifying non canonical splice sites and alternative splicing – two problems that remain difficult to solve.

The system described here for gene identification is in theory applicable to any genome. In practice it has been applied to the annotation here and has been adopted by the Sanger Centre at least on an interim basis for the unpublished *P. vivax* annotation. If it is used in the published genome this system may see wider adoption.

10.3 Gene organisation

The work in this thesis was based on the December 2004 version of the genome. Since then a number of additional genes (\sim 200) have been identified. These additional genes may affect the results of the gene organisation along the chromosomes so these results should be interpreted with caution. Similarly they will affect the findings concerning the intergenic distances but this is less likely to be significant given the known marked existing variability.

An extant and singularly difficult problem is the identification of promoters and other transcription modifying sequences with the genome. Examination here of this genome suggests that while in general these control elements are here likely to lie with 2000 bases of the translation initiation codon on average, some may lie within the coding sequences of adjacent genes. Experimental confirmation of these predictions seems desirable.

The distribution of protein lengths found here needs to be compared with those in other genomes – prokaryotic, eukaryotic and archeabacterial – as the pattern found here may or may not be unique. While it is anticipated that the bimodal distribution of predicted pI values found here is found in all genomes, this also needs confirmation. If this assumption is correct, then on theoretical grounds, it seems likely that the trough in the distribution reflects the cellular pH. This last hypothesis will also need experimental confirmation. The linear relationship between acidic and basic residues appears to hold in other eukaryotic genomes but the slope – 0.9 for *P. falciparum* – appears to vary between genomes. The cause of this variation in the slope of the line is under investigation.

10.4 Analysis of codon use

Codon base use was an essential part of the annotation methods used here. As noted earlier these distributions lay within four standard deviations of the mean rather than three as would be expected of a unimodal distribution. Currently the data available does not allow us to separate the genes into those that are transcribed solely in the human and those that are transcribed only in the mosquito. This is clearly an investigation that needs to be examined once the data is available. It may be that within each group of genes base and codon use lies within three standard deviations of the group mean; alternatively there may be such an overlap that these genes cannot be so divided. In organisms with less complex life cycles one might expect the base use to lie within three standard deviations but this also needs to be examined in other genomes.

The correlation of (A2 - T2) and hydrophobicity was anticipated given the differences codons with A2 and T2 respectively make to the overall hydrophicity of the protein. It is probable that this is a general rule and not one specific to *P. falciparum*. Assuming this is the case and that the slope of the regression is constant between genomes, this correlation may prove useful seeking possible membrane spanning regions within genes: such regions must be at least moderately hydrophobic: and the relationship found here would require that they are rich in T2 and poor in A2 content.

Frappet *et al* (144) found that codon use in 40 species of eukaryotes and 5 chloroplasts when ranked in order of their occurrence in the genome could be fitted by a combination of a constant, a linear function and an exponential function. The authors found that the parameters of their equations were heavily dependent on the GC content of the genome - consistent with what is known about codon use. These findings are similar to those here where codon and amino acid use ranked in order of their use can be fitted by an exponential function with a small number of outliers. The fact that this is found in several genomes suggests similar functions may control codon use in all eukaryotic genomes. If this is true for all eukaryotic genomes then this would appear to be a previously unrecognised constraint on codon use in these genomes. It seems worth noting that this relationship has only been tested only in eukaryotes and may not apply to other genomes: this possibility needs to be investigated.

The results of the correspondence analysis are typical of eukaryotic genomes and contrast with the results obtained from bacterial ones where significant trends may be found. Codon and amino acid use near the N and C ends of the proteins are often - if not universally - biased. In the *Escherichia coli* genome AAA in the codon second position is both the most common codon overall and the most common second position codon in the highly expressed genes. (475) A over representation of G in the first codon position is also found in the highly expressed proteins. (181) These genes also avoid the use of NGG codons close to the initial ATG (168). These patterns of codon use are limited to the first 20-30 codons and similar patterns are also found in other bacteria (399).

Additionally Rocha *et al* (399) found that the TAA stop codon was preferentially found in genes of high expression, that its use correlated with high GC content within the gene, that the terminal region of the proteins in general was enriched in A content and that the amino acid composition of the terminal region of the protein was atypical of the whole. Similar results have been obtained in other genomes (323, 438, 554). A non random association of the base following the stop codon similar to that found here has also been reported before (495).

The trends found here in the amino acid and base use near the initiation and termination codons appear to be consistent with those found in other organisms. While the results here are similar to those found in other genomes they have not previously reported before for any *Plasmodium* species. Since translation is a chemical process and is influenced by the temperature of the reaction, it is possible that the patterns in the genes expressed solely in the human and the mosquito may differ significantly: the core temperature for humans is ~38 degrees Centigrade while the mosquito prefers a temperature between 20 and 30 degrees. (27) Although there is insufficient data currently to test this hypothesis, this question may be worth revisiting when this data becomes available.

10.5 Intron organisation and structure

Introns were found in 42.8% of the genes annotated here. Correlations of the strength found between intron number and chromosome length do not appear to have been reported before. This finding may be unique to this genome and needs investigation in other intron containing genomes. Introns, with the exception of the single intron of the

erythrocyte membrane protein 1 genes, tended to be less than 500 bases long. Their mean AT content was 87.5% - significantly higher than that of the coding sequence surrounding them. The preferred 5' splice site was AG – GT and that of the 3' splice site was TAG – N. The GT site was almost invariably followed by a purine rich run of bases and the 3' site was invariably preceded by a pyrimidine tract. These features are common to many eukaryotes (401).

These combination of features combined with the presence of a run of alternating purines and pyrimidines of 5 or more made intron recognition here a relatively straight forward task. The reason for the presence of this alternating sequence of bases remained obscure until work here with the Gibbs sampler suggested that this sequence may be acting as the lariat site of the splicing mechanism. While this remains an untested hypothesis, it is compatible with the presence of this sequence in the introns and experimental testing of this hypothesis now seems indicated.

There is no obvious relationship between the GC content of the introns and the GC or GC3 content of the surrounding exons. Nor is there any obvious relationship between the GC or GC3 content of the exons surrounding the intron. If mutations in the third base position or within most of the intron are neutral as the neutral theory of evolution suggests then an approximately linear relationship might be expected between these values. The low GC content of the introns (~12.5%) and of the exons (~25%) here may perhaps have complicated the relationship here and a more linear relationship may be more obvious in other genomes. The absence of this finding does not contradict the neutral theory *per se* but suggests that if such a relationship does exist in this genome, it is likely be complex. This needs investigation in other genomes and in particular in *Plasmodium* species.

The correlation of the A and T content within the introns suggests that upper and lower limits to intronic AT content exist in this genome. This finding may be unique to this genome and needs investigation in other genome. Combined with the presence of long blocks of A and T tending to lie at opposite ends of the intron this suggests that these features may be connected with the intron splicing mechanism. The mechanism of splicing to date has not been studied in this organism so the findings here remain putitive.

The assumption of neutrality in the evolution of the introns in this genome has been used to date the origin of this organism as a serious human parasite (519). This

assumption is at variance with the statistically significant clustering of bases and their preferential location with the intron found here.

Base use around the introns splice sites in this genome is biased - consistent with the findings from other genomes (282). Intron phase is similarly biased; this is again a pattern found in other genomes (340). Arguments have been advanced that these findings are consequences of the competing 'intron late' or 'intron early' theories: the merits of this debate are beyond the scope of this work.

Intron organisation appears to be conserved between *Plasmodium* species (499, 516). Examination of the splice sites and intron structures in the other *Plasmodium* species may suggest features sufficiently distinct from those of the human host that may be exploitable as drug targets.

10.6 An application of the annotation

Unlike the genome projects of the model organisms – *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and others – which were created to enhance our understanding of how their biology functions, the malaria genome project was created to help solve a serious problem. Culture and genetic manipulation while possible remain difficult when compared with the model organisms and 'conventional' approaches were limited before this project was begun. The knowledge of the sequence of the proteins encoded within the genome is of considerable help in those wishing to try to develop vaccines as the process of cloning can be shortened somewhat.

The genome annotation itself is arguable more useful to those studying the metabolic pathways and identifying potential drug targets. Once a potential pathway has been identified the genes can be cloned, expressed and their crystal structure studied to provide insight into new drug development (124, 333). Alternatively drugs known to be effective against malaria but whose mechanism of killing may not be understood can be explored more easily. Here the enzyme calpain was identified in the genome. This enzyme is known to act on a variety of protein substrates preferentially if not exclusively with a PEST sequence. These sequences may or may not also target the protein towards the proteasome but this latter property is still under investigation.

While calpain inhibitors are known to be active against malaria their killing mechanism is not yet known. It appears that these drugs inhibit merozoite invasion of

the erythrocyte (189) and hence block its replication. Despite these morphological studies the location of calpain within the merozoite remains a matter of speculation. It seems probable that it is located within the micronemes and is released during the invasion process but this has yet to be confirmed. The proteins it acts upon either in the erythrocyte or the merozoite are not yet known with any degree of certainty but the presence of the PEST sequences within a surface exposed protein makes these likely candidates. Here a variety of proteins known to be involved in erythrocyte binding and/or invasion have been identified. Further investigation of this pathway may enhance the potency of these agents.

In additional to these surface exposed proteins a number of proteins – many of which can currently only be identified as 'hypothetical proteins' – posses PEST sequences and are involved in processes other than invasion of the erythrocyte. These other process also may be inhibited by these calpain inhibitors and provide a second drug target. The presence of multiple drug targets for a single agent reduces the likelihood of resistance developing at least in the short term so these other proteins may be worthy of investigation. The action of these inhibitors on the *P. falciparum* calpain which is distinctly different from the human forms of calpain has yet to be investigated.

While the existing inhibitors can be freeze dried and stored at room temperature for a considerable period of time, they have to be administered by a parental route -a method which limits their effectiveness. With the gene sequence now available it should now be possible to clone, express and crystallize the protein with the intention of developing selective oral agents.

10.7 Conclusions

The purpose of this work was to provide the basis for the understanding of the biology of this important organism at the level of its genome. Because the annotation work was shared with Dr. Huestis certain topics that might have formed part of the work here were reserved for his thesis. Specifically these were a comprehensive comparison between the Malaria Genome Project's annotation and the version that was used here and a comparison of the gene rearrangements including the Robertsonian translocations between the various *Plasmodium* species for which there are partial sequences available.

Several of the results here were both novel and unexpected. The presence of isochores in a parasite genome and the association with the constant presence of the erythrocyte membrane protein 1 genes with these regions could not have been anticipated. The deviation from Chargaff's second rule for both the mitochondrion and plastid genomes lead directly to the discovery that Chargaff's second rule holds only for four of the five types of double stranded DNA genome and for no other genome type (316). This lead to the discovery that these deviations may be associated with particular forms of replication and that base location within genomes follow certain laws (314, 342).

Other results were more predictable. Since calpain inhibitors have been known for some time to block the invasion of erythrocytes (353) it was not surprising to find a calpain gene in the genome. Similarly since the association of calpain and PEST sequences has been known for some time, the presence of these motifs within the proteins was anticipated and it is expected that these findings may shed some light on some of the molecular biology of this organism.

While the work here has resulted in the discovery of a number of novel features both in this genome and in others, ultimately the value of a scientific work lies in its usefulness to others. It is hoped that the annotation methods described here, the annotation itself and some of the other results here will prove of use.

173

Appendix

A.1 Types of errors

The presence of errors in genomes while a well know problem does not seem to have yet received the degree of attention that it might perhaps be expected to receive. Zhang *et al* (559) systematically examined a number of published human genes in the databases. After examining the intron structure and gene organization the authors verified a number of their predictions by cloning the genes, reverse transcribing the messenger RNA or comparison with the mouse or rat genome. Evidence was found of base insertion, base deletion, inclusion of non coding sequences into coding sequences, incomplete coding sequences, incorrect stop codons and incorrect initiation codons (Leu instead of Met). Genes were found to be lacking either N or C terminal ends and sometimes both. Additional genes were found in the human genome using cDNA and comparisons with mouse orthologues. Different mistakes were found in copies of the 'same' gene. They noted that these errors had not been noted in the papers recoding the sequencing of the human genome and suggested that computer predictions alone were unlikely to produce accurate annotations.

Other authors have reported problems in genomes from all three branches of life: these include among the eukaryotes the human (220, 268, 350), the mosquito (102, 271), the yeast (103), *Dictyostelium discoidium* (123), *Theileria parva* (434), *Drosophila melanogaster* (178), *P. falciparum* (515) and the chicken (257) genomes; among the bacteria a number of *Nitrobacter* genomes (517) and the *Shigella flexneri* 2a genome (270); and among the arechbacteria the *Pyrococcus furiosus* genome (369). The errors found in these genomes vary but include errors in sequencing and gene identification. The authors agree that (a) computational annotation alone has its limits (b) that this should be supplemented by other data - including transcriptional – when this is available and (c) manual annotation remains a superior but a time consuming method of gene annotation. They appear to have reached these conclusions independently. The remainder of this appendix will illustrate some of the problems with the *P. falciparum* that were discovered in the course of this project.

A.2 Summary of differences between genome versions

The following four tables (Tables A1 - 4) given below were provided by Dr. Robert Huestis. They were created as part of his PhD thesis and have been reproduced here with his permission. The Victoria Bioinformatics Consortium (VBC) is an *ad hoc* group of Australian scientists based in Monash University, Melbourne and in the University of Queensland who are interested in bioinformatics in general and in malaria in particular.

The number of genes given in these tables has risen slightly since the December 2004 version which was used as the basis for the material presented in this thesis. While the genome was published in 2002, there have been a number of revisions to the sequences of chromosomes 6 and 13 in 2004 and to chromosomes 7 and 8 in 2005 with the addition of previously unmapped contigs being added to the chromosomes. There are a number of contigs that are known to be part of the nuclear genome but which have not yet been mapped: once these are mapped the numbers will change again. Further minor variations may also occur later.

Some differences exist between the versions in the PlasmoDB database and GenBank. PlasmoDB is the official database for the *P. falciparum* genome and the annotation there has been subject to a number of revisions which started largely after some of the material in this thesis was presented at the British Society for Parasitology in 2004. The GenBank version to date does not yet seem to have received the same degree of revision as that hosted by the PlasmoDB which explains at least some of the differences found here. To quantify the differences it would be necessary to compare the PlasmoDB and GenBank versions but this has not been done to date.

Table A.1 gives the number of currently annotated (2006) genes in each chromosome, the number of genes listed in GenBank in each chromosome and the percentage agreement between the two annotations. Agreement between the annotations for the nuclear chromosomes lies between 50 and 76%.

	Total VBC	VBC genes	%VBC genes	Total GB	GB genes	%GB genes
	genes	in GB	in GB	genes	in VBC	in VBC
chr 1	154	98	64	150	98	65
chr 2	228	113	50	224	113	50
chr 3	247	187	76	246	187	76

chr 4	257	180	70	257	180	70
chr 5	326	237	73	325	237	73
chr 6	327	228	70	324	228	70
chr 7	332	218	66	305	218	71
chr 8	323	199	62	320	199	62
chr 9	374	243	65	366	243	66
chr 10	404	245	61	404	245	61
chr 11	485	300	62	490	300	61
chr 12	537	380	71	533	380	71
chr 13	709	539	76	718	539	75
chr 14	774	501	65	774	501	65
plastid irA	17	14	82	17	14	82
plastid irB	40	28	70	42	28	67
mitochondrion	21	21	100	22	21	95
unmapped chr 4	7					
unmapped chr 6	27					
unmapped chr 7	7					
unmapped chr 8	2					
unmapped chr 678	59					
unmapped chr 13	25					
partial rRNAs				3		
chr 7						
partial rRNAs chr 8				2		
total genes	5683	3732	67	5522	3732	67

TABLE A.1. Comparison of GenBank MGP genes 2002, 2005 and 2005 with the current annotation of genes. Chromosomes 6 and 13 2004; chromosomes 7 and 8 2005; others as published in 2002. Abbreviations: VBC – Victoria Bioinformatics Consortium, GB - GenBank, Chr – chromosome; ir – inverted repeat; unmapped 678 – contigs that belong to one or more of the chromosomes 6, 7 or 8.

Table A.2 gives the number of annotated exons in each chromosome, the number of exons listed in GenBank for each chromosome and the percentage agreement between

the two annotations. Agreement between the annotations for the nuclear chromosomes lies between 68 and 82%.

	VBC exons	VBC exons	%VBC exons	GB exons	GB exons	%GB exons
		in GB	in GB		in VBC	in VBC
chr 1	370	284	77	376	284	76
chr 2	569	345	61	511	345	68
chr 3	629	546	87	643	546	85
chr 4	629	504	80	611	504	82
chr 5	770	613	80	748	613	82
chr 6	837	658	79	827	658	80
chr 7	768	567	74	698	567	81
chr 8	817	622	76	854	622	73
chr 9	970	733	76	926	733	79
chr 10	1036	654	63	894	654	73
chr 11	1307	825	63	1091	825	76
chr 12	1316	1053	80	1284	1053	82
chr 13	1728	1431	83	1711	1431	84
chr 14	2038	1385	68	1763	1385	79
plastid irA	17	14	82	18	14	78
plastid irB	41	28	68	42	28	67
mitochondrion	21	21	100	22	21	95
unmapped chr4	7					
unmapped chr6	50					
unmapped chr7	21					
unmapped chr8	3					
unmapped chr678	120					
unmapped chr13	40					
partial rRNAs chr7				3		
partial rRNAs chr8				2		
total exons	14103	10284	74	13024	10284	79

TABLE A.2. Comparison of GenBank MGP exons 2002, 2004 and 2005 with the current annotation of exons. Chromosomes 6 and 13 2004; chromosomes 7 and 8 2005; others as published in 2002. Abbreviations: VBC – Victoria Bioinformatics Consortium, GB - GenBank, Chr – chromosome; ir – inverted repeat; unmapped 678 – contigs that belong to one or more of the chromosomes 6, 7 or 8.

Table A.3 gives the number of annotated genes in each chromosome, the number of genes listed in PlasmoDB for each chromosome and the percentage agreement between the two annotations. Agreement between the annotations for the nuclear chromosomes lies between 50 and 82%.

	Total VBC	VBC genes	%VBC genes	PDB	PBD genes	%PDB genes
	Genes	In PDB	in PDB	genes	in VBC	in VBC
chr 1	154	93	60	146	93	64
chr 2	228	113	50	223	113	51
chr 3	247	185	75	243	185	76
chr 4	257	175	68	244	175	72
chr 5	326	234	72	314	234	75
chr 6	327	225	69	312	225	72
chr 7	332	218	66	305	218	71
chr 8	323	199	62	320	199	62
chr 9	374	243	65	367	243	66
chr 10	404	240	59	402	240	60
chr 11	485	300	62	489	300	61
chr 12	537	378	70	526	378	721
chr 13	709	532	75	698	532	76
chr 14	775	407	53	771	407	53
plastid irA	17	14	82	17	14	82
plastid irB	40	28	70	42	28	67
mitochondrion	21	21	100	22	21	95
unmapped chr 4	7					
unmapped chr 6	27					
unmapped chr 7	7					
unmapped chr 8	2					
unmapped chr 678	59					
unmapped chr 13	25					
partial rRNAs chr 7				3		
partial rRNAs chr 8				2		
total genes	5683	3605	67	5446	3605	67

TABLE A.3. Comparison of PlasmoDB 2005 MGP genes with the current annotation of genes. Chromosomes 6 and 13 2004; chromosomes 7 and 8 2005; others as published in 2002. Abbreviations: VBC – Victoria Bioinformatics Consortium, GB -

	VBC	VBC exons	% VBC exons	PDB	PDB exons	%PDB exons
	Exons	in PDB	in PDB	exons	in VBC	in VBC
chr 1	370	281	76	402	281	70
chr 2	569	345	61	510	345	68
chr 3	629	544	86	658	544	83
chr 4	629	499	79	601	499	83
chr 5	770	619	80	755	619	82
chr 6	837	653	78	810	653	81
chr 7	768	567	74	698	567	81
chr 8	817	622	76	854	622	73
chr 9	970	732	75	931	732	79
chr 10	1036	638	62	877	638	73
chr 11	1307	824	63	1088	824	76
chr 12	1316	1050	80	1270	1050	83
chr 13	1728	1435	83	1702	1435	84
chr 14	2037	1173	58	1775	1173	66
plastid irA	17	14	82	18	14	78
plastid irB	41	28	68	42	28	67
mitochondrion	21	21	100	22	21	95
unmapped chr4	7					
unmapped chr6	50					
unmapped chr7	21					
unmapped chr8	3					
unmapped chr678	120					
unmapped chr13	40					
partial rRNAs chr7				3		
partial rRNAs chr8				2		
total exons	14103	10043	74	13018	10043	79

GenBank, Chr – chromosome; ir – inverted repeat; PDB – PlasmoDB database; unmapped 678 – contigs that belong to one or more of the chromosomes 6, 7 or 8.

TABLE A.4. Comparison of PlasmoDB 2005 MGP exons with the current annotation of exons. Chromosomes 6 and 13 2004; chromosomes 7 and 8 2005; others as published in 2002. Abbreviations: VBC – Victoria Bioinformatics Consortium, GB - GenBank, Chr – chromosome; ir – inverted repeat; PDB – PlasmoDB database; unmapped 678 – contigs that belong to one or more of the chromosomes 6, 7 or 8.

Table A.4 gives the number of annotated exons in each chromosome, the number of exons listed in PlasmoDB for each chromosome and the percentage agreement between the two annotations. Agreement between the annotations for the nuclear chromosomes lies between 58 and 83%.

A.3 Probable origin errors in the P. falciparum genome

Several problems were evident in the published *P. falciparum* genome including a number of minor lapses. All the Crick stand exons were incorrectly ordered when the genome was released. When this was pointed out and the exons were reordered within a week. This appears to have been the result of at leasy two separate coding errors. A mismatch between the number of predicted proteins in the coding sequence file and the translated protein file was found. The source of this difference was found to be the method used to predict genes. Some bases had not been identified and were marked with symbols – 'M' (PF10_0399w and PF10_0399w) and 'X' (PF11_0478c, PFL0880c, PFL0880c, PF13_0063c and PF13_0063c). These predicted coding sequences had not been translated and were hence omitted. This difference was not mentioned in the paper.

In addition to these minor problems more systematic errors were found which it seems likely arose as a consequence of the programmes used to prepare the annotation - GlimmerM and Phat. GlimmerM identifies long coding regions without difficulty and provided there is no ambiguity chooses the correct splice sites. It has an unfortunate tendency to overcall introns and in particular it gives great weight to an AG-GT site without reference to other features of the putative intron. This results in a tendency to give rise to unbiological predictions. It tends to call single genes with a long intron as two separate genes, presumably because of its inability to identify the splice sites correctly. In some cases the sensitivity of the intron search process has been increased resulting in artificial 5' exons: a 5' GT site has been created at the first GT within the coding sequence, the coding sequence and a variable amount of the 5' untranslated sequence has been included as an intron. Presumably this arose as a consequence of the optimization of the programme which was set to maximize the gene length. It is not known as a fact whether GlimmerM was indeed set to maximize the coding length but is rather a conjecture based on observation of its output.

A similar error is its choice of incorrect AG sites. Genuine intronic AG site are preceded by a pyrimidine tract of varying length - a rule violated many times by GlimmerM which frequently chooses AGs lacking the required pyrimidine tracts. While the true AG site is always the first after the pyrimidine tract, GlimmerM may also choose the wrong AG. While it is possible that these AG sites are used as alternative splicing sites – the data presently available is insufficient to decide this – all the experimentally confirmed splice sites in *P. falciparum* obey this rule.

On occasion GlimmerM choses introns ending in G AG. While these are theoretically possible 3' splice sites, there is no experimental evidence that this triplet is actually used in *P. falciparum*. At present the reason for its avoidance is not known.

Phat has a tendency to break up both introns and exons into smaller pieces resulting in artificial introns within genes that lack in frame stop codons resulting in a tendency to call a single gene as two distinct genes. This is likely to be the result of the assumption of a geometric distribution of both intron and exon length.

The situation was somewhat complicated as the parameters used by TIGR and Sanger in their programmes were been varied between their publications which accounts for some of the changes between versions in the annotations. While it is sensible to improve the programmes when new data becomes available, this makes the changes in annotations more difficult to analyze.

181

A.4 Examples of errors found

Some of the problems concerning intron recognition will now be illustrated with examples from the published 2002 version of the genome and the probable causes of these problems discussed.

A.4.1 Unbiological introns

Too short: This 'intron' from PFA0085c on Chromosome 1 is 14 bases long.

GTA TGA AGA GAT AG

For physical reasons splicing is impossible. There is no pyrimidine tract at the 3' end. If the intron is included in the reading frame there are two stop codons. It seems likely that this 'intron' was created to bypass these. Overall these finding suggest that the reading frame in this part of the gene is incorrect and that the entire gene should be re-examined.

Atypical content: This 'intron' sequence from PFC0195w (Chromosome 3) is 58 bases long – long enough to be spliced.

GTT TGA CAT ATC TTT TAA GGA AAT GGT GAG CTA GCC AAA AAC AAT GAA ATA AAT CTA G

The GC content of is 31.0% - higher than many of the coding sequences. The pyrimidine content is 39.7% - suggestive of a coding sequence. There is neither a pyrimidine tract nor lariat site. This intron has been miscalled and again it seems likely that this 'intron' was created to avoid the in frame TGA and TAA.

False cryptic introns: Cryptic introns are introns that lack in frame stop codons: these are uncommon and may be miscalled coding sequences. The example here contains no in frame stop codons, has no pyrimidine tract, has no A or T runs, lacks any $(RY)_{5+}$ sequences and has a GC content of 43.1%. This is a coding sequence where the G-GT-AA sequence has been miscalled as a 5' splice site. This problem

may have arisen because of the limited window size of GlimmerM (8 bases or less). This example is from PFD1080w (Chromosome 4)

GTA AGT TGT GCA TCT ATG CAT CTG TGC ATC TGT GCG CCT ACA CAT CTA CAT ATG CAC CTA CGC ACC TAC ATA TTC ATG TAT ACC TTT GGC TAT TTG TTG CAA TTG CAG

Additional sequence: This example is from MAL8P1.107w (hypothetical protein) on chromosome 8. There is an additional codon included at the 3'end. If this intron is correct 'the first AG after the pyrimidine tract 'rule the correct splice site requires the first TAG is the 3' site and not the TAG as given.

A second example from chromosome 9 (PFI0085c – a hypothetical protein) has TAG instead of CAG. TAG is more common as a 3' site: this is not surprising given the AT content of the introns.

GTATATACAT CAATATAATT TAACTCATAA AAGATGTCTA ACAGAACGTA TTT CAG *TAG*

These errors may have occurred because of the algorithm used by GlimmerM begins at the 3' end of a putative intron and moves in the 5' direction. The [T/C] AG here may have been identified as lying within in a good context for a 3' splice and the programme ceased to look further. Since TAG is generally preferred over CAG as a 3' splice site, this may be an example of the local minimum problem. This conjecture is unlikely as this would have been an elementary programming error that should have been avoided by the standard methods. Without examining the code of the programme it is difficult to know how this problem arose.

A.4.2 Misreads

These sequences with probable misreads are particularly difficult for automated annotation. While miscalls in these sequences are foreseeable, the difficulties in creating algorithms to handle them correctly are considerable.

Miscalled bases: The gene annotated as PFA0120c (lysophospholipase) on Chromosome 1 appears to have a misread within a run of As. While the gene appears to be coding along its entire length there is an in frame AAA TAA AAA. The orthologue of this gene in *P. yoelli* (Py2923) has AAA TCT AAA at this point. The sequence in *P. falciparum* is either a miscalled base (the most likely possibility), the gene is becoming a pseudogene in the 3D7 strain - and possibly in other strains of *P. falciparum* - or the TAA codon is part of the regulatory apparatus found in *P. falciparum* and not in *P. yoelli*. Assuming there is a miscalled base it seems most likely that is the in second codon position: TAA should instead be TCA as both of these codons encode serine. The only way to decide this issue seems to be to resequence the gene.

ACT TTA GAA ATA TTA GGA AAA *TAA* AAA GAT CGA

Base miscalls may lead to misidentification of introns. This otherwise very typical intron in the teleomerase gene (tert) on chromosome 13 has a 5' GT, a solid run of As in the left half of the intron, a run of $(RY)_9$ in the center and a solid 3' pyrimidine tract immediately 5' of the putative 3' splice site - which instead of the canonical AG is given as AA. Comparison with the published sequence and its orthologue in *P. yoelli* confirms the initial impression of a sequence miscall.

Frameshifts: In PfL2385c (a hypothetical protein) on chromosome 12, a missing A in a run of three - resulting in a frameshift - can be inferred fairly confidently. Without

the additional base the gene terminates rapidly while with the extra base, the coding region extends in the 3' direction for a considerable distance. The annotator's trilemma presents itself again: is the sequence in error; is this a recent mutation in this strain or species; or is this some novel regulatory device? Knowing that sequences errors occur, while a sequence error seems the most probable option the other possibilities cannot yet be ruled out.

Exon

GGA GA 2025900 T GAA AAT AAT TTA TCG GAT ACA TTA ATA ATG TTA AAA ACA ACA ATA GTT GCT GTT ACT ACA TCT AAA GGA GAA GCT TTG CGT TTG GCT GAT [GAa] ATT AAT ATG ACA TGT AAA ATA TGT CAT AAG AAA AGA AAA GCA ACC AAA TTT TCT ATA

Without the inferred A – here marked with an '[]' the sequence terminates after three codons

GAT GA [] A TTA ATA TGA

A. 4.3 Excessively long introns

Anomalous splicing site identification has resulted in a number of incorrect gene models. Long introns (800-1000 bases) within the *P. falciparum* genome are found exclusively in the *var* genes. Within PFD0665c (26S proteasome ATPase) two separate genes have been spliced together with the help of a large (almost 2000 bases) 'intron.' This sequence has very atypical base use for an intron when compared with that of a genuine intron.

exon GGA TTT TCA ATT AC

intron

GTTACCATT TGAAAAAGGA GCAAATCAAT ATGTTGATGA AAATTTATCT TTTAAATTCC ATTATTTTTT TACAAGAAAA TTTTGGCTAG TTATTTATC TCTAGCATTT ATTATAATGC CAGGAGGATT CGGAACTTTA GAGAACTAA

TGGAAATTCT	ТАСАТТАААА	CAATGTAAAA	GATTTAAAAG	CATGTACCA
ATCATATTAT	TTGGAAAACA	ATTTTGGACA	TCCATTCTTA	ATTTTGTAT
GCTAGCTGAA	TATGGATTAA	TATCTAAAGA	CGATTTAGCT	AGTTTATTA
TTACAGATTC	TATTGAAGAA	GCATATGAAT	GTGTTATCAA	TTTTTGAAA
AATTCAAATC	CTACAACCCC	AAAAGAGATC	AACAATTCTA	GCTAAATAC
CAAAATGTTT	TTCTCACCTT	TCAGCTAGCT	AAAAGTCATA	TTTTTTTTT
TTTTTTTTTT	TTTATAATTT	TATACATATA	TATAACTCAT	AAAAATTAT
AGTTTTACAT	GTTCTCTCTA	TATGTGTGAT	TATTTCACAT	ATAAAAGAA
CAGAACACAT	TAATTTTTTAT	GGGGGTTATA	AACTTATTAT	GATGTTGTT
СТАТТТАТАА	ATCATTAGGG	TATTTCATGT	TAATATTTGA	GTTTTTTTA
ATATGTTTAA	TTTTATGTAT	TTTTTCAATA	TTTGCTTTTT	TGTGTTTTT
TTAATATTTA	ATTTTATAGA	TTTTTAGTTA	GTTGATTTGA	TATTTCAAT
TTGTGTTTGT	TCCATTTTTT	TATATGAAAT	AAAAATGTTT	TAAACATGT
TTTGTTAATT	ATCCATTTTT	CATATGTTCG	TCAAGTAGCC	ААТААСАТА
ААААААТАТ	AAAATTTATA	GTATGAGTTA	СТАТАТАААА	CTTTGTAAT
TTCAATAAGT	TCATTTCACA	ATTCATAAAG	CGATGAATTC	AAACGTTGC
ATTATGCAAT	TATAAGAAGC	TTTTTGTCAA	GTCATATAAT	TTTCCCATG
ACTATTTATT	TCGATTAAGT	GAAATCGTCA	TTTAGGAAAA	TAATAAGTA
GTATTTTTAA	АТАСССАААТ	TATATGTACA	ATATTATATA	TATATCATA
ATTATATATT	GTATATAGGA	AGTATACTTG	CACATACACA	CCAAGCCTT
ACATATGCAT	АААСТАТААС	GGGTTTGTTT	TGTAAAAATT	GGAAAAAAT
ACAAATTTAA	GGATTTGGAA	ATTTTGATAA	TGCAAAAAAA	АААТАТААС
GGTTGAAAAA	TATTTTACTT	ААСАТАААТА	TGCATATATA	GTAAGATCC
TTACATAACC	ATGTTTATAT	GTTAAAACAT	ТТААСТААТТ	AAAATATTT
AAGATATTTA	AAATTCTGGA	TAAAGTGGAA	AAATGTATTAA	AATTTGTAT
GCAACATTTG	TGTATATATT	TTATGTATTG	CATTTTTTTT	TTTTTTTGC
ATTCATATAC	TTTATTATAA	TTATTTTCCT	TTTCTCTTCA	GAAAAATAT
TTTATATACA	ATAATATTCT	ААТАСАТААА	ТАААТААТАА	ATAAATATT
ATATGGCATC	ATGCATAAAT	ATATACATTT	TTGCGAAACT	TTTCAACAA
САТАТААТАА	ААААААААТ	ATATCATAAT	AAGTGAAAGC	TACGAAAAT
ATCTATAATT	AAATTTTTGT	TTGTTTTTTT	TTTTTCCATA	ААААТААСА
TGAGAGTAAT	TTTTTAAATT	AATCAAAAAG	ААААСААТАТ	GGAAAAATG
GAAAATGTTA	GTAAGTACCT	TAAGGAGGAA	GATTATTATA	TAAAATGAA
AATTCTGAAG	AAGCAACTTG	ATATTTTAAA	TATTCAGGTG	AGACAAAAT

AACAATTTCC TTTTCATTTT TCCATATCTA TATGTATATG TGTGTAGGA ATGCATATCC ATAGTATGCT ATTACAAAAG TAGTATATTT TTATCGATT ATGTGTTTTT ATTATATTTA CATATAATTC TGATTACGAT CCATGCCAT AATATGTTTT TAAAATTGTA TTTGAATATA AG

exon

A TAC ATT GAA AAT ATG GA

A.4.4 Inconsistent sequences

A number of sequences are not consistent between the previously released contigs and the final draft. All the sequencing centers suffered from this problem to some extent. The first example here is from a gene on chromosome 6 (PFF0085w). The original contig sequence (c_6) and the final release are shown. While the remainder of the gene sequence agrees between the versions, there are 10 differences in 36 bases including a new stop codon. The final version may well be the correct one but with 27.8% of the bases being changed, this does not inspire confidence in the quality of the sequencing and subsequent assembly.

c6_0064 ta ttt aca gaa tgg atg aaa gag ctc caa gga tta
gaa aag taa*

PFF0085w GG TTT ACA GAA AGG ATG AAT AAT CCA AAA *TGA* TTA GAT ATG TAA*

This second example is from chromosome 12 (PfL1565c – a hypothetical protein). The 2001 release is in lower case and the final draft in upper case. In the final draft an additional base and a transversion in the second codon position has been added. While the additional base appears to be correct, it is not possible to comment on the transversion.

24oct01: t aaa tta gtt aat tta ata aaa -ac aaa PfL1565c: Т AAA TTA AAA GGT TAC AAT AAA TTA ATA

A.4.5 Misassembly

Misassembly is difficult both to identify or to prove because not many uniquely identifiable proteins have been cloned that span two contigs. It does not seem likely that misassembly is a major problem as most of the genes cloned since the MGP reported tend to agree with their reported assembly. Nonetheless an unequivocal example was found in chromosome 13.

The normocyte binding protein 2b gene was cloned two years before the publication of the *P. falciparum* genome. (3) Examination of chromosome 13 shows part of the gene is located there but the sequence changes completely within the middle of an exon. It seems that this gene which could have been used as a scaffold for the assembly was never consulted. As additional genes are found further examples may be discovered.

A.5 Discussion

The material above is not an exhaustive listing of the problems in the genome annotation but includes examples chosen to illustrate the types of the problems discovered. Some of the problems are based on experimental errors: sequencing is known to be difficult where the genome is either exceptionally rich or poor in AT content and this may lead to 'artificial' indel mutations and misannotations. It is clear from the disparity between the number of predicted genes and the number of translated proteins that the authors were aware of these problems. This being so it was curious that these caveats were not mentioned in the papers. The paper by Zhang *et al* (559) commented on the absence of such caveats in the human genome.

Given that the authors appear to have been aware of the sequencing problems their reliance on programmatic means to identify genes is somewhat surprising. Writing programmes for genome annotation is difficult. To allow for the possibility that the bases might not be correctly identified or that the sequence might be missing bases or non existent ones might be included would make gene prediction by computer currently impossible because the number of possibilities increase exponentially with the length of the sequence. Accordingly all gene prediction programmes make the assumption that the sequence input is error free. When these programmes are used on 'real' (error prone) sequence data the true prediction rate is reduced. The extent of this

reduction is difficult to predict as it is dependent on the quality of the sequence data. Even if the gene prediction programmes were capable of perfect accuracy the gene predictions would still need review because of this problem. Judging from the published data, it does not appear to be the case that such reviews are done either by the authors or the reviewers so caution seems indicated in relying on this data.

Even with 'manual' review indels and sequence errors are difficult to be identify. Two methods were used here to identify them here: firstly by alignment with gene sequences either cloned from the same organism or a similar gene. The second method was by inspection. Genuine pseudogenes have stop codons in all three frames after an indel mutation. Frequently much of the gene either 5' or 3' of the mutation has been deleted. Where there is an artifactual indel mutation, the coding sequence 3'of the indel is usually recognizable and the indel can be corrected by hand. Similar strategies were employed to determine possible sequence errors. These methods require practice and are currently very difficult to programme. Zhang *et al* appear to have used similar methods.

Huestis and Fisher (210) like Zhang *et al* noted that there examples of exons were missing in the both in the five and three prime ends of genes. Both papers described similar methods of identifying these: in both the sequence has to be examined for the presence of coding sequences that can be spliced in frame into the known coding sequence. In the case of *P. falciparum* the exon will be located within 400-600 bases of the remainder of the gene but in the human genome the search may need be more extensive because of the greater size of the introns. Computionally this appears likely to be a problem with a local optimum: this may be correctable if the weighings used in the programmes are changed.

The examples given above concentrate on the introns. Finding the initial and final exons of any gene may be difficult: inclusion of the possibility of indels and sequence errors is probably beyond the capabilities of current computers. In constrast once the coding region is identified the annotation programmes should be capable of identifying these correctly.

The presence of very short introns in the annotation should have given rise to concern to the authors. A miminal intron has a 5' splice site (GT + 2-3 bases), a lariat site (5-7 bases), a TC tract (5-6 bases) and a 3' AG site – total of 16-20 bases. The intron from PFA0085c is 14 bases long and aside from a GT and AG site lacks the other expected features of an intron. What appears to have happened here is that the programme

noticed an in frame TGA. Instead of rechecking the frame it decided that the GT site was reasonable (GT AT GAA GAG) and that there was a potential 3' T AG close by that would enable to coding frame to be extended. This determination does not appear to take into consideration the absence of a TC tract or the length of the intron. It seems likely that the programme has been tasked to maximize the length of the coding sequence.

This impression is reinforced by the presence of introns with atypical base content. The purine content within coding sequences is greater than their pyrimidine content with the conserve being true for introns and intergenic regions. Also the GC content of introns – at least in this genome – tend to be lower than that of the coding sequences. The example given above suggest that this 'intron' was created to avoid an in frame stop codon. The sequence has the typical appearance of being coding with the high GC and purine content which suggests that the frame chosen of this sequence is incorrect. The papers were not clear on how the reading frame is chosen by the programmes but it seems likely that the frame is chosen to maximize the length of the coding sequence and that introns are created to allow for its maximization.

The false cryptic intron shown above appears to be a false intron created by Phat. Phat is constrained to ensure that the predicted introns and exons follow an arbitrary length distribution. This Phat does by artificially truncating introns and exons as appears to have happened here.

The occasional problem of excessive long introns is difficult to explain. This is unlikely to be due to Phat; nor is it typical of GlimmerM. As these introns are easily recognized because of their length it is very surprising that these were not re examined.

In sum there were a variety of problems evident in the published genome. Many of these relate to sequence errors – misreads, indels, misassembly - which no existing programme can handle and which can be difficult even for experienced annotators. A second class involved missing exons at either end of the genes. This too can be a problem even for human annotators but it may be a correctable problem if the search space for coding sequence is extended in both directions. A more serious problem was the misidentification of introns. This latter problem seems to have arisen from a number of source: in part from an attempt maximize the coding length of the 'gene' regardless of the introns created; in part from the arbitrary intron and length distributions programmed in to Phat; and in part to causes that are not yet understood.

References

- 1. Adam, I., M. E. Osman, G. Elghzali, G. I. Ahmed, L. L. Gustafssons, and M. I. Elbashir. 2004. Efficacies of chloroquine, sulfadoxine-pyrimethamine and quinine in the treatment of uncomplicated, *Plasmodium falciparum* malaria in eastern Sudan. Ann. Trop. Med. Parasitol. **98:**661-666.
- 2. Adebajo, A. O., D. J. Smith, B. L. Hazleman, and T. G. Wreghitt. 1994. Seroepidemiological associations between tuberculosis, malaria, hepatitis B, and AIDS in West Africa. J. Med. Virol. 42:366-368.
- 3. Afanasiev, V. I. 1879. On the pathology of malarial infection. Protocol of Meeting of Russian Physicians in St. Petersburg 10:380.
- 4. Ahlborg, N., R. Andersson, S. Stahl, M. Hansson, I. Andersson, P. Perlmann, and K Berzins. 1994. B- and T-cell responses in congenic mice to repeat sequences of the malaria antigen Pf332: effects of the number of repeats. Immunol. Lett. 40:147-155.
- 5. Aikawa, M., L. H. Miller, J. Johnson, and J. Rabbege. 1978. Erythrocyte entry by malarial parasites. A moving junction between erythrocyte and parasite. J. Cell Biol. 77: 72-82.
- 6. Aiyar, S. E., R. L. Gourse, and W. Ross. 1998. Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase alpha subunit. Proc. Natl. Acad. Sci. U S A **95**:14652 14657.
- Ajaiyeoba, E. O., C. O. Falade, O. I. Fawole, D. O. Akinboye, G. O. Gbotosho, O. M. Bolaji, J. S. Ashidi, O. O. Abiodun, O. S. Osowole, O. A. Itiola, O. Oladepo, A. Sowunmi, and A. M. Oduola. 2004. Efficacy of herbal remedies used by herbalists in Oyo State Nigeria for treatment of *Plasmodium falciparum* infections - a survey and an observation. Afr. J. Med. Med. Sci. 33:115-119.
- Akompong, T., J. VanWye, N. Ghori and K. Haldar. 1999. Artemisinin and its derivatives are transported by a vacuolar-network of *Plasmodium falciparum* and their anti-malarial activities are additive with toxic sphingolipid analogues that block the network. Mol. Biochem. Parasitol. 101:71-79.
- Alaii, J. A., H. W. van den Borne, S. P. Kachur, H. Mwenesi, J. M. Vulule, W. A. Hawley, M. I. Meltzer, B. L. Nahlen, and P.A. Phillips-Howard. 2003. Perceptions of bed nets and malaria prevention before and after a randomized controlled trial of permethrin-treated bed nets in western Kenya. Am. J. Trop. Med. Hyg. 68:142-148.
- Alonso, P. L., J. Sacarlal, J. J. Aponte, A. Leach, E. Macete, J. Milman, I. Mandomando, B. Spiessens, C. Guinovart, M. Espasa, Q. Bassat, P. Aide, O. Ofori-Anyinam, M. M. Navia, S. Corachan, M. Ceuppens, M. C. Dubois, Demoitie MA, F. Dubovsky, C. Menendez, N. Tornieporth, W. R. Ballou, R. Thompson, and J. Cohen. 2004. Efficacy of the RTS,S/AS02A vaccine against *Plasmodium falciparum* infection and disease in young African children: randomised controlled trial. Lancet 364:1411-1420.
- 11. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.
- 12. **Ambroise-Thomas, P.** 2004. Genomic, molecular biology and malaria: new medical perspectives? Bull. Soc. Pathol. Exot. **97:**155-160.

- Aminov, R. I., J. C. Chee-Sanford, N. Garrigues, A. Mehboob, and R. I. Mackie. 2004. Detection of tetracycline resistance genes by PCR methods. Methods Mol. Biol. 268:3-13.
- 14. Andersson, G. E., and P. M. Sharp. 1996. Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology 142:915-925.
- 15. Andersson, S. G., and P. M. Sharp. 1996. Codon usage and base composition in *Rickettsia prowazekii*. J. Mol. Evol. **42**:525-536.
- 16. Andersson, S. G., and P. M. Sharp. 1996. Codon usage and base composition in *Rickettsia prowazekii*. Mol. Evol. **42**:525-536.
- Anyanwu, E. C., J. E. Ehiri, I. Kanu, M. Morad, S. Ventegodt, and J. Merrick. 2004. Assessing the health effects of long-term exposure to insecticide-treated mosquito nets in the control of malaria in endemic regions. ScientificWorldJournal 4:978-988.
- Arnott, S., and E. Selsing. 1974. Structures for the polynucleotide complexes poly(dA) with poly (dT) and poly(dT) with poly(dA) with poly (dT). J. Mol. Biol. 88:509 - 521.
- Baird, J. K., T. R. Jones, E. W. Danudirgo, B. A. Annis, M. J. Bangs, H. Basri, Purnomo, and S. Masbar. 1991. Age-dependent acquired protection against *Plasmodium falciparum* in people having two years exposure to hyperendemic malaria. Am. J. Trop. Med. Hyg. 45:65-76.
- Ballou, W. R., M. Arevalo-Herrera, D. Carucci, T. L. Richie, G. Corradin, C. Diggs, P. Druilhe, B. K. Giersing, A. Saul, D. G. Heppner, K. E. Kester, D. E. Lanar, J. Lyon, A. V. Hill, W. Pan, and J. D. Cohen. 2004. Update on the clinical development of candidate malaria vaccines. Am. J. Trop. Med. Hyg. 71:239-247.
- Banerjee, T., S. Basak, S. K. Gupta, and T. C. Ghosh. 2004. Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. J. Biomol. Struct. Dyn. 22:13-24.
- Bannister L.H., a. G. H. M. 1989. The fine structure of secretion by *Plasmodium knowlesi* merozoites during red cell invasion. J. Protozool. 36:362-367.
- Bannister, L. H., J. M. Hopkins, R. E. Fowler, S. Krishna, and G. H. Mitchell. 2000. A brief illustrated guide to the ultrastructure of *Plasmodium falciparum* asexual blood stages. Parasitol. Today 16:427-433.
- 24. Bao, W., W. Zhang, Q. Yang, Y. Zhang, B. Han, M. Gu, Y. Xue, and Z. Cheng. 2006. Diversity of centromeric repeats in two closely related wild rice species, *Oryza officinalis* and *Oryza rhizomatis*. Mol. Genet. Genomics 275:421-430.
- 25. Baruch, D. I., B. L. Pasloske, H. B. Singh, X. Bi, X. C. Ma, M. Feldman, T. F. Taraschi, and R. J. Howard. 1995. Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82:77-87.
- 26. **Bastien, O., J. C. Aude, S. Roy, and E. Marechal.** 2004. Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics. Bioinformatics **20**:534-537.
- Beier, M. S., J. C Beier, A. A. Merdan, B. M. el Sawaf, and M. A. Kadder. 1987. Laboratory rearing techniques and adult life table parameters for *Anopheles sergentii* from Egypt. J. Am. Mosq. Control Assoc. 3:266-270.
- 28. Bejon, P., L. Andrews, R. F. Andersen, S. Dunachie, D. Webster, M. Walther, S. C. Gilbert, T. Peto, and A. V. Hill. 2005. Calculation of liver-

to-blood inocula, parasite growth rates, and preerythrocytic vaccine efficacy, from serial quantitative polymerase chain reaction studies of volunteers challenged with malaria sporozoites. J. Infect. Dis. **191:**619-626.

- 29. Bemelman, F., K. De Blok, P. De Vries, S. Surachno, I. and Ten Berge. 2004. Falciparum malaria transmitted by a thick blood smear negative kidney donor. Scand. J. Infect. Dis. **36:**769-771.
- Bennetzen, J. L., C. Coleman, R. Liu, J. Ma, and W. Ramakrishna. 2004. Consistent over-estimation of gene number in complex plant genomes. Curr. Opin. Plant Biol. 7:732-736.
- 31. Bensch, S., J. Perez-Tris, J. Waldenstrom, and O. Hellgren. 2004. Linkage between nuclear and mitochondrial DNA sequences in avian malaria parasites: multiple cases of cryptic speciation? Evolution Int. J. Org. Evolution 58:1617-1621.
- 32. Bergmann-Leitner, E. S., S. Scheiblhofer, R. Weiss, E. H. Duncan, W. W. Leitner, D. Chen, E. Angov, F. Khan, J. L. Williams, D. B. Winter, J. Thalhamer, J. A. Lyon, and G. C. Tsokos. 2005. C3d binding to the circumsporozoite protein carboxy-terminus deviates immunity against malaria. Int. Immunol. 17:245-255.
- 33. **Bernardi, G.** 2000. Isochores and the evolutionary genomics of vertebrates. Gene **241:**3 -17.
- Berry, A. E., M. J. Gardner, G. J. Caspers, D. S. Roos, and M. Berriman. 2004. Curation of the *Plasmodium falciparum* genome. Trends Parasitol. 20:548 - 552.
- Berry, A. E., M. J. Gardner, G. J. Caspers, D. S. Roos, and M. Berriman. 2004. Curation of the *Plasmodium falciparum* genome. Trends Parasitol. 20:548-552.
- Bertram, G., S. Innes, O. Minella, J. Richardson, and I. Stansfield. 2001. Endless possibilities: translation termination and stop codon recognition. Microbiology 147:255-269.
- 37. Bhisutthibhan, J., X. Q. Pan, P. A. Hossler, D. J. Walker, C. A. Yowell, J. Carlton, J. B. Dame, and S. R. Meshnick. 1998. The *Plasmodium falciparum* translationally controlled tumor protein homolog and its reaction with the antimalarial drug artemisinin. J. Biol. Chem. 273:16192-16198.
- 38. **Birkemeyer, C., A. Luedemann, C. Wagner, A. Erban, and J. Kopka.** 2005. Metabolome analysis: the potential of *in vivo* labeling with stable isotopes for metabolite profiling. Trends Biotechnol. **23**:28-33.
- 39. Bischoff, E., M. Guillotte, O. Mercereau-Puijalon, and S. Bonnefoy. 2000. A member of the *Plasmodium falciparum* Pf60 multigene family codes for a nuclear protein expressed by readthrough of an internal stop codon. Mol. Microbiol. 35:1005-1016.
- 40. **Bjerknes, J.** 1969. Atmospheric teleconnections from the Equatorial Pacific. Mon. Weather Rev. **97:**163-172.
- 41. **Blackman, M. J.** 2000. Proteases involved in erythrocyte invasion by the malaria parasite: function and potential as chemotherapeutic targets. Curr. Drug Targets 1:59-83.
- 42. **Blount, R. E.** 1967. Management of chloroquine resistant falciparum malaria. Trans. Am. Clin. Climatol. Assoc. **78**:196-204.
- 43. Bojang, K. A., P. J. M. Milligan, M. Pinder, L. Vigneron, A. Alloueche, K. E. Kesterd, W. R. Balloud, D. J. Conway, W. H. H. Reece, P. Gothard, L. Yamuah, M. Delchambre, G. Voss, B. M. Greenwood, A. Hill, McAdama,

K. P. W. J. Nadia Tornieporth, J. D. Cohen, T. Doherty and RTS, S Malaria Vaccine Trial Team. 2001. Efficacy of RTS, S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial Lancet **358**:927-1934.

- Bojang, K. A., P. J. Milligan, M. Pinder et al. 2001. Efficacy of RTS,S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial. Lancet 358:1927-1934.
- Bond, J. G., J. C. Rojas, J. I. Arredondo-Jimenez, H. Quiroz-Martinez, J. Valle, and T. Williams. 2004. Population control of the malaria vector *Anopheles pseudopunctipennis* by habitat manipulation. Proc. R. Soc. Lond. B Biol. Sci. 271:2161-2169.
- 46. **Bonfils, C., P. Gaudet, and A. Tsang.** 1999. Identification of *cis*-regulating elements and *trans*-acting factors regulating the expression of the gene encoding the small subunit of ribonucleotide reductase in *Dictyostelium discoideum* J. Biol. Chem. **274:**20384 20390.
- 47. Borrmann, S., S. Issifou, G. Esser, A. A. Adegnika, M. Ramharter, P. B. Matsiegui, S. Oyakhirome, D. P. Mawili-Mboumba, MA Missinou, J. F. Kun, H. Jomaa, and P. G. Kremsner. 2004. Fosmidomycin-Clindamycin for the treatment of *Plasmodium falciparum* malaria. J. Infect. Dis. 190:1534-1540.
- 48. **Bottius, E., N. Bakhsis, and A. Scherf.** 1998. *Plasmodium falciparum* telomerase: *de novo* telomere addition to telomeric and nontelomeric sequences and role in chromosome healing. Mol. Cell. Biol. **18**:919 925.
- Boudjelal, M., Z. Wang, J. J. Voorhees, and G. J. Fisher. 2000. Ubiquitin/proteasome pathway regulates levels of retinoic acid receptor gamma and retinoid X receptor alpha in human keratinocytes. Cancer Res. 60:2247-2252.
- Bouma, M. J. 2003. Methodological problems and amendments to demonstrate effects of temperature on the epidemiology of malaria. A new perspective on the highland epidemics in Madagascar, 1972-89. Trans. R. Soc. Trop. Med. Hyg. 97:133-139.
- 51. Bouma, M. J., and C. Dye. 1997. Cycles of malaria associated with *El Nino* in Venezuela. JAMA 278:1772-1774.
- 52. **Bouma, M. J., and H. J. van der Kaay.** 1996. The *El Nino* Southern Oscillation and the historic malaria epidemics on the Indian subcontinent and Sri Lanka: an early warning system for future epidemics? Trop. Med. Int. Health **1**:86-96.
- 53. Bozdech, Z., J. Zhu, M. P. Joachimiak, F. E. Cohen, B. Pulliam, and J. L. DeRisi. 2003. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. Genome Biol. 4:R9.
- 54. Bradley-Moore, A. M., B. M. Greenwood, A. K. Bradley, A. Bartlett, D. E. Bidwell, A. Voller, J. Craske, B. R. Kirkwood, and H. M. Gilles. 1985. Malaria chemoprophylaxis with chloroquine in young Nigerian children. II. Effect on the immune response to vaccination. Ann. Trop. Med. Parasitol. 79:563-573.
- 55. Breman, J. G., M. S. Alilio, and A. Mills. 2004. Conquering the intolerable burden of malaria: what's new, what's needed: a summary. Am. J. Trop. Med. Hyg. 71:1 15.

- 56. Brook, J. H., C. A. Genese, P. B. Bloland, J. R. Zucker, and K. C. Spitalny. 1994. Brief report: malaria probably locally acquired in New Jersey. N. Engl. J. Med. 331:22-23.
- 57. Bruna-Romero, O., and A. Rodriguez. 2001. Dendritic cells can initiate protective immune responses against malaria. Infect. Immun. 69:5173-5176.
- Bruneel, F., M. Thellier, O. Eloy, D. Mazier, G. Boulard, M. Danis, and J. P. Bedos. 2004. Transfusion-transmitted malaria. Intensive Care Med. 30:1851-1852.
- 59. Brustoski, K., U. Moller, M. Kramer, A. Petelski, S. Brenner, D. R. Palmer, M. Bongartz, P. G. Kremsner, A. J. Luty, and U. Krzych. 2005. IFN-gamma and IL-10 mediate parasite-specific immune responses of cord blood cells induced by pregnancy-associated *Plasmodium falciparum* malaria. J. Immunol. 174:1738-1745.
- Buckin, V. A., B. I. Kankiya, N. V. Bulichov, A. V. Lebedev, I. Ya.
 Gukovsky, V. P. Chuprina, A. P. Sarvazyan, and A. R. Williams. 1989.
 Measurement of anomalously high hydration of (dA)n.(dT)n double helices in dilute solution. Nature 340:321 322.
- 61. **Bunn, A., R. Escombe, M. Armstrong, C. J. Whitty, and J. F. Doherty.** 2004. Falciparum malaria in malaria-naive travellers and African visitors. QJM. **97:**645-649.
- 62. Burge, C., and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268:78–94.
- 63. **Burns, J. M. J., P. R. Flaherty, P. Nanavati, and W. P. Weidanz.** 2004. Protection against *Plasmodium chabaudi* malaria induced by immunization with apical membrane antigen 1 and merozoite surface protein 1 in the absence of gamma interferon or interleukin-4. Infect. Immun. **72:**5605-5612.
- 64. **Burset, M., and R. Guigo.** 1996. Evaluation of gene structure prediction programs. Genomics **34**:353-367.
- 65. **Butler, D.** 2003. Mosquito production mooted as fast track to malaria vaccine. Nature **425**:437.
- 66. **Calabrese, P. P., R. T. Durrett, and C. F. Aquadro.** 2001. Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. Genetics **159:**839 952.
- 67. Caldas de Castro, M., Y. Yamagata, D. Mtasiwa, M. Tanner, J. Utzinger, J. Keiser, and B. H. Singer. 2004. Integrated urban malaria control: a case study in Dar es Salaam, Tanzania. Am. J. Trop. Med. Hyg. 71:103-117.
- Cancilla, M. R., A. J. Hillier and B. E. Davidson. 1995. Lactococcus lactis glyceraldehyde-3-phosphate dehydrogenase gene, gap - further evidence for strongly biased codon usage in glycolytic pathway genes. Microbiology 141:1027-1036.
- 69. Cardoso, R. F., R. C. Cooksey, G. P. Morlock, P. Barco, L. Cecon, F. Forestiero, C. Q. Leite, D. N. Sato, L. Shikama Mde, E. M. Mamizuka, R. D. Hirata, and M. H. Hirata. 2004. Screening and characterization of mutations in isoniazid-resistant *Mycobacterium tuberculosis* isolates obtained in Brazil. Antimicrob. Agents Chemother. 48:3373-3381.
- 70. **Carlton, J. M., R. Muller, C. A. Yowell**, *et al.* 2001. Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. Mol. Biochem. Parasitol. **118**:201-210.

- 71. **Carlton, J. M., S. V. Angiuoli, B. B. Suh** *et al.* 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature **419:**512-519.
- 72. Carruthers, V. B., and M. J. Blackman. 2005. A new release on life: emerging concepts in proteolysis and parasite invasion. Mol. Microbiol 55:1617-1630.
- 73. Carter, J. A., A. J. Ross, B. G. Neville, E. Obiero, K. Katana, V. Mung'ala-Odera, JA Lees, and C. R. Newton 2005. Developmental impairments following severe falciparum malaria in children. Trop. Med. Int. Health 10:3 - 10.
- 74. Carter, R., and K. N. Mendis. 2002. Evolutionary and historical aspects of the burden of malaria. Clin. Microbiol. Rev. 15:564 594.
- 75. Casey, G. J., M. Ginny, M. Uranoli, I. Mueller, J. C. Reeder, B. Genton, and A. F. Cowman. 2004. Molecular analysis of *Plasmodium falciparum* from drug treatment failure patients in Papua New Guinea. Am. J. Trop. Med. Hyg. 70:251-255.
- 76. **Cattell, R. B.** 1966. The meaning and strategic use of factor analysis: In Handbook of multivariate experimental psychology. Rand McNally, Chiago.
- 77. Cawley, S. E., A. I. Wirth, and T. P. Speed. 2001. Phat a gene finding program for *Plasmodium falciparum*. Mol. Biochem. Parasitol. **118**:167-174.
- 78. Charlesworth, B., and P. Sniegowski, and W.Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215 220.
- 79. **Chastel, C.** 2004. When the Egyptian mummies are speaking about the infections that have made them ill. Hist. Sci. Med. **38**:147-155.
- 80. Chen, J., M. Sun, W. J. Kent, X. Huang, H. Xie, W. Wang, G. Zhou, R. Z. Shi, and J. D. Rowley. 2004. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res. 32:4812-4820.
- 81. Chen, J. Y., C. P. Fung, F. Y. Chang, L. Y. Huang, J. C. Chang, and L. K. Siu. 2004. Mutations of the *rpoB* gene in rifampicin-resistant *Streptococcus pneumoniae* in Taiwan. J. Antimicrob. Chemother. **53**:375-378.
- 82. Chen, L., F. Y. Qu, and Y. C. Zhou. 1982. Field observations on the antimalarial piperaquine. Chin. Med. J. 95:281-286.
- Chevaillier, P. 1993. Pest sequences in nuclear proteins. Int. J. Biochem. 25:479-482.
- 84. Clark, R. L., T. E. White, S. A. Clode, I. Gaunt, P. Winstanley, and S. A. Ward. 2004. Developmental toxicity of artesunate and an artesunate combination in the rat and rabbit. Birth Defects Res. B 71:380-394.
- 85. Clay, O. 2001. Standard deviations and correlations of GC levels in DNA sequences. Gene 276:33 38.
- Coban, C., K. J. Ishii, D. J. Sullivan, and N. Kumar. 2002. Purified malaria pigment (hemozoin) enhances dendritic cell maturation and modulates the isotype of antibodies induced by a DNA vaccine. Infect. Immun. 70:3939-3943.
- Coban, C., K. J. Ishii, T. Kawai, H. Hemmi, S. Sato, S. Uematsu, M. Yamamoto, O. Takeuchi, S. Itagaki, N. Kumar, T. Horii, and S. Akira. 2005. Toll-like receptor 9 mediates innate immune activation by the malaria pigment hemozoin. J. Exp. Med. 201:19-25.
- 88. **Codd, E.** 1969. Derivability, redundancy, and consistency of relations stored in large data banks. IBM Research Report **RJ599**.

- Coetzee, M., M. Craig, and D. le Sueur. 2000. Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. Parasitol. Today 16:74-77.
- 90. Collins, L. J., A. M. Poole, and D. Penny. 2003. Using ancestral sequences to uncover potential gene homologues. Appl. Bioinformatics 2:S85-95.
- Corcoran L. M., J. K. T., D. Walliker, and D. J. Kemp. 1988. Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in *P. falciparum*. Cell 53:807-813.
- 92. Cosgriff, T. M., C. L. Pamplin, C. J. Canfield, and G. P. Willet. 1985. Mefloquine failure in a case of falciparum malaria induced with a multidrugresistant isolate in a non-immune subject. Am. J. Trop. Med. Hyg. 34:692-693.
- 93. Cosgriff, T. M., E. F. Boudreau, CL Pamplin, E. B. Doberstyn, R. E. Desjardins, and C. J. Canfield. 1982. Evaluation of the antimalarial activity of the phenanthrenemethanol halofantrine (WR 171,669). Am. J. Trop. Med. Hyg. 31:1075-1079.
- 94. **Cox, F. E.** 2002. History of human parasitology. Clin. Microbiol. Rev. **15**:595 612.
- 95. Craig, M. H., I. Kleinschmidt, J. B. Nawn, D. Le Sueur, and B. L. Sharp. 2004. Exploring 30 years of malaria case data in KwaZulu-Natal, South Africa: Part I. The impact of climatic factors. Trop. Med. Int. Health 9:1247-1257.
- 96. Creasey A, B. F., A. Walker, S. Thaithong, S. Oliveira, S. Mutambu, and D. Walliker. 1990. Genetic diversity of *Plasmodium falciparum* shows geographical variation. Am. J. Trop. Med. Hyg. 42:403-413.
- 97. **D'Onofrio, G., and G. Bernardi.** 1992. A universal compositional correlation among codon positions. Gene **110:**81-88.
- 98. **D'Onofrio, G., K. Jabbari, H. Musto, and G. Bernardi.** 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene **238:**3-14.
- 99. **Daglar, O.** 2004. Health situation of the armies in the Crimean war and a document related to this. Tip Tarihi Arastirmalari **12:**41 52.
- 100. Dahl, E. L., J. L. Shock, B. R. Shenai, J. Gut, J. L. DeRisi, and P. J. Rosenthal. 2006. Tetracyclines specifically target the apicoplast of the malaria parasite *Plasmodium falciparum*. Antimicrob. Agents Chemother. 50:3124-3131.
- 101. Dame, J. B., D. E. Arnot, P. F. Bourke, et al. 1996. Current status of the *Plasmodium falciparum* genome project. Mol. Biochem. Parasitol. **79:1-12**.
- 102. Dana, A. N., M. E. Hillenmeyer, N. F. Lobo, M. K. Kern, P. A. Romans, and F. H. Collins. 2006. Differential gene expression in abdomens of the malaria vector mosquito, *Anopheles gambiae*, after sugar feeding, blood feeding and *Plasmodium berghei* infection. BMC Genomics 7:119.
- 103. David, L., W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. 2006. A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. USA 103:5320-5325.
- 104. Davis, T. M., T. Y. Hung, I. K. Sim, H. A. Karunajeewa, and K. F. Ilett. 2005. Piperaquine: a resurgent antimalarial drug. Drugs 65:75-87.
- 105. De Bolle, X., C. D. Bayliss, D. Field, T. van de Ven, N. J. Saunders, D. W. Hood, and E. R. Moxon 2000. The length of a tetranucleotide repeat tract in

Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol. Microbiol. **35:**211 - 222.

- 106. de Miranda, A. B., F. Alvarez-Valin, K. Jabbari, W. M. Degrave, and G. Bernardi. 2000. Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J. Mol. Evol. 50:45-55.
- 107. **de Zulueta**, J. 1994. Malaria and ecosystems: from prehistory to posteradication. Parassitologia 36:7-15.
- 108. Dechering, K. J., K. Cuelenaere, R. N. Konings, and J. A. Leunissen 1998. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. Nucleic Acids Res. 26:4056 - 4062.
- 109. Denis, M. B., T. M. Davis, S. Hewitt, S. Incardona, K. Nimol, T. Fandeur, Y. Poravuth, C. Lim, and D. Socheat. 2002. Efficacy and safety of dihydroartemisinin-piperaquine (Artekin) in Cambodian children and adults with uncomplicated falciparum malaria. Clin. Infect. Dis. 35:1469-1476.
- 110. Deparis, X., B. Frere, M. Lamizana, R. N'Guessan, F. Leroux, P. Lefevre, L. Finot, J. M. Hougard, P. Carnevale, P. Gillet, and D. Baudon. 2004. Efficacy of permethrin-treated uniforms in combination with DEET topical repellent for protection of French military troops in Cote d'Ivoire. J. Med. Entomol. 41:914-921.
- 111. **Deshpande, P., and P. Shastry.** 2004. Modulation of cytokine profiles by malaria pigment hemozoin: role of IL-10 in suppression of proliferative responses of mitogen stimulated human PBMC. Cytokine **28**:205-213.
- 112. **Deutsch, M., and M. Long.** 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. **27**:3219-3228.
- 113. Dicko, A., C. Mantel, M. A. Thera, S. Doumbia, M. Diallo, M. Diakite, I. Sagara, and O. K. Doumbo. 2003. Risk factors for malaria infection and anemia for pregnant women in the Sahel area of Bandiagara, Mali. Acta Trop. 89:17-23.
- 114. Dike. S., V. S. B., L. U. Nascimento, Z. Xuan, J. Ou, T. Zutavern, L. E. Palmer, G. Hannon, M. Q. Zhang, and W. R. McCombie. 2004. The mouse genome: Experimental examination of gene predictions and transcriptional start sites. Genome Res. 14:2424-2429.
- 115. Dluzewski, A. R., P. R. Fryer, S. Griffiths, R. J. Wilson, and W. B. Gratzer. 1989. Red cell membrane protein distribution during malarial invasion. J. Cell Sci. 92:691-699.
- 116. **Dobson, M. J.** 1994. Malaria in England: a geographical and historical perspective. Parassitologia **36**:35-60.
- 117. Dokholyan, N. V., S. V. Buldyrev, S. Havlin, and H. E. Stanley. 2000. Distributions of dimeric tandem repeats in non-coding and coding DNA sequences. J. Theor. Biol. 202:273 - 282.
- 118. Donati, D., L. P. Zhang, Q. Chen, A. Chene, K. Flick, M. Nystrom, M. Wahlgren, and M. T. Bejarano. 2004. Identification of a polyclonal B-cell activator in *Plasmodium falciparum*. Infect. Immun. 72:5412-5418.
- dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32:5036-5044.
- 120. Ducasse-Cabanot, S., M. Cohen-Gonsaud, H. Marrakchi, M. Nguyen, D. Zerbib, J. Bernadou, M. Daffe, G. Labesse, and A. Quemard. 2004. In

vitro inhibition of the *Mycobacterium tuberculosis* beta-ketoacyl-acyl carrier protein reductase MabA by isoniazid. Antimicrob. Agents Chemother. **48:**242-249.

- 121. Durrand, V., A. Berry, R. Sem, P. Glaziou, J. Beaudou, and T. Fandeur. 2004. Variations in the sequence and expression of the *Plasmodium falciparum* chloroquine resistance transporter (*Pfcrt*) and their relationship to chloroquine resistance *in vitro*. Mol. Biochem. Parasitol. 136:273-285.
- 122. Eckstein-Ludwig, U., R. J. Webb, I. D. Van Goethem, J. M. East, A. G. Lee, M. Kimura, P. M. O'Neill, P. G. Bray, S. A. Ward, and S. Krishna. 2003. Artemisinins target the SERCA of *Plasmodium falciparum*. Nature 424:957-961.
- 123. El-Halawany, M. S., H. Shibata, K. Hitomi, and M Maki. 2005. Reevaluation of the predicted gene structure of *Dictyostelium* Cystatin A3 (*cpi*C) by nucleotide sequence determination of its cDNA* and its phylogenetic position in the cystatin superfamily. Mol. Biol. Rep. 32:257-264.
- 124. El Omari, K., B. Dhaliwal, M. Lockyer, I. Charles, A. R. Hawkins, and D. K. Stammers. 2006. Structure of *Staphylococcus aureus* guanylate monophosphate kinase. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. 1:949-953.
- 125. Ellinger, T., D. Behnke, R. Knaus, H. Bujard, and J. D. Gralla. 1994. Context-dependent effects of upstream A-tracts. Stimulation or inhibition of *Escherichia coli* promoter function. J. Mol. Biol. 239:466 - 475.
- 126. Enayati, A. A., H. Vatandoost, H. Ladonni, H, Townson, and J. Hemingway. 2003. Molecular evidence for a kdr-like pyrethroid resistance mechanism in the malaria vector mosquito *Anopheles stephensi*. Med. Vet. Entomol. 17:138-144.
- 127. Eyre-Walker, A., and M. Bulmer. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. 21:4599-4603.
- 128. Fanello, C., I. Carneiro, E. Ilboudo-Sanogo, N. Cuzin-Ouattara, A. Badolo, and C. F. Curtis. 2003. Comparative evaluation of carbosulfan- and permethrin-impregnated curtains for preventing house-entry by the malaria vector *Anopheles gambiae* in Burkina Faso. Med. Vet. Entomol. 17:333-338.
- 129. Favello, A., L. Hillier L, R. K. Wilson. 1995. Genomic DNA sequencing methods. Methods Cell Biol. 48:551-569.
- 130. Feagin, J. E., E. Werner, M. J. Gardner, D. H. Williamson, and R. J. Wilson. 1992. Homologies between the contiguous and fragmented rRNAs of the two *Plasmodium falciparum* extrachromosomal DNAs are limited to core sequences. Nucleic Acids Res. 20:879-887.
- 131. Ferdig, M. T., R. A. Cooper, J. Mu, B. Deng, D. A. Joy, X. Z. Su, and T. E. Wellems. 2004. Dissecting the loci of low-level quinine resistance in malaria parasites. Mol. Microbiol. 52:985-997.
- 132. Fernandez, V., A. Zavala, and H. Musto. 2001. Evidence for translational selection in codon usage in *Echinococcus* spp. Parasitology 123:203-209.
- 133. Fernando, S. D., D. M. Gunawardena, M. R. Bandara, D. De Silva, R. Carter, K. N. Mendis, and A. R. Wickremasinghe. 2003. The impact of repeated malaria attacks on the school performance of children. Am. J. Trop. Med. Hyg. 69:582 588.
- 134. Fischer, K., M. Chavchich, R. Huestis, D. W. Wilson, D. J. Kemp, and A. Saul. 2003. Ten families of variant genes encoded in subtelomeric regions of multiple chromosomes of *Plasmodium chabaudi*, a malaria species that

undergoes antigenic variation in the laboratory mouse. Mol. Microbiol. **48:**1209 - 1223.

- 135. **Fleischer, B.** 2004. 100 years ago: Giemsa's solution for staining of plasmodia Trop. Med. Int. Health **9:**755.
- 136. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.
- 137. Fogg, C., F. Bajunirwe, P. Piola, S. Biraro, F. Checchi, J. Kiguli, P. Namiiro, J. Musabe, A. Kyomugisha, and J. P. Guthmann 2004. Adherence to a six-dose regimen of artemether-lumefantrine for treatment of uncomplicated *Plasmodium falciparum* malaria in Uganda. Am. J. Trop. Med. Hyg. 71:525-530.
- 138. Foley, M., L. Tilley, W. H. Sawyer, and R. F. Anders. 1991. The ringinfected erythrocyte surface antigen of *Plasmodium falciparum* associates with spectrin in the erythrocyte membrane. Mol. Biochem. Parasitol. **46**:137-147.
- 139. **Fontenille, D., and F. Simard.** 2004. Unravelling complexities in human malaria transmission dynamics in Africa through a comprehensive knowledge of vector populations. Comp. Immunol. Microbiol. Infect. Dis. **27**:357-375.
- Forsdyke, D. R. 2002. Symmetry observations in long nucleotide sequences: a commentary on the discovery note of Qi and Cuticchia. Bioinformatics 18:215 - 217.
- Forsdyke, D. R., and J. R. Mortimer. 2000. Chargaff's legacy. Gene 261:127-137.
- 142. Foti, M., F. Granucci, and P. Ricciardi-Castagnoli. 2004. A central role for tissue-resident dendritic cells in innate responses. Trends Immunol. 25:650-654.
- 143. Fox, K. R. 1992. Wrapping of genomic polydA.polydT tracts around nucleosome core particles. Nucleic Acids Res. 20:1235 1242.
- 144. Frappat, L., C. Minichini, A. Sciarrino, and P. Sorba. 2003. Universality and Shannon entropy of codon usage. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 68:061910
- 145. Freerksen, E., E. W. Kanthumkumwa, and A. R. Kholowa. 1995. Malaria therapy and prophylaxis with cotrifazid, a multiple complex combination consisting of rifampicin + isoniazid + sulfamethoxazole + trimethoprim. Chemotherapy 41:396-398.
- 146. **Freiberg, C., H. P. Fischer, and N. A. Brunner.** 2005. Discovering the mechanism of action of novel antibacterial agents through transcriptional profiling of conditional mutants. Antimicrob. Agents Chemother. **49:**749-759.
- 147. Freire-Picos, M. A., M. I. Gonzalez-Siso, E. Rodriguez-Belmonte, A. M. Rodriguez-Torres, E. Ramil, and M. E. Cerdan. 1994. Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes. Gene 139:43-49.
- 148. Fried, M., F. Nosten, A. Brockman, B. J. Brabin, and P. E. Duffy. 1998. Maternal antibodies block malaria. Nature 395:851 – 852.
- Fuglsang, A. 2004. The 'effective number of codons' revisited. Biochem. Biophys. Res. Commun. 317:957-964.
- 150. **Fuglsang, A.** 2003. The effective number of codons for individual amino acids: some codons are more optimal than others. Gene **320**:185-190.

- 151. **Fuglsang, A.** 2004. Nucleotides downstream of start codons show marked non-randomness in *Escherichia coli* but not in *Bacillus subtilis*. Antonie Van Leeuwenhoek **86:**149-158.
- 152. **Fuglsang, A.** 2005. On the methodological weakness of 'the effective number of codons': a reply to Marashi and Najafabadi. Biochem. Biophys. Res. Commun. **327:1-3**.
- 153. Fujisaki, S., S. Ohnuma, T. Horiuchi, I. Takahashi, S. Tsukui, Y. Nishimura, T. Nishino, M. Kitabatake, and H. Inokuchi. 1996. Cloning of a gene from *Escherichia coli* that confers resistance to fosmidomycin as a consequence of amplification. Gene 175:83-87.
- Gallup, J. L., and J. D. Sachs. 2001. The economic burden of malaria. Am. J. Trop. Med. Hyg. 64:85 - 96.
- 155. **Garat, B., and H. Musto.** 2000. Trends of amino acid usage in the proteins from the unicellular parasite *Giardia lamblia*. Biochem. Biophys. Res. Commun. **279:**996-1000.
- 156. **Gardner, M. J., H. Tettelin, D. J. Carucci**, *et al.* 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Erratum in: Science (1998) 282(5395):1827. Science **282**:1126-1132.
- 157. Gardner, M. J., J. E. Feagin, D. J. Moore, K. Rangachari, D. H. Williamson, and R. J. Wilson. 1993. Sequence and organization of large subunit rRNA genes from the extrachromosomal 35 kb circular DNA of the malaria parasite *Plasmodium falciparum*. Nucleic Acids Res. 21:1067-1071.
- 158. Gardner, M. J., S. J. Shallom, J. M. Carlton, *et al.* 2002. Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. Nature **419**:531-534.
- 159. Gazarini M. L., A. P. T., T. Pozzan, and C. R. Garcia. 2003. Calcium signaling in a low calcium environment: how the intracellular malaria parasite solves the problem. J. Cell Biol. 161:103-110.
- Gazzinelli, R. T., C. Ropert, and M. A. Campos. 2004. Role of the Toll/interleukin-1 receptor signaling pathway in host resistance and pathogenesis during infection with protozoan parasites. Immunol. Rev. 201:9-25.
- 161. Geary, R. C. 1970. Relative efficiency of count of sign changes for assessing residual autocorrelation. Biometrika 57:123-127.
- 162. Gelfand, A. E., J. Ghosh, S. K. Knight and C. F. Sirmans. 1998. Analyzing real estate data problems using the Gibbs sampler. Real. Estate Econ. 26:469-492.
- 163. Genton, B., F. Al-Yaman, I. Betuela, R. F. Anders, A. Saul, K. Baea, M. Mellombo, J. Taraika, G. V. Brown, D. Pye, D. O. Irving, I. Felger, H. P. Beck, T. A. Smith, and M. P. Alpers. 2003. Safety and immunogenicity of a three-component blood-stage malaria vaccine (MSP1, MSP2, RESA) against *Plasmodium falciparum* in Papua New Guinean children. Vaccine 22:30-41.
- 164. Gilbert, W. 1978. Why genes in pieces? Nature 271:501.
- Giovannola, A. 1935. *Plasmodium ovale* considered as a modification of *Plasmodium vivax* after a long residence in the human host Am. J. Trop. Med. 15:175-186.
- 166. **Goerg, H., S. A. Ochola, and R. Goerg.** 1999. Treatment of malaria tropica with a fixed combination of rifampicin, co-trimoxazole and isoniazid: a clinical study. Chemotherapy **45**:68-76.

- 167. Goetz, R. M., and A. Fuglsang. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. Biochem. Biophys. Res. Commun. 327:4-7.
- 168. Gonzalez de Valdivia, E. I., and L. A. Isaksson. 2004. A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. Nucleic Acids Res. 32:5198-5205.
- 169. Good M. F., D. S. D., H. Xu, S. Elliott, and M. Wykes. 2004. The immunological challenge to developing a vaccine to the blood stages of malaria parasites. Immunol. Rev. 201:254-267.
- 170. Good, M. F., H. Xu, M. Wykes, and C. R. Engwerda. 2004. Development and regulation of cell-mediated immune responses to the blood stages of malaria: Implications for vaccine research. Annu. Rev. Immunol. 23:69-99.
- 171. Graham, A. L., T. J. Lamb, A. F. Read, and J. E. Allen. 2005. Malariafilaria coinfection in mice makes malarial disease more severe unless filarial infection achieves patency. J. Infect. Dis. **191:**410-421.
- 172. Grassi, B. 1900. Studi di un zoologo sulla malaria. Rome.
- 173. Grassi, B., and Feletti, R. 1890. Parasites malariques chez les oiseaux. Arch. Ital. de Biologie 13:297-300.
- 174. **Grassi, G. a. R. F.** 1892. Contribuz. allo studio dei parassiti malarici. Atti. Accad. Sci. Naturali Catania **4**:5.
- 175. Gray, S. A., and M. E. Konkel. 1999. Codon usage in the A/T-rich bacterium *Campylobacter jejuni*. Adv. Exp. Med. Biol. 473:231-235.
- 176. Green, W. H. 1993. Econometric analysis. Prentice Hall.
- 177. Grocock, R. J., and P. M. Sharp. 2001. Synonymous codon usage in *Cryptosporidium parvum*: identification of two distinct trends among genes. Int. J. Parasitol. 31:402-412.
- 178. Gross, S. S., and M. R. Brent. 2006. Using multiple alignments to improve gene prediction. J. Comput. Biol. 13:379-393.
- 179. Guigo, R., P. Agarwal, J. F. Abril, M. Burset, and J. W. Fickett. 2000. An assessment of gene prediction accuracy in large DNA sequences. Genome Res. 10:1631-1642.
- Gupta, S. K., T. K. Bhattacharyya, and T. C. Ghosh. 2004. Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. J. Biomol. Struct. Dyn. 21:527-536.
- 181. **Gutierrez, G., L. Marquez, and A. Marin.** 1996. Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. Nucleic Acids Res. **24**:2525-2527.
- 182. Guyatt, H. L., A. M. Noor, S. A. Ochola, and R. W. Snow. 2004. Use of intermittent presumptive treatment and insecticide treated bed nets by pregnant women in four Kenyan districts. Trop. Med. Int. Health 9:255-261.
- 183. Habtewold, T., A. Prior, S. J. Torr, and G. Gibson. 2004. Could insecticide-treated cattle reduce Afrotropical malaria transmission? Effects of deltamethrin-treated Zebu on *Anopheles arabiensis* behaviour and survival in Ethiopia. Med. Vet. Entomol. 18:408-417.
- 184. **Hackett, L. W.** 1937. Malaria in Europe. An ecological study. Oxford University Press, London.
- 185. Hall, A. P., E. B. Doberstyn, A. Nanokorn, and P. Sonkom. 1975. Falciparum malaria semi-resistant to clindamycin. Br. Med. J. 2:12-14.

- 186. Hall, N., A. Pain, M. Berriman, et al. 2002. Sequence of *Plasmodium* falciparum chromosomes 1, 3 9 and 13. Nature. 419:527-531.
- 187. Haniford, D. B., and D. E. Pulleyblank. 1985. Transition of a cloned d(AT)n-d(AT)n tract to a cruciform *in vivo*. Nucleic Acids Res. 13:4343-4363.
- 188. Hanna, J. N., S. A. Ritchie, D. P. Eisen, R. D. Cooper, D. L. Brookes, and B. L. Montgomery. 2004. An outbreak of *Plasmodium vivax* malaria in Far North Queensland, 2002. Med. J. Aust. 180:24-28.
- 189. Hanspal, M., V. K. Goel, S. S. Oh, and A. H. Chishti. 2002. Erythrocyte calpain is dispensable for malaria parasite invasion and growth. Mol. Biochem. Parasitol. 122:227-229.
- 190. Hargreaves, K., R. H. Hunt, B. D. Brooke, J. Mthembu, M. M. Weeto, T. S. Awolola, and M. Coetzee. 2003. Anopheles arabiensis and An. quadriannulatus resistance to DDT in South Africa. Med. Vet. Entomol. 17:417-422.
- Hartl, D. L. 2004. The origin of malaria: mixed messages from genetic diversity. Nat. Rev. Microbiol. 2:15 - 22.
- 192. **Hastings, I. M.** 2004. The origins of antimalarial drug resistance. Trends Parasitol. **20**:512-518.
- 193. Hata, S., K. Nishi, T. Kawamoto, H. J. Lee, H. Kawahara, T. Maeda, Y. Shintani, H. Sorimachi, and K. Suzuki. 2002. Both the conserved and the unique gene structure of stomach-specific calpains reveal processes of calpain gene evolution. J. Mol. Evol. 53.
- 194. Hay, S. I., C. A. Guerra, A. J. Tatem, A. M. Noor, and RW Snow. 2004. The global distribution and population at risk of malaria: past, present, and future. Lancet Infect. Dis. 4:327 - 336.
- 195. Hay, S. I., C. A. Guerra, A. J. Tatem, P. M. Atkinson, and R. W. Snow. 2005. Opinion - Tropical Infectious Diseases: Urbanization, malaria transmission and disease burden in Africa. Nat. Rev. Microbiol. 3:81-90.
- 196. Hayward, R., K. J. Saliba, and K. Kirk. 2005. *Pfmdr1* mutations associated with chloroquine resistance incur a fitness cost in *Plasmodium falciparum*. Mol. Microbiol. 55:1285-1295.
- 197. Hien T. T., a. N. J. W. 1993. Qinghaosu. Lancet 341:603-608.
- 198. **Hill, M. O.** 1974. Correspondance analysis: a neglected multivariate method. Appl. Stats. **23**:340-354.
- 199. Hilleren, P. J., and R. Parker. 2003. Cytoplasmic degradation of splicedefective pre-mRNAs and intermediates. Mol. Cell 12:1453-1465.
- 200. Hippocrates. 1923. Airs, Waters, Places. Loeb Classical Library, London.
- 201. Hippocrates. 1923. Epidemics. Loeb Classical Library, London.
- 202. Hirtzlin, J., P. M. Farber, R. M. Franklin, and A. Bell. 1995. Molecular and biochemical characterization of a *Plasmodium falciparum* cyclophilin containing a cleavable signal sequence. Eur. J. Biochem. 232:765-772.
- 203. Hizver, J., H. Rozenberg, F. Frolow, D. Rabinovich, and Z. Shakked. 2001. DNA bending by an adenine-thymine tract and its role in gene regulation. Proc. Natl. Acad. Sci. U S A 98:84900 - 84905.
- 204. Honigsbaum, M. 2003. The fever trail: In search of the cure for malaria. Pan.
- 205. **Hooper, S. D., and O. G. Berg.** 2000. Gradients in nucleotide and codon usage along *Escherichia coli* genes. Nucleic Acids Res. **28**:3517-3523.
- 206. **Hori, R., and R. A. Firtel.** 1994. Identification and characterization of multiple A/T-rich cis-acting elements that control expression from *Dictyostelium* actin promoters: the *Dictyostelium* actin upstream activating

sequence confers growth phase expression and has enhancer-like properties. Nucleic Acids Res. **22:**5099 - 5111.

- 207. **Hosfield, C. M., J. S. Elce, P. L. Davies, and Z. Jia.** 1999. Crystal structure of calpain reveals the structural basis for Ca(2+)-dependent protease activity and a novel mode of enzyme activation. EMBO J. **18**:6880-6889.
- 208. Hoshen, M. B., and A. P. Morse. 2004. A weather-driven model of malaria transmission. Malar. J. 3:132.
- 209. **Hsu, E.** 2006. Reflections on the 'discovery' of the antimalarial qinghao. Br. J. Clin. Pharmacol. **61**:666-670.
- Huestis, R., and K. Fischer. 2001. Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2. Mol. Biochem. Parasitol. 118:187-199.
- Huestis R, N. C., M. Tchavtchitch, and A. Saul. 2001. An algorithm to predict 3' intron splice sites in *Plasmodium falciparum* genomic sequences. Mol. Biochem. Parasitol. 112:71-77.
- Huestis. R., a. K. F. 2001. Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2. Mol. Biochem. Parasitol. 118:187-199.
- 213. **Hughes, A. L.** 1991. Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. Genetics **127**:345-353.
- 214. Hughes, M. K., and A. L. Hughes. 1995. Natural selection on *Plasmodium* surface proteins. Mol. Biochem. Parasitol. 71:99-113.
- 215. Hyman, R. W., E. Fung, A. Conway *et al.* 2002. Sequence of *Plasmodium falciparum* chromosome 12. Nature **419**:534-547.
- 216. Hyman, R. W., E. Fung, A. Conway, et al. 2002. Sequence of *Plasmodium* falciparum chromosome 12. Nature 419:534 547.
- 217. Iida, K., M. Seki, T. Sakurai, M. Satou, K. Akiyama, T. Toyoda, A. Konagaya, and K. Shinozaki. 2004. Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. Nucleic Acids Res. 32:5096-5103.
- 218. Ishiura, S., H. Murofushi, K. Suzuki, and K. Imahori. 1978. Studies of a calcium-activated neutral protease from chicken skeletal muscle. I. Purification and characterization. J. Biochem. 84:225-230.
- Iyer, V., and K. Struhl. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J. 14:2570 2579
- 220. JA., N. 2006. Pseudogenes of the human HPRT1 gene. Environ. Mol. Mutagen. 47:212-218.
- 221. Jaeger, A., P. Sauder, J. Kopferschmitt, and F. Flesch. 1987. Clinical features and management of poisoning due to antimalarial drugs. Med. Toxicol. Adverse Drug Exp. 2:242-273.
- 222. Jalajakumari, M. B., C. J. Thomas, , R. Halter, and P. A. Manning. 1989. Genes for biosynthesis and assembly of CS3 pili of CFA/II enterotoxigenic *Escherichia coli*: novel regulation of pilus production by bypassing an amber codon. Mol. Microbiol. 3:1685-1695.
- 223. Janse, C. J., J. M. Carlton, D. Walliker, and A. P. Waters. 1994. Conserved location of genes on polymorphic chromosomes of four species of malaria parasites. Mol. Biochem. Parasitol. 68:285-296.

- 224. Jaramillo, M., I. Plante, N. Ouellet, K. Vandal, P. A. Tessier, and M. Olivier. 2004. Hemozoin-inducible proinflammatory events *in vivo*: potential role in malaria infection. J. Immunol. **172:**3101-3110.
- 225. Jaramillo, M., M. Godbout, and M. Olivier. 2005. Hemozoin induces macrophage chemokine expression through oxidative stress-dependent and independent mechanisms. J. Immunol. 174:475-484.
- Jonsson, A. B., G. Nyberg, and S. Normark. 1991. Phase variation of gonococcal pili by frameshift mutation in *pil*C, a novel gene for pilus assembly. EMBO J. 10:477 - 488.
- 227. Jowett, M., and N. J. Miller. 2005. The financial burden of malaria in Tanzania: implications for future government policy. Int. J. Health Plan. Manage. 20:67 - 84.
- 228. Kannan, R., K. Kumar, D. Sahal, S. Kukreti, and V. S. Chauhan. 2004. Reaction of artemisinin with hemoglobin: implications for antimalarial activity. Biochem. J. 385:409-418.
- 229. **Kapp, C.** 2004. New international convention allows use of DDT for malaria control. Bull. World Health Organ. **82:**472-473.
- 230. Kazadi, W., J. D. Sexton, M. Bigonsa, B. W'Okanga, and M. Way. 2004. Malaria in primary school children and infants in Kinshasa, Democratic Republic of the Congo: surveys from the 1980s and 2000. Am. J. Trop. Med. Hyg. 71:97 - 102.
- 231. Keller, C. C., J. B. Hittner, B. K. Nti, J. B. Weinberg, P. G. Kremsner, and D. J. Perkins. 2004. Reduced peripheral PGE2 biosynthesis in *Plasmodium falciparum* malaria occurs through hemozoin-induced suppression of blood mononuclear cell cyclooxygenase-2 gene expression via an interleukin-10-independent mechanism. Mol. Med. 10:45-54.
- 232. Kerr, A. R., J. F. Peden, and P. M. Sharp. 1997. Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. Mol. Microbiol. **25**:1177–1179.
- 233. Kessl, J. J., P. Hill, B. B. Lange, S. R. Meshnick, B. Meunier, and B. L. Trumpower. 2004. Molecular basis for atovaquone resistance in *Pneumocystis jirovecii* modeled in the cytochrome bc(1) complex of *Saccharomyces cerevisiae*. J. Biol. Chem. 279:2817-2824.
- 234. Khasnis, A. A., and D. R. Karnad. 2003. Human immunodeficiency virus type 1 infection in patients with severe falciparum malaria in urban India. J. Postgrad. Med. **49:**114-117.
- 235. Killeen, G. F. 2003. Following in Soper's footsteps: northeast Brazil 63 years after eradication of *Anopheles gambiae*. Lancet Infect. Dis. **3**:663-666.
- 236. King, L. J. 1969. Statistical analysis in geography. Prentice Hall.
- Knapp, B., K. Gunther, and K. Lingelbach. 1991. In vitro translation of Plasmodium falciparum aldolase is not initiated at an unusual site. EMBO J. 10:3095-3097.
- 238. Knapp, B., U. Nau, E. Hundt, and H. A. Kupper. 1991. Demonstration of alternative splicing of a pre-mRNA expressed in the blood stage form of *Plasmodium falciparum*. J. Biol. Chem. 266:7148-7154.
- 239. Knauer, A., J. Sirichaisinthop, F. F. Reinthaler, G. Wiedermann, G. Wernsdorfer, and W. H. Wernsdorfer. 2003. In vitro response of Plasmodium falciparum to the main alkaloids of Cinchona in northwestern Thailand. Wien Klin Wochenschr. 115:39-44.

- 240. **Kochetov, A. V.** 2004. AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. Bioinformatics **21:8**37-840.
- 241. Kovats, R. S., M. J. Bouma, S. Hajat, E. Worrall, and A. Haines. 2003. *El Nino* and health. Lancet **362**:1481-1489.
- 242. **Kozak, M.** 1997. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J. **16**:2482-2492.
- 243. Kremsner, P. G., and S. Krishna. 2004. Antimalarial combinations. Lancet 364:285-294.
- 244. Krishnegowda, G., A. M. Hajjar, J. Zhu, E. J. Douglass, S. Uematsu, S. Akira, A. S. Woods, and D. C. Gowda. 2005. Induction of proinflammatory responses in macrophages by the glycosylphosphatidylinositols (GPIs) of *Plasmodium falciparum*: Cell signaling receptors, GPI structural requirement, and regulation of GPI activity. J. Biol. Chem. 280:8606-8616.
- 245. **Krogh, A.** 2000. Using database matches with for HMMGene for automated gene detection in Drosophila. Genome Res. **10**:523-528.
- 246. Krotoski, W. A., P. C. Garnham, R. S. Bray, D. M. Krotoski, R. Killick-Kendrick, C. C. Draper, G. A. Targett, and M. W. Guy. 1982. Observations on early and late post-sporozoite tissue stages in primate malaria. I. Discovery of a new latent form of *Plasmodium cynomolgi* (the hypnozoite), and failure to detect hepatic forms within the first 24 hours after infection. Am. J. Trop. Med. Hyg. **31**:24-35.
- 247. Krudsood, S., M. Imwong, P. Wilairatana, S. Pukrittayakamee, A. Nonprasert, G. Snounou, N. J. White, and S. Looareesuwan. 2005. Artesunate-dapsone-proguanil treatment of falciparum malaria: genotypic determinants of therapeutic response. Trans. R. Soc. Trop. Med. Hyg. 99:142-149.
- 248. **Kruger, C.** 2004. Malaria intermittent preventive treatment and EPI coverage. Lancet **363**:2000 2001.
- 249. Kruglyak, S., RT Durrett, M. D. Schug, and C. F. Aquadro 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. U S A. 95:10774-10778.
- Krzycki, J. A. 2004. Function of genetically encoded pyrrolysine in corrinoiddependent methylamine methyltransferases. Curr. Opin. Chem. Biol. 8:484-491.
- 251. Kuhn, K. G., D. H. Campbell-Lendrum, B. Armstrong, and C. R. Davies. 2003. Malaria in Britain: past, present, and future. Proc. Natl. Acad. Sci. U S A. 100:9997-10001.
- 252. Kurtis, J. D., M. R. Hollingdale, A. J. Luty, D. E. Lanar, U. Krzych, and P. E. Duffy. 2001. Pre-erythrocytic immunity to *Plasmodium falciparum*: the case for an LSA-1 Vaccine. Trends Parasitol. 17:219-223.
- 253. Kussmaul, A. 1874. Zur Lehre vom Diabetes mellitus. Über eine eigenthümliche Todesart bei Diabetischen, über Acetonämie, Glycerin-Behandlung des Diabetes und Einspritzungen von Diastase in's Blut bei dieser Krankheit. Deutsches Archiv für klinische Medicin, Leipzig 14.
- 254. Kyte, J., and R. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105-132.

- 255. Lafay, B., A. T. Lloyd, M. J. McLean, K. M. Devine, P. M. Sharp, and K. H. Wolfe. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. 27:1642-1649.
- Lafay, B., J. C. Atherton, and P. M. Sharp. 2000. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology 146:851-860.
- 257. Lagerstrom, M. C., A. R. Hellstrom, D. E. Gloriam, T. P. Larsson, H. B. Schioth, and R. Fredriksson. 2006. The G protein-coupled receptor subset of the chicken genome. PLoS Comput. Biol. 2:e54.
- 258. Laird, J., P. Armengaud, P. Giuntini, V. Laval, and J. J. Milner. 2004. Inappropriate annotation of a key defence marker in *Arabidopsis*: will the real PR-1 please stand up? Planta **219**:1089-1092.
- 259. Lang, B., C. I. Newbold, G. Williams, N. Peshu, K. Marsh, and C. R. Newton. 2005. Antibodies to voltage-gated calcium channels in children with falciparum malaria. J. Infect. Dis. **191:**117-121.
- 260. Langenfeld, J., H. Kiyokawa, D. Sekula, J. Boyle, and E. Dmitrovsky. 1997. Posttranslational regulation of cyclin D1 by retinoic acid: a chemoprevention mechanism. Proc Natl Acad Sci U S A. 94:12070-12074.
- 261. Langhorne, J. R., F. Albano, M. Hensmann, L. Sanni, E. Cadman, C. Voisine, and A. M. Sponaas. 2004. Dendritic cells, pro-inflammatory responses, and antigen presentation in a rodent malaria infection. Immunol. Rev. 201.
- 262. Laufer, M. K., and C. V. Plowe. 2004. Withdrawing antimalarial drugs: impact on parasite resistance and implications for malaria treatment policies. Drug Resist. Updat. 7:279-288.
- 263. Laveran, A. 1880. Nouveau parasite du sang. Bull. Acad. Med. 9:1235-1236.
- 264. Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208-214.
- 265. Laxminarayan, R. 2004. Act now or later? Economics of malaria resistance. Am. J. Trop. Med. Hyg. 71:187-195.
- 266. Lee, D. H., and A. L. Goldberg. 1998. Proteasome inhibitors: valuable new tools for cell biologists. Trends Cell Biol. 8:397-403.
- 267. Lee, E. W., M. N. Huda, T. Kuroda, T. Mizushima, and T. Tsuchiya. 2003. EfrAB, an ABC multidrug efflux pump in *Enterococcus faecalis*. Antimicrob. Agents Chemother. **47**:3733-3738.
- 268. Lee, L. J., T. R. Hughes, and B. J. Frey. 2006. How many genes are there? Science 311:1709 1711.
- 269. Leibundgut, M., C. Frick, M. Thanbichler, A. Bock, and N. Ban. 2005. Selenocysteine tRNA-specific elongation factor SelB is a structural chimaera of elongation and initiation factors. EMBO J. 24:11-22.
- 270. Lerat, E., and H. Ochman. 2004. Psi-Phi: exploring the outer limits of bacterial pseudogenes. Genome Res. 14:2273-2278.
- 271. Li, J., M. M. Riehle, Y. Zhang, J. Xu, F. Oduol, S. M. Gomez, K. Eiglmeier, BM Ueberheide, J. Shabanowitz, D. F. Hunt, J. M. Ribeiro, and K. D. Vernick. 2006. Anopheles gambiae genome reannotation through synthesis of *ab initio* and comparative gene prediction algorithms. Genome Biol. 7:R24.

- 272. Li, W. 2001. Delineating relative homogeneous G+C domains in DNA sequences. Gene 276:57-72.
- 273. Lindblade, K. A., E. D. Walker, A. W. Onapa, J. Katungu, and M. L. Wilson. 1999. Highland malaria in Uganda: prospective analysis of an epidemic associated with *El Nino*. Trans. R. Soc. Trop. Med. Hyg. 93:480-487.
- 274. Lindblade, K. A., T. P. Eisele, J. E. Gimnig, J. A. Alaii, F. Odhiambo, F. O. ter Kuile, W. A. Hawley, K. A. Wannemuehler, P. A. Phillips-Howard, D. H. Rosen, B. L. Nahlen, D. J. Terlouw, K. Adazu, J. M. Vulule, and L. Slutsker. 2004. Sustainability of reductions in malaria transmission and infant mortality in western Kenya with use of insecticide-treated bednets: 4 to 6 years of follow-up. JAMA 291:2571-2580.
- Lindsay, S. W., R. Bodker, R. Malima, H. A. Msangeni, and W. Kisinza.
 2000. Effect of 1997-98 *El Nino* on highland malaria in Tanzania. Lancet
 355:989-990.
- Liu, Q., and Q. Xue. 2004. Computational identification and sequence analysis of stop codon readthrough genes in *Oryza sativa*. Biosystems 77:33-39.
- 277. Liu, Y. 2001. Dendritic cell subsets and lineages, and their functions in innate and adaptive immunity. Cell 106.
- 278. Lloyd, A. T., and P. M. Sharp. 1992. Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res. 20:5289-5295.
- 279. Lobry, J. R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13:660 665.
- 280. Lobry, J. R., and C. Gautier. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. **22**:3174-3180.
- Lomsadze, A., V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky. 2005. Gene identification in novel eukaryotic genomes by selftraining algorithm. Nucleic Acids Res. 33:6494-506.
- 282. Long, M., and C. Rosenberg. 2000. Testing the "proto-splice sites" model of intron origin: evidence from analysis of intron phase correlations.
- Mol. Biol. Evol. 17:1789-1796.
- 283. Long, M., E. Betran, K. Thornton, and W. Wang. 2003. The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. 4:865-875.
- 284. Looareesuwan, S., C. Viravan, H. K. Webster, D. E. Kyle, D. B. Hutchinson, and C. J. Canfield. 1996. Clinical studies of atovaquone, alone or in combination with other antimalarial drugs, for treatment of acute uncomplicated malaria in Thailand. Am. J. Trop. Med. Hyg. 54:62-66.
- 285. Lyke, K. E., R. Burges, Y. Cissoko, L. Sangare, M. Dao, I. Diarra, A. Kone, R. Harley, C. V. Plowe, O. K. Doumbo, and M. B. Sztein. 2004. Serum levels of the proinflammatory cytokines Interleukin-1 beta, IL-6, IL-8, IL-10, tumor necrosis factor alpha, and IL-12 (p70) in Malian children with severe *Plasmodium falciparum* malaria and matched uncomplicated malaria or healthy controls. Infect. Immun. 72:5630-5637.
- 286. MacDonald, S. M., J. Bhisutthibhan, T. A. Shapiro, S. J. Rogerson, T. E. Taylor, M. Tembo, J. M. Langdon, and S. R. Meshnick. 2001. Immune mimicry in malaria: *Plasmodium falciparum* secretes a functional histamine-

releasing factor homolog *in vitro* and *in vivo*. Proc. Natl. Acad. Sci. U S A. **98**:10829-10832.

- 287. Mahadevan, R., and B. Palsson. 2004. Properties of metabolic networks: structure vs. function. Biophys. J. 88:L07-9.
- 288. Majumdar, S., S. K. Gupta, V. S. Sundararajan, and T. C. Ghosh. 1999. Compositional correlation studies among the three different codon positions in 12 bacterial genomes. Biochem. Biophys. Res. Commun. 266:66-71.
- 289. Maki, M., Y. Kitaura, H. Satoh, S. Ohkouchi, and H. Shibata. 2002. Structures, functions and molecular evolution of the penta-EF-hand Ca2+binding proteins. Biochim. Biophys. Acta 1600:51-60.
- 290. Malaguarnera, L., and S. Musumeci. 2002. The immune response to *Plasmodium falciparum* malaria. Lancet Infect Dis. 2:472-478.
- Mandelbaum-Schmid, J. 2004. HIV/AIDS, hunger and malaria are the world's most urgent problems, say economists. Bull. World Health Organ. 82:554 - 555.
- Mao, P. L., T. F. Liu, K. Kueh, and P. Wu. 2004. Predicting the efficiency of UAG translational stop signal through studies of physicochemical properties of its composite mono- and dinucleotides. Comput. Biol. Chem. 28:245-256.
- 293. **Marashi, S. A., and H. S. Najafabadi.** 2004. How reliable re-adjustment is: correspondence regarding A. Fuglsang, "The 'effective number of codons' revisited". Biochem. Biophys. Res. Commun. **324:**1-2.
- 294. Martin-Galiano, A. J., J. M. Wells, and A. G. De La Campa. 2004. Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. Microbiology 150:2313-2325.
- 295. Martin, D. C., and J. D. Arnold. 1968. Treatment of acute falciparum malaria with sulfalene and trimethoprim. JAMA 203:476-480.
- 296. Martin, D. C., and J. D. Arnold. 1967. Trimethoprim in therapy of acute attacks of malaria. J. Clin. Pharmacol. J. New Drugs 7:336-341.
- 297. Martin, P. W. 2006. Introduction to Basic Legal Citation. Cornell Law School.
- 298. Martin, R. E., and K. Kirk. 2004. The malaria parasite's chloroquine resistance transporter is a member of the drug/metabolite transporter superfamily. Mol. Biol. Evol. 21:1938-1949.
- 299. Martinelli, A., P. Hunt, R. Fawcett, P. V. Cravo, D. Walliker, and R. Carter. 2005. An AFLP-based genetic linkage map of *Plasmodium chabaudi chabaudi*. Malar. J **4**:11.
- McCallum-Deighton, N., and A. A. Holder. 1992. The role of calcium in the invasion of human erythrocytes by *Plasmodium falciparum*. Mol. Biochem.Parasitol. 50:317-323.
- 301. McClellan J. A., E. P., and D. M. Lilley. 1986 (A-T)n tracts embedded in random sequence DNA--formation of a structure which is chemically reactive and torsionally deformable. Nucleic Acids Res. 14:9291 9309.
- 302. McGill, M. A., and C. J. McGlade. 2003. Mammalian numb proteins promote Notch1 receptor ubiquitination and degradation of the Notch1 intracellular domain. J Biol Chem. 278:23196-23203.
- 303. McGregor, I. A. 1984. Epidemiology, malaria and pregnancy. Am. J. Trop. Med. Hyg. 33:517-525.

- McInerney, J. O. 1997. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microbial. Comp. Genom. 2:1 – 10.
- 305. **McInerney, J. O.** 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA **95**:10698-10703.
- McLean, M. L., K. H. Wolfe, and K. M. Devine. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47:691 – 696.
- Meintjes, P. L., and A. G. Rodrigo. 2005. Evolution of relative synonymous codon usage in human immunodeficiency virus type-1. Bioinform. Comput. Biol. 3:157-168.
- 308. Meraldi, V., I. Nebie, A. B. Tiono, D. Diallo, E. Sanogo, M. Theisen, P. Druilhe, G. Corradin, R. Moret, and B. S. Sirima. 2004. Natural antibody response to *Plasmodium falciparum* Exp-1, MSP-3 and GLURP long synthetic peptides and association with protection. Parasite Immunol. 26:265-272.
- 309. Miller, L. H., R. H. Glew, D. J. Wyler, W. A. Howard, W. E. Collins, P. G. Contacos, and F. A. Neva. 1974. Evaluation of clindamycin in combination with quinine against multidrug-resistant strains of *Plasmodium falciparum*. Am. J. Trop. Med. Hyg. 23:565-569.
- 310. Millet, J., S. Alibert, M. Torrentino-Madamet, C. Rogier, C. Santelli-Rouvier, P. Bigot, J. Mosnier, E. Baret, J. Barbe, D. Parzy, and B. Pradines. 2004. Polymorphism in *Plasmodium falciparum* drug transporter proteins and reversal of *in vitro* chloroquine resistance by a 9,10dihydroethanoanthracene derivative. Antimicrob. Agents Chemother. 48:4869-4872.
- 311. Missinou, M. A., S. Borrmann, A. Schindler, S. Issifou, A. A. Adegnika, P. B. Matsiegui, R. Binder, B. Lell, J. Wiesner, T. Baranek T, H. Jomaa, and P. G. Kremsner. 2002. Fosmidomycin for malaria. Lancet 360:1941-1942.
- 312. **Missiroli, A.** 1948. *Anopheles* control in the Mediterranean area. Proceedings of the 4th International Congress on Tropical Medicine and Malaria (Washington, DC):1566-1575.
- 313. Mitchell, D., and A. Bell. 2003. PEST sequences in the malaria parasite *Plasmodium falciparum*: a genomic study. Malar. J. 2:16.
- 314. Mitchell, D., and R. Bridge. 2006. An investigation of genomic base distribution. Biochem. Biophys. Res. Commun. 344:612 616.
- 315. Mitchell, D., and R. Bridge. Some rules of genome composition. Submitted.
- 316. Mitchell, D., and R. Bridge. 2006. A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. 340:90 94
- 317. Modiano, D., V. Petrarca, B. S. Sirima, I. Nebie, D. Diallo, F. Esposito, and M. Coluzzi. 1996. Different response to *Plasmodium falciparum* malaria in west African sympatric ethnic groups. Proc. Natl. Acad. Sci. U S A. 93:13206-13211.
- 318. Modiano, D., V. Petrarca, B. S. Sirima, I. Nebie, G. Luoni, F. Esposito F, and M. Coluzzi. 1998. Baseline immunity of the population and impact of insecticide-treated curtains on malaria infection. Am. J. Trop. Med. Hyg. 59:336-340.
- 319. Mohan, K., and M. M. Stevenson. 1998. Acquired immunity to asexual blood stages. ASM Press.

- 320. Montero, L. M., J. Salinas, G. Matassi, and G. Bernardi. 1990. Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res. 18:1859-1867.
- 321. Moorthy, V. S., E. B. Imoukhuede, P. Milligan, K. Bojang, S. Keating, P. Kaye, M. Pinder, S. C. Gilbert, G. Walraven, B. M. Greenwood, and A. S. Hill. 2004. A Randomised, Double-Blind, Controlled Vaccine Efficacy Trial of DNA/MVA ME-TRAP Against Malaria Infection in Gambian Adults. Plos Med. 1:e33.
- 322. Moritz, E., S. Seidensticker, A. Gottwald, W. Maier, A. Hoerauf, J. T. Njuguna, and A. Kaiser 2004. The efficacy of inhibitors involved in spermidine metabolism in *Plasmodium falciparum*, *Anopheles stephensi* and *Trypanosoma evansi*. Parasitol. Res. 94:37-48.
- 323. **Mottagui-Tabar, S., and L. A. Isaksson.** 1998. The influence of the 5' codon context on translation termination in *Bacillus subtilis* and *Escherichia coli* is similar but different from *Salmonella typhimurium*. Gene **212**:189-196.
- 324. Mouchet, J., S. Manguin, J. Sircoulon, S. Laventure, O. Faye, A. W. Onapa, P. Carnevale, J. Julvez, and D. Fontenille. 1998. Evolution of malaria in Africa for the past 40 years: impact of climatic and human factors. J. Am. Mosq. Control Assoc. 14:121-130.
- 325. Mouchiroud, D., G. D'Onofrio, B. Aissani, G. Macaya, C. Gautier, and G. Bernardi. 1991. The distribution of genes in the human genome. Gene 100:181-187.
- 326. Mourier, T., A. Pain, B. Barrell, and S. Griffiths-Jones. 2005. A selenocysteine tRNA and SECIS element in *Plasmodium falciparum*. RNA 11:119-122.
- 327. Muehlen, M., J. Schreiber, S. Ehrhardt, R. Otchwemah, T. Jelinek, U. Bienzle, and F. P. Mockenhaupt. 2004. Short communication: Prevalence of mutations associated with resistance to atovaquone and to the antifolate effect of proguanil in *Plasmodium falciparum* isolates from northern Ghana. Trop. Med. Int. Health 9:361-363.
- 328. Muhia, D. K., C. A. Swales, U. Eckstein-Ludwig, S. Saran, S. D. Polley, J. M. Kelly, P. Schaap, S. Krishna, and D. A. Baker. 2003. Multiple splice variants encode a novel adenylyl cyclase of possible plastid origin expressed in the sexual stage of the malaria parasite *Plasmodium falciparum*. J. Biol. Chem. 278:22014-22022.
- 329. Mulligan, H. W. 1935. Descriptions of two species of monkey *Plasmodium* isolated from *Silenus irus*. Archiv. Protistenk. **84**:285-314.
- 330. **Musto, H., H. Romero, and A. Zavala.** 2003. Translational selection is operative for synonymous codon usage in *Clostridium perfringens* and *Clostridium acetobutylicum*. Microbiology **149:8**55-863.
- 331. Musto, H., H. Romero, and H. Rodriguez-Maseda. 1998. Heterogeneity in codon usage in the flatworm *Schistosoma mansoni*. J. Mol. Evol. 46:159-167.
- 332. Nadeau, J. G., and D. M. Crothers. 1989. Structural basis for DNA bending. Proc. Natl. Acad. Sci. U S A 86:26220 - 26226.
- 333. Nakama, T., O. Nureki, and S. Yokoyama. 2001. Structural basis for the recognition of isoleucyl-adenylate and an antibiotic, mupirocin, by isoleucyl-tRNA synthetase. J. Biol. Chem. 276:47387 47393.
- Namy, O., I. Hatin, and J. P. Rousset. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep. 2:787-793.

- 335. Naya, H., A. Zavala, H. Romero, H. Rodriguez-Maseda, and H. Musto. 2004. Correspondence analysis of amino acid usage within the family *Bacillaceae*. Biochem. Biophys. Res. Commun. 325:1252-1257.
- 336. Ndao, M., E. Bandyayera, E. Kokoskin, D. Diemert, T. W. Gyorkos, J. D. Maclean, R. St John, and B. J. Ward. 2005. Malaria "epidemic" in Quebec: diagnosis and response to imported malaria. CMAJ. 172:46-50.
- 337. Nebert, D. W., and H. M. Wain. 2003. Update on human genome completion and annotations: Gene nomenclature. Hum. Genomics 1:66-71.
- 338. Nelson, D. R. 2004. 'Frankenstein genes', or the Mad Magazine version of the human pseudogenome. Hum. Genomics 1:310-316.
- 339. Newman, R. D., M. E. Parise, A. M. Barber, and R. W. Steketee. 2004. Malaria-related deaths among U.S. travelers, 1963-2001. Ann. Intern. Med. 141:547-555.
- 340. Nguyen, H. D., M. Yoshihama, and N. Kenmochi. 2006. Phase distribution of spliceosomal introns: implications for intron origin. BMC Evol. Biol. 6:69.
- 341. Nielsen H., A. K. a. T. G. T. 1986. Suppression of blood monocyte and neutrophil chemotaxis in acute human malaria. Parasite Immunol. 8:541–550.
- Nikolaou C, a. Y. A. 2006. Deviations from Chargaff's second parity rule in organellar DNA. Insights into the evolution of organellar genomes. Gene 381C:34-41.
- 343. Nikou, D., H. Ranson, and J. Hemingway. 2003. An adult-specific CYP6 P450 gene is overexpressed in a pyrethroid-resistant strain of the malaria vector, *Anopheles gambiae*. Gene **318**:91-102.
- 344. Nilsson, D., and B. Andersson 2005. Strand asymmetry patterns in trypanosomatid parasites. Exp. Parasitol. 109:143 149.
- 345. Noor, A. M., D. Zurovac, S. I. Hay, S. A. Ochola, and R. W. Snow. 2003. Defining equity in physical access to clinical services using geographical information systems as part of malaria planning and monitoring in Kenya. Trop. Med. Int. Health 8:917-926.
- 346. O'Donnell, R. A., L. H. Freitas-Junior, P. R. Preiser, D. H. Williamson, M. Duraisingh, T. F. McElwain, A. Scherf, A. F. Cowman, and B. S. Crabb. 2002. A genetic screen for improved plasmid segregation reveals a role for Rep20 in the interaction of *Plasmodium falciparum* chromosomes. EMBO J. 21:1231 - 1239.
- 347. O'Neil-Dunne, I., R. N. Achur, S. T. Agbor-Enoh, M. Valiyaveettil, R. S. Naik, C. F. Ockenhouse, A. Zhou, R. Megnekou, R. Leke, D. W. Taylor, and D. C. Gowda. 2001. Gravidity-dependent production of antibodies that inhibit binding of *Plasmodium falciparum*-infected erythrocytes to placental chondroitin sulfate proteoglycan during pregnancy. Infect. Immun. 69:7487-7492.
- 348. Ocana-Morgner, C., M. M. Mota, and A. Rodriguez. 2003. Malaria blood stage suppression of liver stage immunity by dendritic cells. J. Exp. Med. 197:143-151.
- 349. Ogwang, S., M. Engl, M. Vigl, H. Kollaritsch, G. Wiedermann, and W. H. Wernsdorfer. 2003. Clinical and parasitological response of *Plasmodium falciparum* to chloroquine and sulfadoxine/pyrimethamine in rural Uganda. Wien Klin Wochenschr. 115:45-49.
- 350. Ohler, U., N. Shomron, and C. B. Burge. 2005. Recognition of unknown conserved alternatively spliced exons. PLoS Comput. Biol. 1:113-122.

- 351. Okoko, B. J., G. Enwere, and M. O. Ota. 2003. The epidemiology and consequences of maternal malaria: a review of immunological basis. Acta Trop. 87:193-205.
- 352. Olaya, P., and M. Wasserman. 1991. Effect of calpain inhibitors on the invasion of human erythrocytes by the parasite *Plasmodium falciparum*. Biochim. Biophys. Acta **1096**:217-221.
- Olaya, P. a. M. W. 1991. Effect of calpain inhibitors on the invasion of human erythrocytes by the parasite *Plasmodium falciparum*. Biochim. Biophys. Acta 1096:217–221.
- 354. Oliver, J. L., P. Bernaola-Galvan, P. Carpena, and R. Roman-Roldan. 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276:47-56.
- 355. **Omumbo, J. A., C. A. Guerra, S. I. Hay, and R. W. Snow.** 2005. The influence of urbanisation on measures of *Plasmodium falciparum* infection prevalence in East Africa. Acta Trop. **93**:11-21.
- 356. **OO., O.** 2003. The status of malaria among pregnant women: a study in Lagos, Nigeria. Afr. J. Reprod. Health **7:**77 **8**3.
- 357. Over, M., B. Bakote'e, R. Velayudhan, P. Wilikai, and P. M. Graves. 2004. Impregnated nets or DDT residual spraying? Field effectiveness of malaria prevention techniques in Solomon islands, 1993-1999. Am. J. Trop. Med. Hyg. 71:214-223.
- Pace, T., M. Ponzi, R. Scotti, and C. Frontali. 1995. Structure and superstructure of *Plasmodium falciparum* subtelomeric regions. Mol. Biochem. Parasitol. 69:257 - 268.
- 359. Paces, J., R. Zika, V. Paces, A. Pavlicek, O. Clay, and G. Bernardi. 2004. Representing GC variation along eukaryotic chromosomes. Gene 333:135 -141.
- Palacios, C. a. J. J. W. 2002. A strong effect of AT mutational bias on amino acid usage in *Buchnera* is mitigated at high expression genes. Mol. Biol. Evol. 19:1575–1584.
- 361. Pasloske, B. L., D. I. Baruch, M. R. van Schravendijk, S. M. Handunnetti, M. Aikawa, H. Fujioka, T. F. Taraschi, J. A. Gormley, and R. J. Howard. 1993. Cloning and characterization of a *Plasmodium falciparum* gene encoding a novel high-molecular weight host membrane-associated protein, PfEMP3. Mol. Biochem. Parasitol. **59**:59-72.
- Paul, R. E., M. Diallo, and P. T. Brey. 2004. Mosquitoes and transmission of malaria parasites - not just vectors. Malar. J. 3:39.
- 363. Paul, R. E., T. N. Coulson, A. Raibaud, and P. T. and Brey 2000. Sex determination in malaria parasites. Science 287:128-131.
- 364. **Peixoto, L., A. Zavala, H. Romero, and H. Musto.** 2003. The strength of translational selection for codon usage varies in the three replicons of *Sinorhizobium meliloti*. Gene **320**:109-116.
- 365. Pelletier, P. J., and J. B. Caventou. 1820. Recherches chimiques sur les quinquinas. Ann. de chimie et de physique 15:289-318.
- Ponzi, M., P. Alano, R. Scotti, and L. Roca. 1993. Chromosomal polymorphism and sexual differentiation in *Plasmodium*. Parassitologia 35:87 - 89.
- 367. **Poole, E. S., C. M. Brown, and W. P. Tate.** 1995. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli*. EMBO J. **14**:151-158.

- 368. Poole, E. S., R. Brimacombe, and W. P. Tate. 1997. Decoding the translational termination signal: the polypeptide chain release factor in *Escherichia coli* crosslinks to the base following the stop codon. RNA 3:974-982.
- 369. Poole, F. L., B. A. Gerwe, R. C. Hopkins, G. J. Schut, M. V. Weinberg, F. E. Jenney Jr, and M. W. Adams. 2005. Defining genes in the genome of the hyperthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. J. Bacteriol. 187:7325-7332.
- 370. **Posner, G. H., and P. M. O'Neill.** 2004. Knowledge of the proposed chemical mechanism of action and cytochrome p450 metabolism of antimalarial trioxanes like artemisinin allows rational design of new antimalarial peroxides. Acc. Chem. Res. **37:**397-404.
- 371. Pouniotis, D. S., O. Proudfoot, V. Bogdanoska, V. Apostolopoulos, T. Fifis, and M. Plebanski. 2004. Dendritic cells induce immunity and long-lasting protection against blood-stage malaria despite an *in vitro* parasite-induced maturation defect. Infect. Immun. 72:5331-5339.
- 372. Poveda, G., W. Rojas, M. L. Quinones, I. D. Velez, R. I. Mantilla, D. Ruiz, J. S. Zuluaga, and G. L. Rua. 2001. Coupling between annual and ENSO timescales in the malaria-climate association in Colombia. Environ. Health Perspect. 109:489-493.
- 373. **Prabhu, V. V.** 1993. Symmetry observations in long nucleotide sequences. Nucleic Acids Res. **21**:2797 - 2800.
- 374. Preiser, P. R., R. J. Wilson, P. W. Moore, S. McCready, M. A. Hajibagheri, K. J. Blight, M. Strath, and D. H. Williamson. 1996. Recombination associated with replication of malarial mitochondrial DNA. EMBO J. 15:684 - 693.
- 375. Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press.
- 376. Price, R. N., A. C. Uhlemann, A. Brockman, R. McGready, E. Ashley, L. Phaipun, R. Patel, K. Laing, S. Looareesuwan, N. J. White, F. Nosten, and S. Krishna. 2004. Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number. Lancet 364 438-447.
- 377. Price, R. N., A. C. Uhlemann, A. Brockman, R. McGready, E. Ashley, L. Phaipun, R. Patel, K. Laing, S. Looareesuwan, N. J. White, F. Nosten, and S. Krishna. 2004. Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number. Lancet 364:438-447.
- 378. Puhl, H. L., S. R. Gudibande, and M. J. Behe. 1991. Poly[d(A.T)] and other synthetic polydeoxynucleotides containing oligoadenosine tracts form nucleosomes easily. J. Mol. Biol. 222:1149 - 160.
- 379. Pukrittayakamee, S., M. Imwong, S. Looareesuwan, and N. J. White. 2004. Therapeutic responses to antimalarial and antibacterial drugs in vivax malaria. Acta Trop. 89:351 - 356.
- 380. Pukrittayakamee, S., R. Clemens, A. Chantra, Nontprasert A, Luknam T, Looareesuwan S, and N. J. White 2001. Therapeutic responses to antibacterial drugs in vivax malaria. Trans. R. Soc. Trop. Med. Hyg. 95:524-548.
- 381. Qazi, K. R., M. Wikman, N. M. Vasconcelos, K. Berzins, S. Stahl, and C. Fernandez. 2005. Enhancement of DNA vaccine potency by linkage of

Plasmodium falciparum malarial antigen gene fused with a fragment of HSP70 gene. Vaccine **23:**1114-1125.

- 382. Qi, D., and A. J. Cuticchia. 2001. Compositional symmetries in complete genomes. Bioinformatics 17:557 559.
- 383. Qin, W., J. Feng, Y. Li, Z. Lin, and B. Shen. 2005. Changes of primary sequence and secondary structure proximal to the 5' end of the stop codon substantially increases the expression of the variable region of an antibody in *E. coli*. Biotechnol. Lett. 27:131-134.
- 384. **Quadrini KJ, B. J.** 2006. EKLF/KLF1 is ubiquitinated *in vivo* and its stability is regulated by activation domain sequences through the 26S proteasome. FEBS Lett. **580**:2285-2293.
- 385. Ranson, H., M. G. Paton, B. Jensen, L. McCarroll, A. Vaughan, J. R. Hogan, J. Hemingway, and F. H. Collins. 2004. Genetic mapping of genes conferring permethrin resistance in the malaria vector, *Anopheles gambiae*. Insect Mol. Biol. 13:379-386.
- 386. Rasheed, F. N., J. N. Bulmer, D. T. Dunn, C. Menendez, M. F. Jawla, A. Jepson, P. H. Jakobsen, and B. M. Greenwood. 1993. Suppressed peripheral and placental blood lymphoproliferative responses in first pregnancies: relevance to malaria. Am. J. Trop. Med. Hyg. 48:154-160.
- 387. Rathore, D., A. M. Wahl, M. Sullivan, and T. F. McCutchan. 2001. A phylogenetic comparison of gene trees constructed from plastid, mitochondrial and genomic DNA of *Plasmodium* species. Mol. Biochem. Parasitol. 114:89-94.
- 388. Rayner, J. C., E. Vargas-Serrato, C. S. Huber, M. R. Galinski, and J. W. Barnwell. 2001. A *Plasmodium falciparum* homologue of *Plasmodium vivax* reticulocyte binding protein (PvRBP1) defines a trypsin-resistant erythrocyte invasion pathway. J. Exp. Med. 194:1571-1581.
- 389. Rayner, J. C., T. M. Tran, V. Corredor, C. S. Huber, J. W. Barnwell, and M. R. Galinski. 2005. Dramatic difference in diversity between *Plasmodium falciparum* and *Plasmodium vivax* reticulocyte binding-like genes. Am. J. Trop. Med. Hyg. 72:666-674.
- 390. Read, J. A., K. W. Wilkinson, R. Tranter, R. B. Sessions, and R. L. Brady. 1999. Chloroquine binds in the cofactor binding site of *Plasmodium falciparum* lactate dehydrogenase. J. Biol. Chem. 274:10213-10218.
- 391. Rechsteiner, M., and S. W. Rogers. 1996. PEST sequences and regulation by proteolysis. Trends Biochem. Sci. 21:267-271.
- 392. Reece, W. H., M. Plebanski, P. Akinwunmi, P. Gothard, K. L. Flanagan, E. A. Lee, M. Cortina-Borja, A. V. Hill, and M. Pinder. 2002. Naturally exposed populations differ in their T1 and T2 responses to the circumsporozoite protein of *Plasmodium falciparum*. Infect. Immun. 70:1468-1474.
- Reed, R. 2003. Coupling transcription, splicing and mRNA export. Curr. Opin. Cell Biol. 15:326-331.
- 394. Reese, M. G., D. Kulp, H. Tammana, and D. Haussler. 2000. Genie gene finding in *Drosophila melanogaster*. Genome Res. 10:529-538.
- 395. Reverte, C. G., M. D. Ahearn, and L. E. Hake. 2001. CPEB degradation during *Xenopus* oocyte maturation requires a PEST domain and the 26S proteasome. Dev. Biol. 231:447-458.
- 396. Ricke, C. H., T. Staalsoe, K. Koram, B. D. Akanmori, E. M. Riley, T. G. Theander, and L. Hviid 2000. Plasma antibodies from malaria-exposed

pregnant women recognize variant surface antigens on *Plasmodium falciparum*-infected erythrocytes in a parity-dependent manner and block parasite adhesion to chondroitin sulfate A. J. Immunol. **165**:3309 - 3316.

- 397. Rieckmann, K. H., R. L. Beaudoin, J. S. Cassells, and K. W. Sell. 1979. Use of attenuated sporozoites in the immunization of human volunteers against falciparum malaria. Bull. W.H.O. **57:**261-265.
- 398. Riehle, M. A., P. Srinivasan, C. K. Moreira, and M. Jacobs-Lorena. 2003. Towards genetic manipulation of wild mosquito populations to combat malaria: advances and challenges. J. Exp. Biol. 206:3809-3816.
- 399. Rocha, E. P., A. Danchin, and A. Viari. 1999. Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res. 27:3567-3576.
- 400. Rogers, S., R. Wells, and M. Rechsteiner. 1986. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science 234:364-368.
- Rogozin, I. B., A. V. Sverdlov, V. N. Babenko, and E. V. Koonin. 2005. Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform. 6:118-134.
- 402. **Rojas, F. J., and I. Moretti-Rojas.** 2000. Involvement of the calcium-specific protease, calpain, in the fertilizing capacity of human spermatozoa. Int. J. Androl. **23:**163-168.
- 403. **Romanowsky, D. L.** 1891. Zur frage der Parasitologie und Therapie der Malaria. St. Petersb. Med. Wschr. **16**:297-307.
- 404. **Romero, H., A. Zavala, and H. Musto.** 2000. Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica*. Gene **242**:307-311.
- 405. Romero, H., A. Zavala, H. Musto, and G. Bernardi. 2003. The influence of translational selection on codon usage in fishes from the family *Cyprinidae*. Gene **317**:141-147.
- 406. Roper, C., R. Pearce, S. Nair, B. Sharp, F. Nosten, and T. Anderson. 2004. Intercontinental spread of pyrimethamine-resistant malaria. Science **305**:1124.
- 407. Rosen, J. B., and J. G. Breman. 2004. Malaria intermittent preventive treatment in infants, chemoprophylaxis, and childhood vaccinations. Lancet 363:1386 1388.
- 408. **Rosenthal, P. J. a. N., R. G.** 1992. Isolation and characterization of a cysteine proteinase gene of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **51**:143–152.
- 409. Ross, R. 1911. The Prevention of Malaria. John Murray, London.
- 410. **Ross, R.** 1898. Report on the cultivation of Proteosoma, Labbe, in grey mosquitoes. Government Press, Calcutta.
- 411. **Roth, E. J.** 1990. *Plasmodium falciparum* carbohydrate metabolism: a connection between host cell and parasite. Blood Cells **16**:453-460.
- 412. Rowland, M., G. Downey, A. Rab, T. Freeman, N. Mohammad, H. Rehman, N. Durrani, H. Reyburn, C. Curtis, J. Lines, and M. Fayaz. 2004. DEET mosquito repellent provides personal protection against malaria: a household randomized trial in an Afghan refugee camp in Pakistan. Trop. Med. Int. Health 9:335-342.
- 413. Rudner, R., J. D. Karkas, and E. Chargaff. 1968. Separation of *B. subtilis* DNA into complementary strands. Proc. Natl. Acad. Sci. USA 60:921 922.

- 414. Ruvinsky, A., S. T. Eskesen, F. N. Eskesen, and L. D. Hurst. 2005. Can codon usage bias explain intron phase distributions and exon symmetry. J. Mol. Evol. 60:99-104.
- 415. Ryan, E. T., and K. C. Kain. 2000. Health advice and immunizations for travelers. N. Engl. J. Med. 342.
- 416. Sabchareon, A., T. Burnouf, D. Ouattara, P. Attanath, H. Bouharoun-Tayoun, P. Chantavanich, C. Foucault, T. Chongsuphajaisiddhi, and P. Druilhe. 1991. Parasitologic and clinical human response to immunoglobulin administration in falciparum malaria. Am. J. Trop. Med. Hyg. 45:297-308.
- 417. Sachs, J., and P. Malaney. 2002. The economic and social burden of malaria. Nature 415:680 - 685.
- 418. Sahu, K., S. K. Gupta, T. C. Ghosh, and S. Sau. 2004. Synonymous codon usage analysis of the mycobacteriophage Bxz1 and its plating bacteria *M. smegmatis*: identification of highly and lowly expressed genes of Bxz1 and the possible function of its tRNA species. Biochem. Mol. Biol. **37:**487-492.
- 419. Sallares, R., and s. Gomzi. 2001. Biomolecular archaeology of malaria. Ancient Biomolecules 3:195-213.
- 420. Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26:544-548.
- 421. Sam-Yellowe, T. Y. 1996. Rhoptry organelles of the apicomplexa: Their role in host cell invasion and intracellular survival. Parasitol. Today 12:308-316.
- 422. Sam-Yellowe, T. Y., L. Florens, J. R. Johnson, *et al.* 2004. A *Plasmodium* gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling. Genome Res. **14**:1052-1059.
- 423. Satchwell, S. C., H. R. Drew, and A. A. Travers. 1986. Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol. 191:659 675.
- 424. Saul, A., and D. Battistutta. 1990. Analysis of the sequences flanking the translational start sites of *Plasmodium falciparum*. Mol. Biochem. Parasitol. 42:55-62.
- 425. Schad, E., A. Farkas, G. Jekely, P. Tompa, and P. Friedrich. 2002. A novel human small subunit of calpains. Biochem. J. 362:383-388.
- 426. Schadt, E. E., S. W. Edwards, D. Guhathakurta, *et al.* 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol. **5:**R73.
- 427. Schechter, I., and A. Berger. 1967. On the size of the active site in proteases. I. Papain. Biochem. Biophys. Res. Commun. 27:157-162.
- 428. Schlueter, S. D., M. D. Wilkerson, E. Huala, S. Y. Rhee, and V. Brendel. 2005. Community-based gene structure annotation. Trends Plant Sci. 10:9-14.
- 429. Schneider, T. D. 1997. Information content of individual genetic sequences. J. Theor. Biol. 189:427-441.
- 430. Schneider, T. D., and R. M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18:6097-6100.
- Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht. 1986. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188:415-431.
- 432. Schwab, S. R., J. A. Shugart, T. Horng, S. Malarkannan, and N. Shastri. 2004. Unanticipated antigens: translation initiation at CUG with leucine. PLoS Biol. 2:e366.

- 433. Seligmann, H., and D. Pollock. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene feading. DNA Cell Biol. 23:701-705.
- 434. Shah, T., E. de Villiers, V. Nene, B. Hass, E. Taracha, M. J. Gardner, C. Sansom, R. Pelle, and R. Bishop. 2006. Using the transcriptome to annotate the genome revisited: Application of massively parallel signature sequencing (MPSS). Gene 366:104-108.
- 435. Sharma, S. 1996. Applied multivariate techniques. John Wiley and Sons.
- 436. Sharp, P. M., and K. M. Devine. 1989. Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res. 17:5029-5039.
- 437. Sharp, P. M., and M. Bulmer. 1988. Selective differences among translation termination codons. Gene 63:141-145.
- 438. Sharp, P. M., and M. Bulmer. 1988. Selective differences among translation termination codons Gene 63:141-145.
- 439. Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden, and R. E. Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33:1141-1153.
- 440. Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. 16:8207-8211.
- 441. Sharp, P. M., T. M. F. Tuohy, and K. R.Mosurski. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.
- Shenai, B. R., P. S. Sijwali., A. Singh, and P. J. Rosenthal. 2000.
 Characterization of native and recombinant falcipain-2 a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. J. Biol. Chem. 275:29000–29010.
- 443. Shields, D. C., and P. M. Sharp. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res. 15:8023-8040.
- 444. Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature **407**:81-86.
- 445. Shortt, H. E., N. H. Fairley, G. Covell, P. G. Shute, and P. C. C. Garnham 1949. The pre-erythrocytic stage of *Plasmodium falciparum*. A preliminary note. BMJ 2:1006-1008.
- 446. Shortt, H. E., P. C. C. Garnham, G. Covell, and P. G. Shute. 1948. The pre-erythrocytic stage of human malaria, *Plasmodium vivax*. BMJ 1:547.
- 447. Shumway S. D., M. M., and S. Miyamoto. 1999. The PEST domain of Ikappa B alpha is necessary and sufficient for *in vitro* degradation by mucalpain J. Biol. Chem. 274:30874-30881.
- 448. Siddique, M. E., and S. Ahmed. 1995. Serum complement C4 levels during acute malarial infection and post-treatment period. Indian J. Pathol. Microbiol. 38:335-339.
- Siddiqui, W. A., J. V. Schnell, and S. Richmond-Crum. 1974. In vitro cultivation of *Plasmodium falciparum* at high parasitemia. Am. J. Trop. Med. Hyg. 23:1015 - 1018.

- 450. Sijwali, P. S., B. R. Shenai, J. Gut, A. Singh, and P. J. Rosenthal. 2001. Expression and characterization of the *Plasmodium falciparum* haemoglobinase falcipain-3. Biochem. J. 360:481–489.
- 451. Sims, A. H., M. E. Gent, G. D. Robson, N. S. Dunn-Coleman, and S. G. Oliver. 2004. Combining transcriptome data with genomic and cDNA sequence alignments to make confident functional assignments for *Aspergillus nidulans* genes. Mycol. Res. **108**:853-857.
- 452. Sinden, R. E., and R. H. Hartley. 1985. Identification of the meiotic division of malarial parasites. J. Protozool. 32:742-744.
- 453. Singh, B., L. Kim Sung, A. Matusop, A. Radhakrishnan, S. S. Shamsul, J. Cox-Singh, A. Thomas, and D. J. Conway. 2004. A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. Lancet 363:1017-1024.
- 454. Singh, B., L. Kim Sung, A. Matusop, A. Radhakrishnan, S. S. Shamsul, J. Cox-Singh, A. Thomas, and D. J. Conway. 2004. A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. Lancet 363:1017 1024.
- 455. Singh, D., A. Kumar, E. V. Raghu Ram, and S. Habib. 2005. Multiple replication origins within the inverted repeat region of the *Plasmodium falciparum* apicoplast genome are differentially activated. Mol. Biochem. Parasitol. **139**:99 106.
- 456. Singh, N., O. Kataria, and M. P. Singh. 2004. The changing dynamics of *Plasmodium vivax* and *P. falciparum* in central India: trends over a 27-year period (1975-2002). Vector Borne Zoonotic Dis. **4**:239-248.
- 457. Singh, N., P. Preiser, L. Renia, B. Balu, J. Barnwell, P. Blair, W. Jarra, T. Voza, I. Landau, and J. H. Adams. 2004. Conservation and developmental control of alternative splicing in *maebl* among malaria parasites. J. Mol. Biol. 343:589-599.
- 458. Sinton, J. A. 1935. What malaria costs India, nationally, socially and economically. Rec. Mal. Surv. India 5:223 264.
- 459. Sinton, J. A. a. H. W. M. 1933. A critical review of the literature relating to the identification of the malarial parasites recorded from monkeys of the families *Cercopithecidae* and *Colobidae*. Rec. Malar. Surv. India III:381-443.
- 460. Skorokhod, O. A., M. Alessio, B. Mordmuller, P. Arese, and E. Schwarzer. 2004. Hemozoin (malarial pigment) inhibits differentiation and maturation of human monocyte-derived dendritic cells: a peroxisome proliferator-activated receptor-gamma-mediated effect. J. Immunol. 173:4066-4074.
- 461. **Slater, L. B.** 2004. Malaria chemotherapy and the "kaleidoscopic" organisation of biomedical research during world war II. Ambix. **51**:107-134.
- 462. Smith, E., T. E. Meyerrose, T. Kohler, M. Namdar-Attar, N. Bab, O. Lahat, T. Noh, J. Li, M. W. Karaman, J. G. Hacia, T. T. Chen, J. A. Nolta, R. Muller, I. Bab, and B. Frenkel. 2005. Leaky ribosomal scanning in mammalian genomes: significance of histone H4 alternative translation *in vivo*. Nucleic Acids Res. 33:1298-1308.
- 463. Smithuis, F., M. Shahmanesh, M. K. Kyaw, O. Savran, S. Lwin, and N. J. White. 2004. Comparison of chloroquine, sulfadoxine/pyrimethamine, mefloquine and mefloquine-artesunate for the treatment of falciparum malaria in Kachin State, North Myanmar. Trop. Med. Int. Health 9:1184-1190.

- 464. Snow, R. W., C. A. Guerra, A. M. Noor, H. Y. Myint, and S. I. Hay. 2005. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. Nature **434**:214-217.
- 465. Snow, R. W., M. H. Craig, U. Deichmann, and K. Marsh. 1999. Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. Bull. World Health Organ. 77:624 640.
- 466. Snow, R. W., M. H. Craig, U. Deichmann, D. le Sueur. 1999. A preliminary continental risk map for malaria mortality among African children. Parasitol. Today 15:99 - 104.
- 467. Sokhna, C., J. Y. Le Hesran, P. A. Mbaye, J. Akiana, P. Camara, M. Diop, A. Ly, and P. Druilhe. 2004. Increase of malaria attacks among children presenting concomitant infection by *Schistosoma mansoni* in Senegal. Malar. J. 3:43.
- 468. Song, H., P. Mugnier, H. M. Webb, D. R. Evans, M. F. Tuite, B. A. Hemmings, and D. Barford. 2000. The crystal structure of human eukaryotic release factor eRF1 – mechanism of stop codon recognition and peptidyltRNA hydrolysis. Cell 100:311-321.
- 469. **Sowunmi, A.** 1998. Efficacy of chloroquine plus chlorpheniramine in chloroquine-resistant falciparum malaria during pregnancy in Nigerian women: a preliminary study. J. Obstet. Gynaecol. **18:**524-527.
- 470. **Spearman, C.** 1904. General intelligence, objectively determined and measured. Am. J. Psych. **15**:201-293.
- 471. **Spencer, M. L., M. Theodosiou, and D. J. Noonan.** 2004. NPDC-1, a novel regulator of neuronal proliferation, is degraded by the ubiquitin/proteasome system through a PEST degradation motif. J Biol Chem. **279:**37069-37078.
- 472. Staedke, S. G., A. Mpimbaza, M. R. Kamya, B. K. Nzarubara, G. Dorsey, and P. J. Rosenthal 2004. Combination treatments for uncomplicated falciparum malaria in Kampala, Uganda: randomised clinical trial. Lancet 364:1950-1957.
- 473. Stahl, H. D., D. J. Kemp, P. E. Crewther, D. B. Scanlon, G. Woodrow, G. V. Brown, A. E. Bianco, R. F. Anders, and R. L. Coppel. 1985. Sequence of a cDNA encoding a small polymorphic histidine- and alanine-rich protein from *Plasmodium falciparum*. Nucleic Acids Res. 13:7837-7846.
- 474. **Stenico, M., A. T. Lloyd, and P. M. Sharp.** 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. **22:**2437–2446.
- 475. **Stenstrom, C. M., H. Jin, L. L. Major, W. P. Tate, and L. A. Isaksson.** 2001. Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. Gene **263**:273-284.
- 476. **Stephens, J. W. W.** 1922. A new malaria parasite of man. Ann. Trop. Med. Parasitol. **16**:383-386.
- 477. **Stephens, R. M., and T. D. Schneider.** 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J. Mol. Biol. **228**:1124-1136.
- 478. Stepniewska, K., W. R. Taylor, M. Mayxay, R. Price, F. Smithuis, J. P. Guthmann, K. Barnes, H. Y. Myint, M. Adjuik, P. Olliaro, S. Pukrittayakamee, S. Looareesuwan, T. T. Hien, J. Farrar, F. Nosten, N. P. Day, and N. J. White. 2004. *In vivo* assessment of drug efficacy against *Plasmodium falciparum* malaria: Duration of follow-up. Antimicrob. Agents Chemother. 48:4271-4280.

- 479. Stibitz, S., W. Aaronson, D. Monack, and S. Falkow. 1989. Phase variation in *Bordetella pertussis* by frameshift mutation in a gene for a novel two-component system. Nature. **338:**266 269.
- 480. Stivanello, E., P. Cavailler, F. Cassano, S. A. Omar, D. Kariuki, J. Mwangi, P. Piola, and J. P. Guthmann. 2004. Efficacy of chloroquine, sulphadoxine-pyrimethamine and amodiaquine for treatment of uncomplicated *Plasmodium falciparum* malaria in Kajo Keji county, Sudan. Trop. Med. Int. Health 9:975-980.
- 481. Stohrer, J. M., S. Dittrich, V. Thongpaseuth, V. Vanisaveth, R. Phetsouvanh, S. Phompida, F. Monti, E. M. Christophel, N. Lindegardh, A. Annerberg, and T. Jelinek. 2004. Therapeutic efficacy of artemether-lumefantrine and artesunate-mefloquine for treatment of uncomplicated *Plasmodium falciparum* malaria in Luang Namtha Province, Lao People's Democratic Republic. Trop. Med. Int. Health 9:1175-1183.
- 482. Stolc, V., Z. Gauhar, C. Mason, et al. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science **306**:655-660.
- 483. **Stormo, G. D., T. D. Schneider and L. Gold.** 1986. Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucleic Acids Res. **14**:6661-6679.
- 484. Streisinger, G., Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi, and M. Inouye 1966. Frameshift mutations and the genetic code. Cold Spring Harb. Symp. Quant. Biol. **31:**77 - 84.
- 485. Strobl, S., C. Fernandez-Catalan, M. Braun, et al. 2000. The crystal structure of calcium-free human m-calpain suggests an electrostatic switch mechanism for activation by calcium. Proc. Natl. Acad. Sci. U S A 97:588-592.
- 486. Struik, S. S., and E. M. Riley. 2004. Does malaria suffer from lack of memory? Immunol. Rev. 201.
- 487. Struik, S. S., F. M. Omer, K. Artavanis-Tsakonas, and E. M. Riley. 2004. Uninfected erythrocytes inhibit *Plasmodium falciparum*-induced cellular immune responses in whole-blood assays. Blood **103**:3084-3092.
- 488. Sugnet, C. W., W. J. Kent, M. Ares Jr., and D. Haussler. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac. Symp. Biocomput.:66-77.
- 489. Suter, B., G. Schnappauf, and F. Thoma. 2000. Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters *in vivo*. Nucleic Acids Res. 28:4083 - 4089.
- 490. Swed, F. A., and C. Eisenharth. 1943. Tables for testing randomness of grouping in a sequence of alternatives. Am. Math. Statist. 14.
- 491. Takahashi, K., M. Maruyama, Y. Tokuzawa, M. Murakami, Y. Oda, N. Yoshikane, K. W. Makabe, T. Ichisaka, and S. Yamanaka. 2005. Evolutionarily conserved non-AUG translation initiation in NAT1/p97/DAP5 (EIF4G2). Genomics 85:360-371.
- 492. Talisuna, A. O., A. Nalunkuma-Kazibwe, P. Langi, T. K. Mutabingwa, W. W. Watkins, E. V. Marck, T. G. Egwang, and U. D'Alessandro. 2004. Two mutations in dihydrofolate reductase combined with one in the dihydropteroate synthase gene predict sulphadoxine-pyrimethamine parasitological failure in Ugandan children with uncomplicated falciparum malaria. Infect. Genet. Evol. 4:321-327.

- 493. **Tanabe K, A. I., M. Kato, A. Miki, and S. Doi.** 1989. Stage-dependent inhibition of *Plasmodium falciparum* by potent Ca2+ and calmodulin modulators. J. Protozool. **36:**139-143.
- 494. Tang. H. L., L. S. Y., N. K. Chen, T. Ripmaster, P. Schimmel, and C. C. Wang. 2004. Translation of a yeast mitochondrial tRNA synthetase initiated at redundant non-AUG codons. J. Biol. Chem. 279:49656-49663.
- 495. **Tate, W. P., and S. A. Mannering.** 1996. Three, four or more: the translational stop signal at length. Mol. Microbiol. **21**:213-219.
- 496. **Tate, W. P., E. S. Poole, M. E. Dalphin, L. L. Major, D. J. Crawford, and S. A. Mannering.** 1996. The translational stop signal: codon with a context, or extended factor recognition element? Biochimie **78**:945-952.
- 497. **Taylor, K., Z. Hradecna, and W. Szybalski.** 1967. Asymmetric distribution of the transcribing regions on the complementary strands of the coliphage λ DNA. Proc. Natl. Acad. Sci. USA **57**:1618-1625.
- 498. **Taylor, P. R., E. Seixas, M. J. Walport, J. Langhorne, and M. Botto.** 2001. Complement contributes to protective immunity against reinfection by *Plasmodium chabaudi* chabaudi parasites. Infect. Immun. **69**:3853-3859.
- 499. Tchavtchitch, M., K. Fischer, R. Huestis, and A. Saul. 2001. The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. Mol. Biochem. Parasitol. 118:211-222.
- 500. Tenney, A. E., R. H. Brown, C. Vaske, J. K. Lodge, T. L. Doering, and M. R. Brent. 2004. Gene prediction and verification in a compact genome with numerous small introns. Genome Res. 14:2330-2335.
- 501. **Tenson, T., M. Lovmar, and M. Ehrenberg.** 2003. The mechanism of action of macrolides, lincosamides and streptogramin B reveals the nascent peptide exit path in the ribosome. J. Mol. Biol. **330**:1005-1014.
- 502. Thijs, G., M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17:1113-1122.
- 503. Todd, B., D. Moore, C. C. Deivanayagam, G. Lin, D. Chattopadhyay, M. Maki, K. K. Wang, and S. V. Narayana. 2003. A structural model for the inhibition of calpain by calpastatin: crystal structures of the native domain VI of calpain and its complexes with calpastatin peptide and a small molecule inhibitor. J. Mol. Biol. 328.
- 504. **Tong, W.** 2004. Analyzing the biology on the system level. Genomics Proteomics Bioinformatics **2:**6-14.
- 505. **Torti, F.** 1775. La terapia speciale delle febbri perniciose. Roma Casa Luigi Possi.
- 506. **Trager, W., and J. B. Jensen.** 1976. Human malaria parasites in continuous culture. Science **193:**673 675.
- 507. Trenholme, C. M., R. L. Williams, R. E. Desjardins, H. Frischer, P. E. Carson, K. H. Rieckmann, and C. J. Canfield. 1975. Mefloquine (WR 142,490) in the treatment of human malaria. Science 190:792-794.
- 508. Tribus, M., and E. C. McIrvine. 1971. Energy and information. Sci. Am. 225:179-188.
- 509. **Trotta, R. F., M. L. Brown, J. C. Terrell, and J. A. Geyer.** 2004. Defective DNA repair as a potential mechanism for the rapid development of drug resistance in *Plasmodium falciparum*. Biochemistry **43**:4885-4891.

- 510. Urban, B. C., D. J. Ferguson, A. Pain, N. Willcox, M. Plebanski, J. M. Austyn, and D. J. Roberts. 1999. *Plasmodium falciparum* infected erythrocytes modulate the maturation of dendritic cells. Nature **400**:73-77.
- Urban, B. C., N. Willcox, and D. J. Roberts. 2001. A role for CD36 in the regulation of dendritic cell function. Proc. Natl. Acad. Sci. U S A. 98:8750-8755.
- 512. Urban, B. C., T. Mwangi, A. Ross, S. Kinyanjui, M. Mosobo, O. Kai, B. Lowe, K. Marsh, and D. J. Roberts. 2001. Peripheral blood dendritic cells in children with acute *Plasmodium falciparum* malaria. Blood **98**:2859-2861.
- 513. Uzureau, P., J. C. Barale, C. J. Janse, A. P. Waters, and C. B. Breton. 2004. Gene targeting demonstrates that the *Plasmodium berghei* subtilisin PbSUB2 is essential for red cell invasion and reveals spontaneous genetic recombination events. Cell Microbiol. 6:65-78.
- 514. Valasek, L., K. H. Nielsen, F. Zhang, C. A. Fekete, and A. G. Hinnebusch. 2004. Interactions of eukaryotic translation initiation factor 3 (eIF3) subunit NIP1/c with eIF1 and eIF5 promote preinitiation complex assembly and regulate start codon selection. Mol. Cell Biol. 24:9437-9455.
- 515. van Lin, L. H., T. Pace, C. J. Janse, C. Birago, J. Ramesar, L. Picci, M. Ponzi, and A. P. Waters. 2001. Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. Nucleic Acids Res. **29**:2059-2068.
- 516. van Lin, L. H., T. Pace, C. J. Janse, C. Birago, J. Ramesar, L. Picci, M. Ponzi, and A. P. Waters. 2001 Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. Nucleic Acids Res. 29:2059-2068.
- 517. Vanparys, B., P. Bodelier, and P. De Vos. 2006. Validation of the correct start codon of norX/nxrX and universality of the norAXB/nxrAXB gene cluster in Nitrobacter species. Curr. Microbiol. **53**:255-257.
- 518. Vennerstrom, J. L., S. Arbe-Barnes, R. Brun, S. A. Charman, F. C. Chiu, J. Chollet, Y. Dong, A. Dorn, D. Hunziker, H. Matile, K. McIntosh, M. Padmanilayam, J. Santo Tomas, C. Scheurer, B. Scorneaux, Y. Tang, H. Urwyler, S. Wittlin, and W. M. Charman. 2004. Identification of an antimalarial synthetic trioxolane drug development candidate. Nature 430:900-904.
- 519. Volkman, S. K., A. E. Barry, E. J. Lyons, K. M. Nielsen, S. M. Thomas, M. Choi, S. S. Thakore, K. P. Day, D. F. Wirth, and D. L. Hartl. 2001. Recent origin of *Plasmodium falciparum* from a single progenitor. Science 293:482-484.
- 520. Volkman, S. K., D. L. Hartl, D. F. Wirth, *et al.* 2002. Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. Science **298**:216-218.
- 521. Vreugdenhil, C. J., F. Y. Scheper, S. R. Hoogstraatte, M. Smolders, S. Gikunda, F. G. Cobelens, and P. A. Kager. 2004. Comparison of the parasitologic efficacy of amodiaquine and sulfadoxine-pyrimethamine in the treatment of *Plasmodium falciparum* malaria in the Bungoma district of Western Kenya. Am. J. Trop. Med. Hyg. 71:537-541.
- 522. Vysochanskiï, D. F., and Y. I. Petunin. 1980. Justification of the 3s rule for unimodal distributions. Theory Prob. Math. Stat. 21:25-36.

- 523. Wahl, M. B., U. Heinzmann, and K. Imai. 2005. LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. Bioinformatics **21**:1393-1400.
- 524. Wang, N., W. Chen, P. Linsel-Nitschke, L. O. Martinez, B. Agerholm-Larsen, D. L. Silver., and A. R. Tall. 2003. A PEST sequence in ABCA1 regulates degradation by calpain protease and stabilization of ABCA1 by apoA-I. J. Clin. Invest. 111:99-107.
- 525. Wang, R., D. L. Doolan, Y. Charoenvit, R. C. Hedstrom, M. J. Gardner, P. Hobart, J. Tine, M. Sedegah, V. Fallarme, J. B. Sacci Jr., M. Kaur, D. M. Klinman, S. L. Hoffman, and W. R. Weiss. 1998. Simultaneous induction of multiple antigen-specific cytotoxic T lymphocytes in nonhuman primates by immunization with a mixture of four *Plasmodium falciparum* DNA plasmids. Infect. Immun. 66:4193–4202.
- 526. Warburg, A., and I. Schneider. 1993. *In vitro* culture of the mosquito stages of *Plasmodium falciparum*. Exp. Parasitol. **76**:121 126.
- 527. Warhurst, D. C., J. C. Craig, I. S. Adagu, D. J. Meyer, and S. Y. Lee. 2003. The relationship of physico-chemical properties and structure to the differential antiplasmodial activity of the cinchona alkaloids. Malar. J. 2:26.
- 528. Wasserman, M., and J. Chaparro. 1996. Intraerythrocytic calcium chelators inhibit the invasion of *Plasmodium falciparum* Parasitol. Res. 82:102-107.
- 529. Wasserman, M., C. Alarcon, and P. M. Mendoza. 1982. Effects of Ca⁺⁺ depletion on the asexual cell cycle of *Plasmodium falciparum*. Am. J. Trop. Med. Hyg. 31:711-717.
- 530. Waters, A. P., D. G. Higgins, and T. F. McCutchan. 1993. Evolutionary relatedness of some primate models of *Plasmodium*. Mol Biol Evol. 10:914 -923.
- 531. Watkins, W. M., A. D. Brandling-Bennett, C. G. Nevill, J. Y. Carter, D. A. Boriga, R. E. Howells, and D. K. Koech. 1988. Chlorproguanil/dapsone for the treatment of non-severe *Plasmodium falciparum* malaria in Kenya: a pilot study. Trans. R. Soc. Trop. Med. Hyg. 82:398-403.
- 532. Weber, J. L. 1987. Analysis of sequences from the extremely A + T-rich genome of *Plasmodium falciparum*. Gene **52**:103-109.
- 533. Wegmann, T. G., H. Lin, L. Guilbert, and T. R. Osmann. 1993. Bidirectional cytokine interactions in the maternal-fetal relationship: is successful pregnancy a T_H2 phenomenon? Immunol. Today 14:353-356.
- Weiser, J. N., J. M. Love, and E. R. Moxon. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. Cell 59:657 - 665.
- 535. Wellems, T. E. 2004. Transporter of a malaria catastrophe. Nat. Med. 10:1169-1171.
- 536. Wernsdorfer, G., and W. H. Wernsdorfer. 2003. Malaria at the turn from the 2nd to the 3rd millenium. Wien Klin Wochenschr. 115:2-9.
- 537. White J. H., a. B. J. K. 1996. DNA replication in the malaria parasite. Parasitol. Today 12:151 - 155.
- 538. White, N. 1999. Antimalarial drug resistance and combination chemotherapy. Philos. Trans. R. Soc. Lond. B Biol. Sci. **354**:739-749.
- 539. Wichmann, O., M. Muehlen, H. Gruss, F. P. Mockenhaupt, N. Suttorp, and T. Jelinek. 2004. Malarone treatment failure not associated with previously described mutations in the cytochrome b gene. Malar. J. **3**:14.

- 540. Wiesner, J., D. Henschker, D. B. Hutchinson, E. Beck, and H. Jomaa. 2002. *In vitro* and *in vivo* synergy of fosmidomycin, a novel antimalarial drug, with clindamycin. Antimicrob. Agents Chemother. **46**:2889-2894.
- Willcox, M. L., and G. Bodeker. 2004. Traditional herbal medicines for malaria. BMJ 329:1156-1159.
- 542. Willems, R., A. Paul, H. G. van der Heide, A. R. ter Avest, and F. R. Mooi. 1990. Fimbrial phase variation in *Bordetella pertussis*: a novel mechanism for transcriptional regulation. EMBO J. 9:2803 2809.
- 543. Williamson, D. H., P. R. Preiser, and R. J. Wilson. 1996. Organelle DNAs: The bit players in malaria parasite DNA replication. Parasitol. Today 12:357 -362.
- 544. Williamson, D. H., P. R. Preiser, P. W. Moore, S. McCready, M. Strath, and R. J. Wilson. 2002. The plastid DNA of the malaria parasite *Plasmodium falciparum* is replicated by two mechanisms. Mol. Microbiol. **45**:533 - 542.
- 545. Williamson, L. M., S. Lowe, E. M. Love, H. Cohen, K. Soldan, DB McClelland, P. Skacel, and J. A. Barbara. 1999. Serious hazards of transfusion (SHOT) initiative: analysis of the first two annual reports. BMJ 319:16-19.
- 546. Wilson, P. E., W. Kazadi, D. D. Kamwendo, V. Mwapasa, A. Purfield, and S. R. Meshnick. 2005. Prevalence of pfcrt mutations in Congolese and Malawian *Plasmodium falciparum* isolates as determined by a new Taqman assay. Acta Trop. 93:97-106.
- 547. Wilson, R. J., P. W. Denny, P. R. Preiser, K. Rangachari, K. Roberts, A. Roy, A. Whyte, M. Strath, D. J. Moore, P. W. Moore, and D. H. Williamson. 1996. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. J. Mol. Biol. 261:155-172.
- 548. Winter, E., and A. Varshavsky. 1989. A DNA binding protein that recognizes oligo(dA).oligo(dT) tracts. EMBO J. 8:1867-1877.
- 549. Winterberg, D. H., P. C. Wever, C. van Rheenen-Verberg, O. Kempers, R. Durand, A. P. Bos, A. H. Teeuw, L. Spanjaard, and J. Dankert. 2005. A boy with nosocomial malaria tropica contracted in a Dutch hospital. Pediatr. Infect. Dis. J. 24:89-91.
- 550. Wiseman, V., W. A. Hawley, F. O. ter Kuile, P. A. Phillips-Howard, J. M. Vulue, B. L. Nahlen, and A. J. Mills. 2003. The cost-effectiveness of permethrin-treated bed nets in an area of intense malaria transmission in western Kenya. Am. J. Trop. Med. Hyg. 68:161-167.
- 551. Wright, F. 1990. The 'effective number of codons' used in a gene. Gene 87:23-29.
- 552. Wu, Y., L. A. Kirkman, and T. E. Wellems. 1996. Transformation of *Plasmodium falciparum* malaria parasites by homologous integration of plasmids that confer resistance to pyrimethamine. Proc. Natl. Acad. Sci. U S A 93:1130 - 1134.
- 553. Wu, Y., X. Wang, X. Liu, and Y. Wang. 2003. Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. Genome Res. 13:601-616.
- 554. Wuitschick, J. D., and K. M. Karrer. 1999. Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in *Tetrahymena thermophila*. J. Eukaryot. Microbiol. **46:**239-247.
- 555. Xu, X., M. Peng, and Z. Fang 2000. The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. 24:396 399.

- 556. Yeh, I., T. Hanekamp, S. Tsoka, P. D. Karp, and R. B. Altman. 2004. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. Genome Res. 14:917-924.
- 557. Ylostalo, J., A. C. Randall, T. A. Myers, M. Metzger, D. J. Krogstad, and F. B. Cogswell. 2005. Transcriptome profiles of host gene expression in a monkey model of human malaria. J Infect Dis. 191:400-409.
- 558. Zavala, A., H. Naya, H. Romero, and H. Musto. 2002. Trends in codon and amino acid usage in *Thermotoga maritime*. J. Mol. Evol. 54:563-568.
- 559. **Zhang, D. L., L. Ji, and Y. D. Li.** 2004. Analysis, identification and correction of some errors of model refseqs appeared in NCBI Human Gene Database by *in silico* cloning and experimental verification of novel human genes. Yi Chuan Xue Bao **31**:431-443.
- 560. **Zhang, H., M. Paguio, and P. D. Roepe.** 2004. The antimalarial drug resistance protein *Plasmodium falciparum* chloroquine resistance transporter binds chloroquine. Biochemistry **43**:8290-8296.
- 561. **Zhou, Y., J. W. Bizzaro, and K. A. Marx** 2004. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. BMC Genomics **5:**95.
- 562. Zhou, Z., L. J. Licklider, S. P. Gygi, and R. Reed. 2002. Comprehensive proteomic analysis of the human spliceosome. Nature 419:182-185.
- Zhu, Z., and D. J. Thiele. 1996. A specialized nucleosome modulates transcription factor access to a *C. glabrata* metal responsive promoter. Cell 87:459 - 470.