Bioinformatics approaches to malaria

Hansen, Daniel Aaen

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Bioinformatics approaches to malaria

– PhD Thesis –

Daniel Aaen Hansen

Center for Biological Sequence analysis
Department of Systems Biology
Technical University of Denmark

May 6, 2011

Preface

This PhD thesis was prepared at the Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark under the supervision of Anders Gorm Pedersen. The work was funded by a grant from the Technical University of Denmark.

Daniel Aaen Hansen
Kgs. Lyngby, May 2011

Contents

Abstract

Bioinformatics approaches to malaria

Malaria is a life threatening disease found in tropical and subtropical regions of the world. Each year it kills 781 000 individuals; most of them are children under the age of five in sub-Saharan Africa. The most severe form of malaria in humans is caused by the parasite *Plasmodium falciparum*, which is the subject of the first part of this thesis.

The PfEMP1 protein which is encoded by the highly variable *var* gene family is important in the pathogenesis and immune evasion of malaria parasites. We analyzed and classified these genes based on the upstream sequence in seven *Plasmodium falciparum* clones. We show that the amount of nucleotide diversity is just as big within each clone as it is between the clones.

DNA methylation is an important epigenetic mark in many eukaryotic species. We are studying DNA methylation in the malaria parasite *Plasmodium falciparum*. The work is still in progress and will be introduced here.

One of the biggest concerns regarding the treatment of malaria is the continued development of resistance to existing drugs. Therefore, new drugs will be needed in the future. The ApiAP2 proteins are a recently discovered family of putative transcription factors. As they might perform important regulatory functions in the parasite, they could be useful as drug targets. Here, we study one of these proteins and describe our work on identifying small compounds that can interfere with its DNA binding abilities.

Specific binding of short peptides by proteins of the major histocompatibility complex (MHC) is an important event in the activation of immune responses to various pathogens. The set of peptides that can bind a specific MHC molecule can be characterized by a binding motif. In the second part of this thesis, we developed an algorithm that can distinguish several binding motifs within a mixture of peptides from different motifs.

Dansk resumé

Bioinformatiske tilgange til malaria

Malaria er en livstruende sygdom som findes i tropiske og subtropiske egne. Hvert år dør 781 000 individer som følge af sygdommen; de fleste af dem er børn under fem års alderen i Afrika syd for Sahara. Den alvorligste form for malaria i mennesker er forårsaget af parasitten *Plasmodium falciparum*, der er emnet for første del af denne afhandling.

PfEMP1 proteinet, der kodes for af den meget variable *var* gen familie er vigtig for parasittens sygdomsfremkaldende egenskaber og dens evne til at undvige immunforsvaret. Vi analyserede og klassificerede disse gener baseret på deres opstrøms sekvens i syv *Plasmodium falciparum* kloner. Vi viser at mængden af nukleotid diversitet er lige så stort inden for hver klon som den er mellem klonerne.

DNA methylering er en vigtig epigenetisk markør i mange eukaryote arter. Vi studerer DNA methylering i malaria parasitten *Plasmodium falciparum*. Dette arbejde er stadig i gang og vil blive introduceret her.

En af de største bekymringer vedrørende behandlingen af malaria er den kontinuerte udvikling af resistens over for de eksisterende lægemidler. Derfor vil der blive brug for nye lægemidler i fremtiden. ApiAP2 proteinerne er en nyligt opdaget familie af mulige transkriptions-faktorer. Da de måske har vigtige regulatoriske funktioner i parasitten, kan de være brugbare mål for nye lægemidler. Vi studerer et af disse proteiner og beskriver vores arbejde med at finde små stoffer, der kan påvirke dens evne til at binde DNA.

MHC proteiners specifikke binding af korte peptider er en vigtig led i aktiveringen af et immun respons mod forskellige pathogener. Et sæt af peptider, der kan binde et specifikt MHC molekyle kan karakteriseres ved et bindings motiv. I anden del af denne afhandling udvikler vi en algoritme, der kan skelne mellem adskillige bindings motiver i en blanding af peptider fra forskellige motiver.

Acknowledgements

A lot of people deserves to be mentioned here for the various ways they contributed to this thesis.

First and foremost my supervisor, Anders Gorm Pedersen for his supervision, inspiration and encouragement during the last three years.

Morten Nielsen for his supervision of the MHC project.

All our collaborators on various projects:

Thomas Rask, Massimo Andreatta, Morten Nielsen, Kasper Jensen, Irene Kouskoumvekaki and Gianni Panagiotou from CBS.

Thomas Lavstsen, Louise Jørgensen, Anja Tatiana Ramstedt Jensen and Thor Grundtvig Theander from the Centre for Medical Parasitology, Department of International Health, Immunology and Microbiology, University of Copenhagen.

Kim Magnussen and Lars H. Hansen from the Department of Microbiology, Institute of Biology, University of Copenhagen.

Manuel Llinas and Erandi K De Silva at the Department of Molecular Biology and the Lewis-Sigler Institute for Integrative Genomics, Princeton University.

Anders, Morten, Massimo, Irene and Kasper for proofreading this thesis.

The CBS administration, especially Dorthe and Lone for making my stay at CBS a smooth and pleasant experience.

The system administration, Kristoffer, John, Peter, Hans-Henrik and Olga for keeping the system running.

All the wonderful people at CBS that makes this place a very special place. I will not start to mention names because I do not know where to stop again.

My daughter, Alberte, for being who she is and giving me a break from the thesis from time to time.

Papers included in the thesis

- **Paper I:** Rask TS, **Hansen DA**, Theander TG, Pedersen AG, Lavstsen T. *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes – divide and conquer. PLoS computational biology, 6(9):e1000933 (2010).

Papers not included in the thesis

- D'Hertog W, Maris M, Ferreira GB, Verdrengh E, Lage K, **Hansen DA**, Cardozo AK, Workman CT, Moreau Y, Eizirik DL, Waelkens E, Overbergh L, Mathieu C. Novel insights into the global proteome responses of insulin-producing INS-1E cells to different degrees of endoplasmic reticulum stress. Journal of proteome research, 9(10):5142-5152 (2010).

- Ferreira GB, van Etten E, Lage K, **Hansen DA**, Moreau Y, Workman CT, Waer M, Verstuyf A, Waelkens E, Overbergh L, Mathieu C. Proteome analysis demonstrates profound alterations in human dendritic cell nature by TX527, an analogue of vitamin D. Proteomics, 9(14):3752-3764 (2009).

- Ferreira GB, Overbergh L, van Etten E, Lage K, D'Hertog W, **Hansen DA**, Maris M, Moreau Y, Workman CT, Waelkens E, Mathieu C. Protein-induced changes during the maturation process of human dendritic cells: A 2-D DIGE approach. Proteomics. Clinical Applications, 2(9):1349-1360 (2008).

- D'Hertog W, Overbergh L, Lage K, Ferreira GB, Maris M, Gysemans C, Flamez D, Cardozo AK, Van den Bergh G, Schoofs L, Arckens L, Moreau Y, **Hansen DA**, Eizirik DL, Waelkens E, Mathieu C. Proteomics analysis of cytokine-induced dysfunction and death in insulin-producing INS-1E cells: new insights into the pathways involved. Molecular and cellular proteomics, 6(12):2180-2199 (2007).

Abbreviations

AIC	Akaike's information criterion
AP2	Apetala2
ApiAP2	Apicomplexan Apetala2
bp	base pair
CTOB	Complementary To Original Bottom
CTOT	Complementary To Original Top
dCTP	Deoxycytidine triphosphate
HLA	human leukocyte antigen
Illumina GAII	Illumina Genome Analyzer II
MHC	major histocompatibility complex
nt	nucleotide
OB	Original Bottom
OT	Original Top
PCR	Polymerase Chain Reaction
PfEMP1	Plasmodium falciparum Erythrocyte Membrane Protein 1
P. falciparum	*Plasmodium falciparum*

Part I

General Introduction

Chapter 1

Malaria

Malaria is an infectious disease caused by the protozoan parasite of the genus *Plasmodium*. It is transmitted to humans by female anopheline mosquitoes. The disease is widespread in tropical and subtropical regions where high temperatures and significant amounts of rainfall ensure optimal conditions for uninterrupted breeding of the mosquitoes. Classical symptoms of malaria are sudden coldness followed by fever, shaking and sweating. These symptoms typically last for a few hours and re-occur every two to three days.

Four species of *Plasmodium* are traditionally known to cause malaria in humans: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae* and *Plasmodium ovale*. Recently, *Plasmodium knowlesi* has also been observed to infect humans [1–5] although its main host is macaques [6]. *P. falciparum* is responsible for the most severe form of malaria in humans and is the focus of part II of this thesis.

Due to an intensified effort by the international community, the number of malaria cases are on decline. This is in large part due to the increased distribution and use of insecticide-treated nets and indoor residual spraying. Nevertheless, malaria is still a major global cause of disease with an estimated 225 million cases and 781 000 deaths in 2009 [7].

1.1 *Plasmodium falciparum* life cycle and pathogenesis

P. falciparum has a complex life cycle as illustrated in figure 1.1. Humans are infected when bitten by an infected *Anopheles* mosquito. The sporozoites are injected into the subcutaneous tissue and migrate to the liver where the first replica-

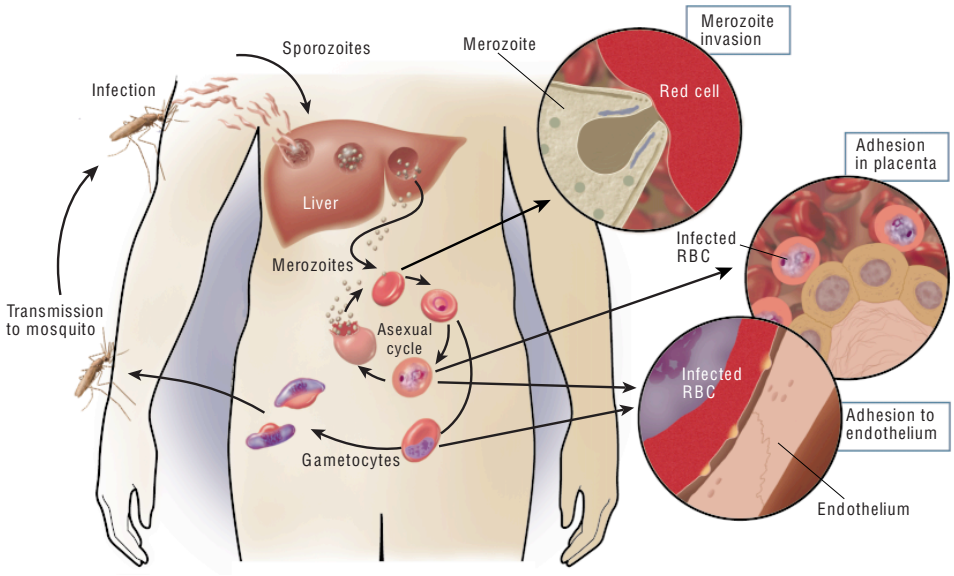


Figure 1.1. The life cycle of *Plasmodium falciparum*. Picture from Miller *et al.*, 2002 [8].

tive stage is established. Inside the liver cells, each sporozoite develops into tens of thousands of merozoites, which are released into the bloodstream, where they invade the red blood cells. Inside the red blood cells, the parasites undergo multiple rounds of asexual reproduction, also known as the intraerythrocytic developmental cycle. It is during this stage that the severe conditions of malaria occur. A small fraction of the asexual parasites develop into gametocytes, which might be ingested by a mosquito taking a blood meal. Inside the mosquito, the sexual reproduction is completed and new sporozoites are formed [8, 9].

Inside the red blood cells, the parasite modifies the red blood cell wall in a way that enables it to adhere to the endothelial cells lining the blood vessels. This is accomplished by the expression of Plasmodium falciparum Erythrocyte Membrane Protein 1 (PfEMP1) on the surface of the red blood cell wall. PfEMP1 has the ability to adhere to receptors found on the endothelial cells in various tissues and causes the sequestration of infected red blood cells that would otherwise have been cleared from circulation in the spleen. Different PfEMP1 variants are responsible for the sequestration in different tissues by specific binding to different receptors, e.g. CD36 in the microvasculature, ICAM-1 in the brain and CSA in the placenta. PfEMP1 binding to uninfected red blood cells leads to the phenomenon called

rosetting, in which several healthy red blood cells are bound by an infected cell. Infected red blood cells also create clumps through platelet binding. It is the sequestration of the infected red blood cells in the various tissues together with the clumping of red blood cells that are responsible for the pathogenicity of malaria [8].

1.2 Genome sequence

The full genome sequence of *P. falciparum* clone 3D7 was published in 2002 [10]. It is being finished and re-annotated at the Wellcome Trust Sanger Institute. The nuclear genome is 23.3 Mb in size and contain 14 chromosomes. The C+G content is unusually low, only approximately 19 %.

The sequencing and annotation of several other *P. falciparum* clones are underway at the Wellcome Trust Sanger Institute and the Broad Institute.

Part II

Malaria

Chapter 2

Classification and diversity of *var* genes

2.1 Introduction

Plasmodium falciparum Erythrocyte Membrane Protein 1 (PfEMP1) is encoded by the *var* gene family. The *var* genes have been classified based on the their upstream seqeunces in the *Plasmodium falciparum* clones 3D7, HB3 and IT4 [10–13]. Here, we extend the previous classification by including four additional *P. falciparum* clones (DD2, PFCLIN, RAJ116 and IGH-CR14) and one *Plasmodium Reichenowi* clone (PREICH).

This project was a collaboration between Thomas Lavstsen and Thor Grundtvig Theander from the Centre for Medical Parasitology, Department of International Health, Immunology and Microbiology, University of Copenhagen; and Thomas Rask, Anders Gorm Pedersen and myself at the Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark.

The project was published as part of a larger project where both the protein coding sequence and the upstream sequence of the *var* genes were analyzed (paper I, appendix A). Thomas Lavstsen and Thomas Rask performed the analysis on the protein coding sequence and I performed the analysis on the upstream sequences. Only the analysis of the upstream sequences is described here.

2.2 Background

The var gene family

PfEMP1, which mediates adhesion of infected red blood cells to the endothelial cells of the blood vessels, is encoded by the *var* gene family [14,15]. This is a highly variable multigene family with ∼60 copies in the 3D7 genome [10]. The majority of the *var* genes are clustered in the sub-telomeric regions while the remaining ones are located centrally in the chromosomes.

The diversity of the *var* genes is important for immune evasion of the parasite. Only a single *var* gene is expressed at a time and transcriptional switching of the expressed variant enables the parasite to evade the immune response [8].

Based on sequence similarity, the sequences upstream of the *var* genes can be divided into upstream (Ups) classes (A, B, C and E). The Ups classes correlate with the chromosomal location of the *var* genes as well as the domain organization of the encoded protein [11, 12]. Subtelomeric UpsA and UpsB genes are oriented tail to tail while central UpsC genes are oriented head to tail in a tandem repeat manner [10]. Based on the classification of the upstream sequences, the *var* genes has been assigned to group A, B and C and two intermediate groups B/A and B/C.

Kraemer and co-workers further subdivided the Ups classes into subclasses (UpsA1–2, UpsB1–4, UpsC1–2 and UpsE) based on clustering of the upstream sequences in the three *P. falciparum* clones 3D7, HB3 and IT4 [13].

2.3 Methods

Data set

A data set containing protein coding sequences of 399 *var* genes in *P. falciparum* clones 3D7, HB3, DD2, IT/FCR3, PFCLIN, RAJ116, IGH-CR14 and *P. Reichenowi* clone PREICH had already been prepared by Thomas Lavstsen and Thomas Rask. Annotated *var* genes were retrieved from PlasmoDB, Broad and Sanger Institute servers. Further *var* genes were identified by BLAST [16] searches with 3D7 *var* sequences against genome contigs retrieved from Broad and Sanger Institute.

For all *var* genes with intact N-terminal segments, we extracted up to 2 000 nucleotides upstream of the coding sequence.

Neighbor joining tree

The sequences were aligned with MAFFT (version 6.240) using the L-INS-i algorithm for multiple sequence alignment [17–19]. The neighbor joining tree was created from the alignment and bootstrapped using Clustalw (version 2.0.9 for

tree construction and version 1.83 for bootstrapping because version 2.0.9 crashed during bootstrap) [20, 21].

Markov clustering

Sequences were clustered using the Markov cluster algorithm (version 08-312) [22, 23]. The Markov cluster algorithm is a graph-theoretical clustering method, which uses an all-against-all pairwise sequence alignment as input, generated with the blastn algorithm implemented in blastall (version 2.2.18) [16]. The inflation parameter of the Markov cluster algorithm was varied in steps of 0.2 from 1.2 to 5.0, and resource scheme 7 (most accurate) was used. A distinct clustering was generated for each value of the inflation parameter, and all the clusters were summarized in a consensus clustering. Briefly, each clustering was converted to a multifurcating tree with a branch representing each cluster. A consensus tree representing the consensus clustering was then constructed, using the majority rule consensus method (include all bipartitions with a frequency larger than 0.5) [24], with the extension that less frequent bipartitions were also included as long as they continued to resolve the tree and did not contradict more frequent groups.

Trees were rendered and edited using Dendroscope (version 2.3) [25].

Nucleotide diversity

In this analysis, only 1 000 nucleotides upstream of the coding sequence were included. Sequences shorter than 1 000 nucleotides were discarded, leaving a total of 293 sequences for the analysis. Global pairwise sequence alignments were obtained using align [26].

2.4 Results

We created a data set of sequences found upstream of *var* genes in seven *Plasmodium falciparum* clones (3D7, HB3, DD2, IT4/FCR3, PFCLIN, RAJ116, IGHCR14) and one *Plasmodium Reichenowi* clone (PREICH). *P. Reichenowi* is the malaria parasite infecting chimpanzees and was thought to be the closest relative of *P. falciparum* although recent reports question this [27, 28]. A data set containing protein coding sequences of 399 *var* genes in these clones had already been prepared by Thomas Lavstsen and Thomas Rask. Using this data set as a starting point, we extracted up to 2 000 nucleotides upstream of the protein coding sequence wherever possible. However, as most of these genomes were not fully assembled and thus only available as contig sequences, it was not always possible to obtain 2 000 nucleotides. In these cases we extracted as much sequence as possible. We were able to extract upstream sequences from a total of 360 *var* genes. Of these, 229 were 2 000 nucleotides long. The length distribution of the upstream sequences is shown in figure 2.1.

Clone	Number of sequences
3D7	62
HB3	41
DD2	48
IT4/FCR3	49
PFCLIN	53
RAJ116	37
IGH-CR14	41
PREICH	20
Other	9

Table 2.1. The distribution of sequences over the different *Plasmodium* clones.

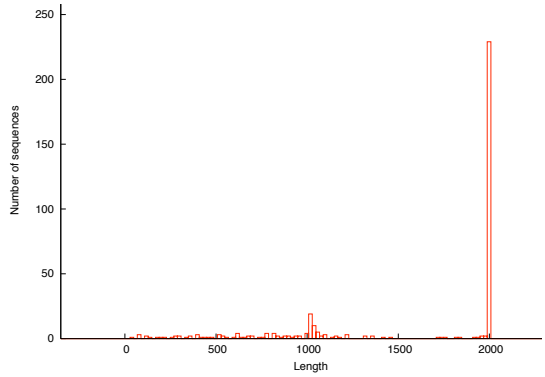


Figure 2.1. The length distribution of the sequences in the data set.

Clustering of *var* upstream sequences

We created a bootstrapped neighbor joining tree of the upstream sequences (figure 2.2 and figure 2.3). Monophyletic groups with a bootstrap support above 0.7 and containing sequences from at least four different *P. falciparum* clones were identified and named in accordance with the assignment by Kraemer *et al.* [13]. These are highlighted with thick red branches in figure 2.2. The assignment by Kraemer *et al.* is shown on figure S2O in appendix B.

Some subclasses were further expanded (without bootstrap support) to form larger monophyletic groups (shown with thick black branches in figure 2.2). UpsA2 and UpsB3 were expanded to include additional sequences annotated to UpsA2 and UpsB3 respectively by Kraemer *et al.* UpsB2 was expanded to include two additional sequences where the coding sequence shared the same domain architec-

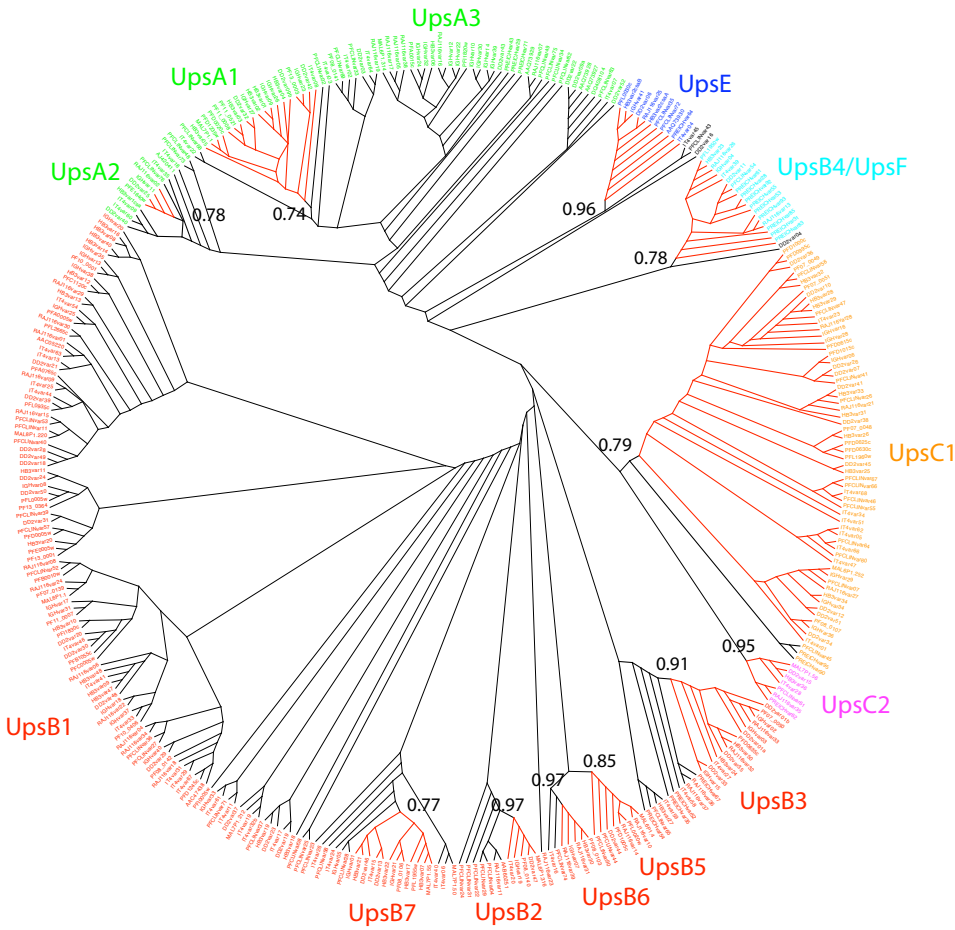


Figure 2.2. Neighbor joining tree of the upstream sequences. The assignment of the sequences to Ups classes is shown. The numbers show the bootstrap support for the monophyletic groups with red branches. See text for details.

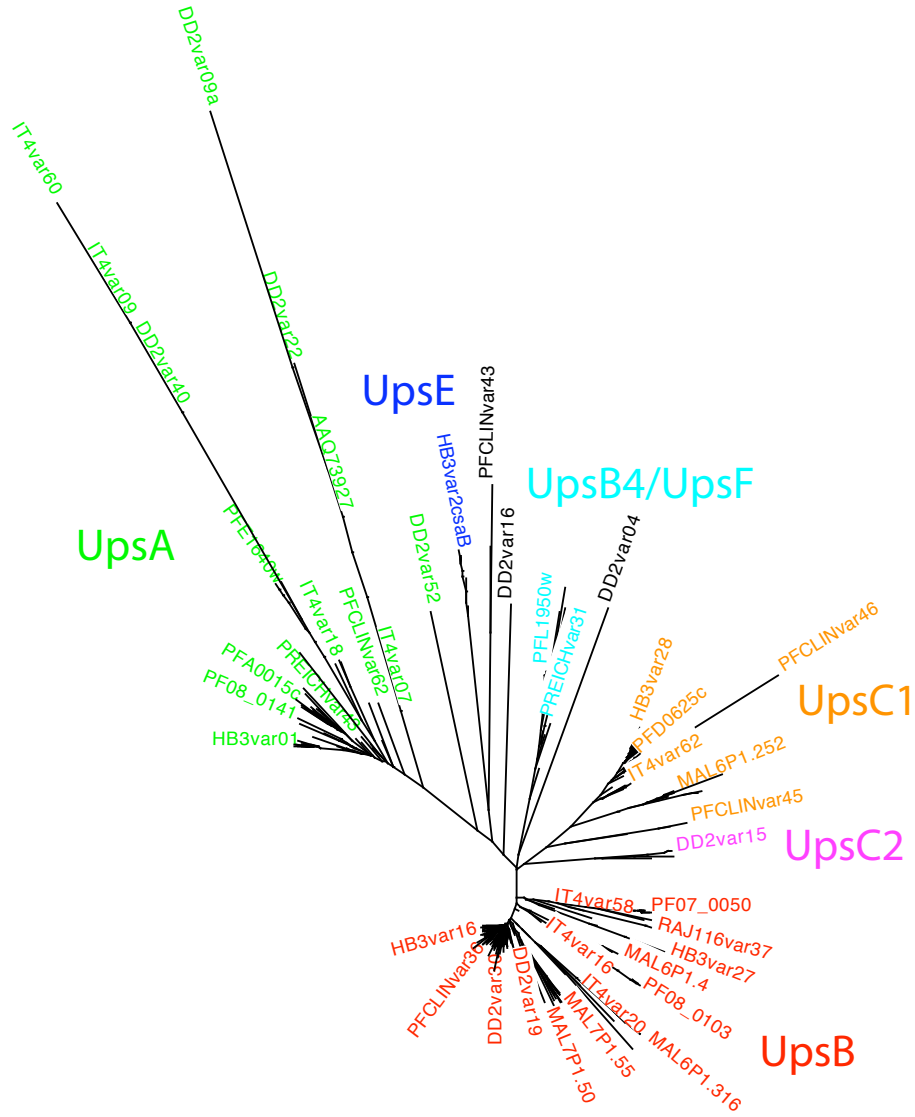


Figure 2.3. Neighbor joining tree of the upstream sequences with branch lengths. This is the same tree as the one in figure 2.2 but the branches are scaled to reflect the branch lengths. Note that only a small fraction of the labels are actually shown. This tree gives a more realistic picture of how the Ups classes cluster in the neighbor joining tree.

ture. UpsC1 was expanded to include three sequences that fell between UpsC1 and UpsC2 but within the larger monophyletic group comprising all UpsC sequences. The additional sequences included by this expansion were denoted with an asterisk in the annotation in table C.1 – table C.5 in appendix C.

A large number of sequences did not fall within any of the identified classes. Some of these clustered with the UpsA sequences and had previously been assigned to UpsA by Kraemer *et al.* [13]. The sequences that clustered with the UpsA sequences and had not been assigned to any other class were assigned to UpsA3. Similarly, sequences that clustered with the UpsB sequences and had not been assigned to any other class were assigned to UpsB1.

All the previously suggested subclasses (UpsA1–2, UpsB1–4, UpsC1–2 and UpsE) were identified, although with some modifications. In addition we identified four new subclasses (UpsA3 and UpsB5–7). It is worth noting that UpsB4 did not cluster with the rest of the UpsB sequences.

We also clustered the sequences using the Markov cluster algorithm [22, 23]. The inflation parameter of the Markov cluster algorithm was varied in steps of 0.2 from 1.2 to 5.0. A distinct clustering was generated for each value of the inflation parameter, and all the clusters were summarized in a consensus clustering, shown as a tree in figure 2.4. The clusters were assigned to Ups classes in accordance with the assignment by Kraemer *et al.* [13]. The assignment by Kraemer *et al.* is shown on figure S2N in appendix B.

All the major classes (UpsA, UpsB, UpsC and UpsE) were also identified with the Markov cluster algorithm but only UpsC was split into supclasses, UpsC1 and UpsC2, and UpsB4 came out as an independent cluster. Based on the observation that UpsB4 did not cluster with the other UpsB sequences in the neighbor joining tree and that it formed its own cluster in the Markov clustering, we suggest that it is renamed to UpsF.

The two methods agreed on the assignment of most sequences. Hence, we derived the consensus annotation shown in table C.1 – table C.5 in appendix C. When the two methods disagreed or they were unable to assign a sequence to any group, the sequence was annotated as ND in the tables. This was only the case for 11 of the 360 sequences.

Distribution of Ups classes

Based on the consensus annotation above, we analyzed the distribution of upstream sequences to Ups classes in the different genomes (figure 2.5). The relative distribution between the Ups classes was similar in all the analyzed genomes. The relative number of UpsC sequences seemed to be a little lower in RAJ116 and IGH-CR14 but this could be explained by a lacking annotation of *var* genes in these two genomes. Only ∼40 *var* genes had been identified in RAJ116 and IGH-CR14.

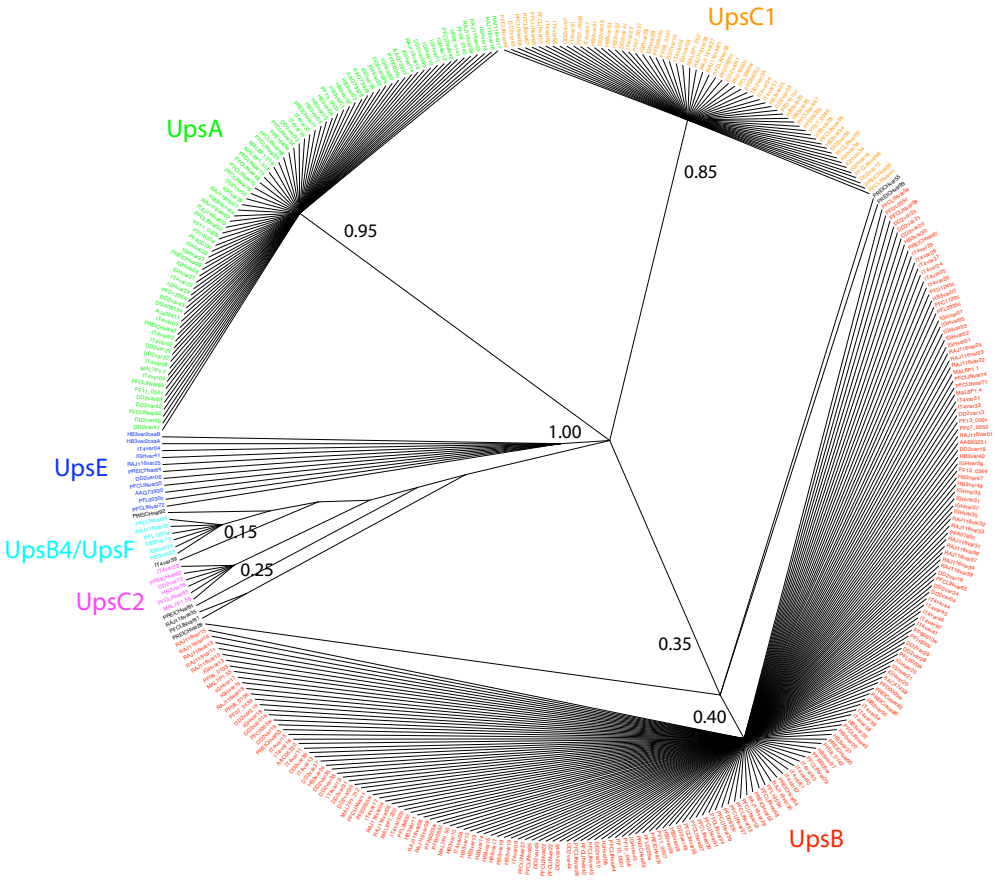


Figure 2.4. Consensus tree of the Markov clustering. The assignment of the sequences to Ups classes is shown. The numbers show the fraction of Markov clusters with this group present. See text for details.

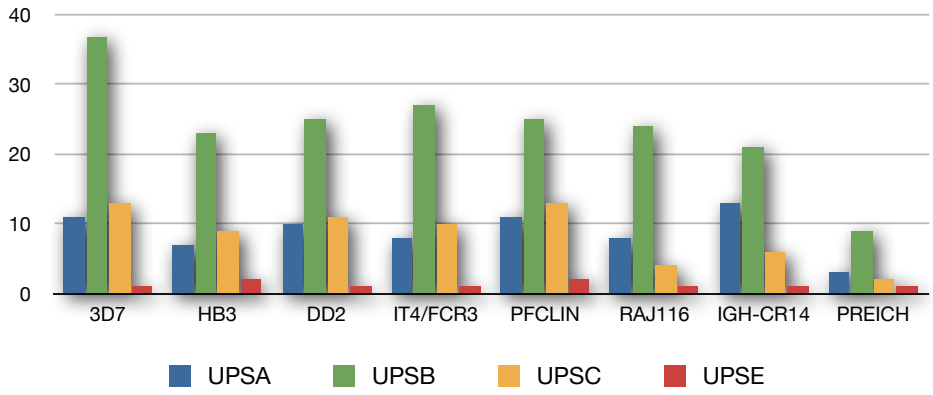


Figure 2.5. The distribution of sequences to UpsA, B, C and E within each of the *Plasmodium* clones.

Comparison	Identical nucleotides
Within each clone	61 %
Between clones	59 %
Within Ups classes	76 %
Between Ups classes	58 %
Within 3D7 subclasses	83 %

Table 2.2. The nucleotide diversity in the upstream sequences of the *var* genes.

Nucleotide diversity

The nucleotide diversity of the upstream sequences of the *var* genes was examined by making pairwise comparisons of the sequences and counting the number of identical nucleotides. The average of all the compared sequences was then calculated. We calculated the diversity within each *Plasmodium* clone, between the clones, within each Ups class, between the Ups classes and within the Ups subclasses in the 3D7 clone (summarized in table 2.2).

The percent identity was ∼60 % both within each clone and between the clones. This means that we saw as much diversity within the upstream sequences within each genome as we saw between genomes. As expected, the diversity was lower within the Ups classes and even lower within the Ups subclasses in 3D7.

2.5 Discussion

We used neighbor joining and Markov clustering to cluster the upstream sequences of the *var* genes and assign them to Ups classes. The two methods yielded congruent results although additional subclusters could be identified in the neighbor joining tree. All the previously suggested subclasses (UpsA1–2, UpsB1–4, UpsC1–2 and UpsE) were identified, although with some modifications. In addition we identified four new subclasses (UpsA3 and UpsB5–7).

UpsB4 did not cluster with the rest of the UpsB sequences in any of the two methods. Hence, we suggest renaming it to UpsF.

Based on the consensus of the neighbor joining tree and the Markov clustering, we derived a consensus annotation of the upstream sequences. The fact that Markov clustering agreed with the neighbor joining tree on most sequences strengthens the support for the major classes and some of the subclasses (UpsC1, UpsC2 and UpsB4/UpsF).

As previously observed, *var2csa* genes had UpsE upstream sequences and *var1csa* had UpsA2 upstream sequences [11, 12]. Type 3 var genes had UpsA3 upstream sequences. The group of *var* genes previously classified as group B/A *var* genes were assigned to domain casette 8 in paper I (appendix A). These are marked with black squares on the figure S2N and S2O in appendix B. Many of these had upstream sequences of the UpsB2 class. The group of *var* genes previously classified as group B/C *var* genes did not have a common type of upstream sequence.

We saw a similar distribution of sequences between the Ups classes in the different genomes. The small differences that could be observed was probably due to a lacking identification of *var* genes in some of the clones. This was most pronounced in RAJ116 and IGH-CR14 where only ∼40 *var* genes had been identified. These *var* genes were identified by BLAST [16] searches with 3D7 *var* sequences against genome contigs. Therefore, it is very likely that several *var* genes had been missed.

The nucleotide diversity in the upstream sequences revealed that on average the same amount of diversity was present among the upstream seqeunces of the *var* genes within each clone as between the clones. As expected the nucleotide diversity was lower within the Ups classes and Ups subclasses.

Chapter 3

DNA methylation in malaria

3.1 Introduction

DNA methylation is an epigenetic mark with importance for many aspects of gene regulation in eukaryotes. Yet, very little is known about DNA methylation in the malaria parasite *Plasmodium falciparum*. A couple of studies back in the 1980'ies were unable to find any evidence of DNA methylation in this parasite [29, 30] and although one instance of DNA methylation was later observed [31], the phenomenon has not been studied using the tools that are available today. Hence, we decided to perform a genome-wide analysis of DNA methylation in *P. falciparum*.

This project is a collaboration between Louise Jørgensen and Anja Tatiana Ramstedt Jensen from the Centre for Medical Parasitology, Department of International Health, Immunology and Microbiology, University of Copenhagen; Kim Magnussen and Lars H. Hansen from the Department of Microbiology, Institute of Biology, University of Copenhagen; and Thomas Rask, Anders Gorm Pedersen and myself at the Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark.

Louise and Anja from the Centre for Medical Parasitology are responsible for delivering the *P. falciparum* DNA, and Kim and Lars from the Department of Microbiology are responsible for bisulfite treatment and sequencing of the DNA. Our group at the Center for Biological Sequence Analysis are overall responsible for the project and the data analysis.

Unfortunately, the project has been delayed for several months and we are still waiting for the DNA to be sequenced. Therefore, I am not able to present any results on DNA methylation in *P. falciparum*. Instead, I will present some

simulations we have done in preparation for the study and describe how to deal with methylated DNA sequences in high-throughput sequencing.

3.2 Background

DNA methylation

DNA methylation is a feature found in many organisms. It involves the addition of a methyl group to the pyrimidine ring of cytosine or the purine ring of adenine. In prokaryotes and some lower eukaryotes both adenine and cytosine residues can be methylated whereas in higher eukaryotes, only cytosine methylation has been observed [31, 32]. Although cytosine methylation is conserved in many eukaryotic groups including plants, animals and fungi, it has been lost from some model organisms including the budding yeast *Saccharomyces cerevisiae* and the nematode worm *Caenorhabditis elegans* [33].

DNA methylation is an inheritable epigenetic mark that adds an additional layer of information to that encoded by the four nucleotides of the DNA code. Yet it is a dynamic feature that can change during the differentiation of cells or in response to diet and other environmental factors [34]. It is known to be important for the correct onset of differentiation, genomic imprinting, X-chromosome inactivation, silencing of transposons and regulation of gene expression. Establishment and maintenance of DNA methylation is performed by the DNA methyltrasferases [33, 34].

DNA methylation has been known and studied for many years but it is only after the invention of next-generation high-throughput DNA sequencing technology, it has become possible to study the genome-wide methylation patterns at single-base resolution.

Types of methylation

DNA methylation is typically divided into three types, depending on the context of the methylated cytosine. These are CG, CHG and CHH where H is either A, C or T. CHG and CHH methylation is also referred to as *non-CG* methylation. Methylation is most frequent in the CG context although it varies between species and non-CG methylation is relatively more frequent in plants than in animals [34]. Regions enriched in CG (or CpG) dinucleotides are often associated with gene regulatory regions and are referred to as CpG islands. Methylation of these regions is usually associated with repression of translation of the downstream gene [35].

In a study of the human methylome, Lister et al. found a relative abundance of non-CG methylation in the H1 human embryonic stem cell line. Non-CG methylation accounted for almost 25 % of the total amount of methylated cytosines. While almost all non-CG methylation disappeared upon induced differentiation of the embryonic stem cells it could be re-established at the same loci in induced

pluripotent stem cells. Non-CG methylation was not evenly distributed across the genome but was enriched within the gene bodies of highly expressed genes [36]. Similarly CG-methylation of gene bodies has also been shown to be positively correlated with gene expression in human cell lines [37].

Malaria parasites and DNA methylation

Very little is known about DNA methylation of the genome of the malaria parasite. In 1982, Pollack reported a lack of 6-methyladenine and 5-methylcytosine in high performance liquid chromatography analysis of hydrolyzed *Plasmodium falciparum* DNA [29]. They confirmed the lack of 5-methylcytosine by analyzing the DNA fragments resulting from digestion of genomic *P. falciparum* DNA with the restriction enzymes *Hpa*II and *Msp*I. Both enzymes recognize the DNA sequence CCGG but while *Msp*I is indifferent to the methylation state, *Hpa*II will not cut when the internal cytosine residue is methylated. By comparing the pattern of fragments produced by the two enzymes, it is possible to estimate the degree of methylation at the internal cytosine in the seqeunce CCGG [29].

In 1987, the lack of 5-methylcytosine in the *P. falciparum* genome was confirmed by restriction enzyme analysis of four genes with *Hpa*II and *Msp*I [30].

However, the low C+G content of the plasmodial genome and a general underrepresentation of CpG dinucleotides coupled with low sensitivity of high performance liquid chromatography could explain why previous attempts had failed to identify any cytosine methylation. Underrepresentation of CpG dinucleotides is a common feature of vertebrate genomes and has been shown to correlate with the degree of cytosine methylation. It is speculated to be caused by mutation of methylcytosine to thymine by deamination [38, 39].

In 1991, Pollack examined the methylation status of the plasmodial DHFR-TS gene using a range of different restriction enzymes. This time he was able to identify partial cytosine methytlation of a single CpG dinucleotide in the GATCGA context [31].

Methods to study DNA methylation

Methods to study DNA methylation on a genome-wide scale can be divided into three types [34]. The first type is based on enrichment of a sample with methylated genomic DNA fragments. This can be achieved through antibody capture of methylcytosines as in methylated DNA immunoprecipitation (MeDIP) [40]. The second type is based on digesting a DNA sample with methylation-sensitive restriction enzymes. The *Hpa*II tiny fragment enrichment by ligation-mediated PCR (HELP) assay is an example of this approach where the genome is digested by a methylation-sensitive restriction enzyme and its methylation-insensitive isoschizomer [41]. Both of these approaches are relatively cheap but suffer from their limited resolution.

Only the third type of methods, which is based on high-throughput sequencing of bisulfite treated DNA, provides genome-wide single-base resolution of DNA methylation. Two similar approaches, MethylC-seq [36,42] and BS-seq [43], exists. They both rely on the conversion of unmethylated cytosines to uracil by sodium bisulfite. After PCR amplification, the uracils have been replaced by thymines and only cytosines that were originally methylated in the genomic DNA will remain as cytosines. These can now be detected by deep sequencing.

Available methylomes

In 2008, MethylC-seq and BS-seq were applied to the genome of the flowering plant *Arabidopsis thaliana* to generate the first genome-wide single-base resolution methylomes [42, 43]. A year later, the human methylome had been determined by MethylC-seq [36] and in 2010, several other eukaryotic methylomes were publised [33, 44, 45], so that the methylome is now available at high resolution for 23 eukaryotic organisms (10 animals, 8 plants and 5 fungi).

3.3 Introduction to MethylC-seq

Bisulfite treatment of DNA coupled to Sanger sequencing has been used for several years to detect the presence of cytosine methylation in DNA. However, with the introduction of next-generation high-throughput DNA sequencing technologies, this technique can now be applied at the genome level. The MethylC-seq method was first published by Lister and co-workers in 2008 [42] and with minor modifications in 2009 [36]. The main points of the experimental procedure of the MethylC-seq method and processing of next-generation sequencing data are described below.

Experimental procedure

The genomic DNA of interest is purified and a small amount of unmethylated Lambda DNA is added. The lambda DNA will be used to estimate the rate of cytosine non-conversion and sequencing error. The DNA is then fragmentet to yield pieces of suitable sizes for sequencing, and the ends are repaired using a mixture of nucleotides without dCTP to avoid the inclusion of cytosines with unknown methylation status. The 3' ends are adenylated to prevent them from ligating to one another during the adapter ligation reaction. Methylated adapters are then ligated to the ends of the DNA fragments. The adapters are methylated to avoid any conversion of their sequence during the bisulfite treatment. The ligation products are purified and size selected before bisulfite treatment. The bisulfite treatment is repeated to get maximal conversion of unmethylated cytosines to uracils. Next, the bisulfite treated DNA fragments are amplified by PCR to enrich them for fragments containing adapter sequences at both ends and to replace the uracils with

thymines. The PCR primers also contain the necessary adapters for sequencing. The number of PCR cycles are minimized to avoid skewing the representation of the library. The final product is then purified, validated and quantified. To generate the appropriate cluster density on the flow cell for sequencing, the DNA concentration has to be precisely adjusted. The DNA is now ready for sequencing on the Illumina Genome Analyzer II (Illumina GAII) platform.

Pre-processing of data

When the raw sequence reads are retrieved from the Illumina GAII platform some pre-processing needs to be done. Usually, the data will be in the form of FASTQ sequences which contain the read sequence together with quality scores for each nucleotide. However, the reads might still contain adapter sequences from the sequencing and it is generally advisable to filter the reads for low quality sequences.

The FASTX Toolkit provided by the Hannon Lab at Cold Spring Harbor Laboratory [46] provides a collection of tools that can be used to pre-process the reads in a FASTQ file. For example, the fastx_clipper script can be used to remove adapter sequences from the 3' end of the reads and the fastq_quality_trimmer will remove the sequence downstream of the first occurrence of a low quality nucleotide (including that nucleotide).

During treatment with sodium bisulfite, some genomic cytosines might not be converted regardless of their methylation state. This is called *non-conversion* and can happen if a short stretch of DNA is not denatured as sodium bisulfite acts on single-stranded DNA. Some studies try to filter reads that result from non-conversion by eliminating reads that contain three consecutive CHH's [43].

However, with the recent findings of the relative abundance of non-CG methylation especially in plants [34], such filtering might not be biologically justifiable and could introduce unwanted biases into the data.

Quality scores

The FASTQ format is a textbased format that contain both the sequence and its corresponding quality scores. Over time, different quality scores have been used, but the most common today is the Phred quality score [47, 48]. The Phred quality score of a nucleotide is:

$$Q_{Phred} = -10 \log_{10} p_e, \tag{3.1}$$

where p_e is the estimated probability of that nucleotide being wrong. The relationship between Phred quality scores and the estimated error probability is shown in table 3.1. The quality scores for raw reads are typically in the range 0–40 and are encoded in ASCII text format. The encoding of quality scores varies with different platforms and versions but in the Illumina 1.3+ pipeline, each quality score is represented by the ASCII character having the value of the Phred score + 64. So a

Q_{Phred}	p_e
10	10^{-1}
20	10^{-2}
30	10^{-3}
40	10^{-4}

Table 3.1. The relationship between Phred quality scores (Q_{Phred}) and error probabilities (p_e).

quality score of 0 is encoded by the ASCII character having the value 64 (0 + 64), which is @. A quality score of 40 is encoded by the ASCII character having the value 104 (40 + 64), which is h. The complete range of characters, representing scores from 0–40 is:

```
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_‘abcdefgh
```

Alignment of reads

When the reads have been preprocessed, they need to be aligned to the reference genome of the organism they came from. Due to the large number of reads typically produced by each run of the sequencing machine, traditional alignment algorithms are too slow for this task. Therefore, special software has been developed that are optimized for fast alignment of millions of short reads to large genome sequences. A popular category of programs is based on a Burrows-Wheeler index of the genome sequence [49].

Bowtie [50] is one such program that has become popular due to its speed and accuracy. In its default mode, Bowtie defines a seed region containing the first 28 nucleotides of each read. A valid alignment between a read and the genome is only allowed to have two mismatches in the seed region and the sum of the quality scores of all mismatches in the entire alignment are not allowed to exceed 70. When this sum is calculated, the quality scores are rounded to the nearest 10 and not allowed to exceed 30.

Due to the bisulfite treatment of the DNA, the alignment of reads in MethylC-seq is more complex than in ordinary sequencing. Unmethylated cytosines are converted to thymines on both DNA strands rendering the two strands non-complementary. Furthermore, the conversion of many cytosines to thymines will lead to many mismatches between the reads and the genome sequences unless some transformations are performed *in-silico* before the alignment. This is illustrated in figure 3.1.

In figure 3.1A, a short stretch of the original methylated DNA is shown. The top strand is shown in green and labeled OT (Original Top) and the complementary

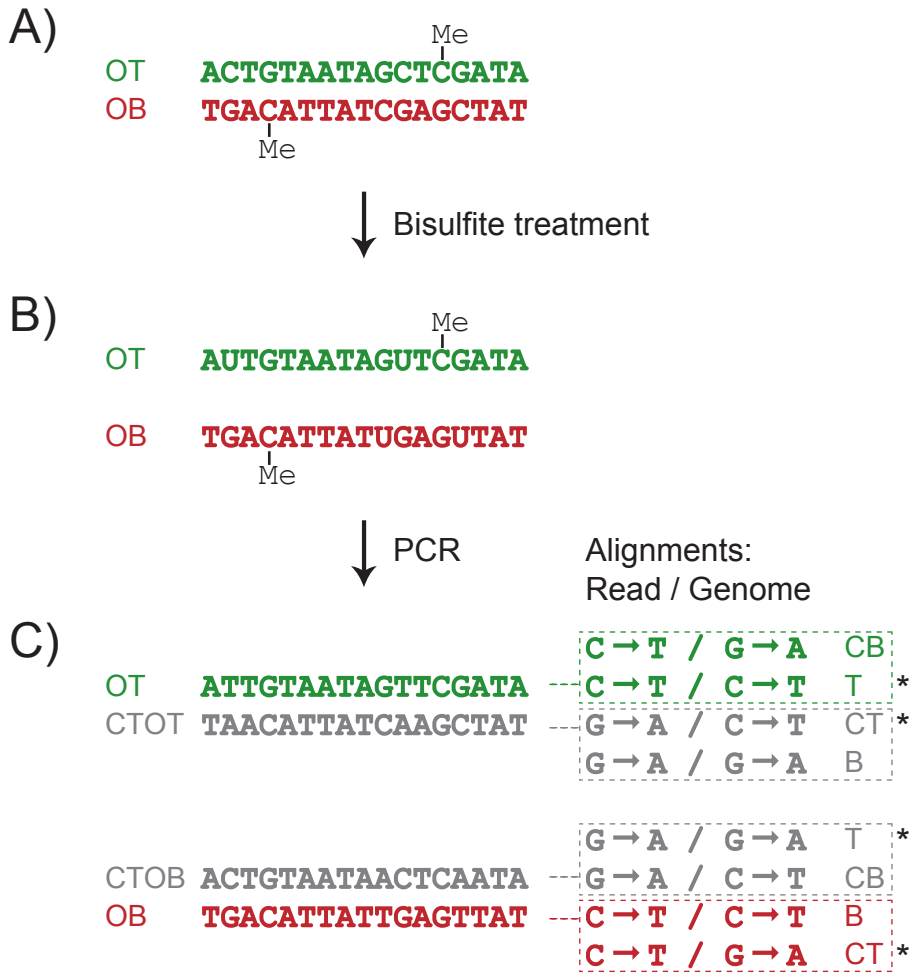


Figure 3.1. The modifications happening to a hypothetical stretch of DNA sequence during the MethylC-seq protocol. (**A**) The original doublestranded piece of DNA with cytosine methylations on both strands. (**B**) After bisulfite treatment, the DNA is singlestranded and unmethylated cytosines have been converted to uracils. (**C**) After PCR amplification, the uracils have been replaced by thymines and new complementary strands have been formed. The boxes on the right hand side of the figure show the possible ways that each sequence can be aligned to the reference genome. The read and the reference genome have to be converted as shown (see text for details). OT: original top; OB: original bottom; CTOT: complementary to original top; CTOB: complementary to original bottom; T: top strand in genome; B: bottom strand in genome; CT: complementary to top strand in genome; CB: complementary to bottom strand in genome; *: alignments used by Bismark (see text for details).

strand is shown in red and labeled OB (Original Bottom). In figure 3.1B, the same sequence is shown after bisulfite treatment. Now, the strands are separated and non-methylated cytosines have been converted to uracils. As a result, the two strands are no longer complementary. After PCR amplification (figure 3.1C), the uracils have been replaced by thymines and two new complementary strands have been generated (shown in dark grey). These are labeled CTOT (Complementary To Original Top) and CTOB (Complementary To Original Bottom) respectively. OT and CTOT provides information about methylation on the original top strand while OB and CTOB carries information about methylation on the original bottom strand.

Each of the four strands (OT, OB, CTOT, CTOB) can be aligned either to the top or to the bottom strand of the reference genome. Some transformations of both the reads and the genome sequence are necessary in order to avoid introducing mismatches in the alignments as a result of the experimental procedure. The possible ways to align each sequence is shown on the right hand side of figure 3.1C. For example, one way to align the OT read (shown in green) is to convert C to T in the read and C to T in the genome. It is then possible to align the read to the top strand of the reference genome (indicated by the T). However, it is also possible to convert C to T in the read and G to A in the genome. The read is then complementary to the bottom strand of the reference genome (indicated by CB).

A convenient choice is to align all four types of reads (OT, OB, CTOT and CTOB) to the top strand of the reference genome. Then it is not necessary to compute the reverse complement of the reference genome and the alignment positions do not need to be mapped from one strand to the other. The four transformations that correspond to this choice are marked with an asterisk in figure 3.1. As it can be seen from the figure, this requires that both C to T and G to A converted reads are made as well as both C to T and G to A converted versions of the top strand of the genome sequence. It is then necessary to align all four possible combinations of these as illustrated in figure 3.2.

For a particular read, it is not known which of the four strands (OT, OB, CTOT or CTOB) it represents. Hence, it is necessary to do all four alignments for each read in the data set. A particular alignment of a read is only accepted if it is the unique best alignment. To fulfill that, the alignment has to be the only valid alignment with a particular number of mismatches. There can be other valid alignments with more mismatches. But the read is discarded if there is more than one valid alignment with the minimum number of mismatches.

Bismark [51] is a flexible tool for the analysis of bisulfite treated DNA sequences. It implements the strategy outlined above for aligning reads to a reference genome. It uses Bowtie [50] to perform the actual alignments. After aligning the reads, Bismark determines the methylation status of all cytosine residues in the reads.

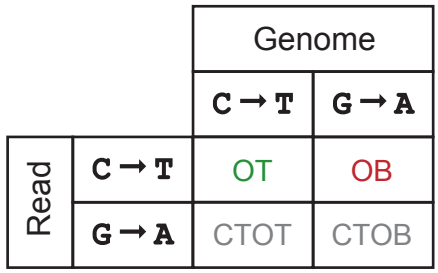


Figure 3.2. Alignment matrix, showing how the different combinations between read sequence and genome sequence corresponds to aligning each of the four possible strands that results after bisulfite treatment and PCR amplification (see figure 3.1) to the genome. The *in silico* conversions made to the read sequence are shown on the left and the *in silico* conversions made to the top strand of the genome sequence are shown on the top. OT: original top; OB: original bottom; CTOT: complementary to original top; CTOB: complementary to original bottom.

Clonal amplification

When analyzing a large number of reads, some genomic positions will be covered by several independent reads, that originate from different parent cells. This can be used to estimate the fraction of cells that had its DNA mehtylated at a given position. It can happen, however, that several reads are obtained from a single cell because of clonal amplification during the PCR reaction. This will put too much weight on the clonal fragments in the estimation of the methylation frequency.

Clonal reads will share the same start position and Charles Berry found that in a sample of 10,000 randomly sampled positions in a MethylC-seq data set, more reads shared the start position than would be expected, if it only happened by chance. He analyzed the effect of the clonal reads on the variance of the estimated fraction of methylated cells, and found that the optimal solution was to keep only one random read when two or more reads shared the same start position [42, see supplementary material].

Post-processing of data

The first step in post-processing of the data is to remove the clonal reads and only keep one random read where two or more reads share the same start position. This can be done using the script deduplicate_bismark_alignment_output.pl which is part of the Bismark package.

Then aligned reads containing too many mismatches should be removed or trimmed. Since the alignment procedure only limits the number of mismatches in the seed region, a read can contain several mismatches in the entire alignment as long as the sum of quality scores is below the limit. One strategy is to trim reads containing more than three mismatches relative to the reference just before the fourth mismatch [36]. Another strategy is to truncate reads at the point where the next four nucleotides contain two or more mismatches [42].

Now, the aligned reads have to be compared to the reference sequence to determine the original nucleotide at each position in the reads. In addition, the methylation status of the cytosines has to be determined. This can be done using the methylation_extractor script, which is part of the Bismark package.

Identification of methylated cytosines

In order to determine the methylation status of each cytosine in the reference genome, it is necessary to sum up the information provided by the reads overlapping with this reference position. The fraction of reads with a methylated cytosine at a given position is an estimate of the fraction of cells that had this cytosine methylated in their genome. But due to sequencing error and non-conversion of cytosines during bisulfite treatment, some reference cytosines might erroneously appear to be methylated. Hence, it is necessary to test if the observed number of methylated cytosines at each site is significant or not. This can be done using the binomial distribution as described below. Since this test is performed at each site in the genome, it is also necessary to control for multiple testing. One way of dealing with multiple testing is to control the false discovery rate of methylated cytosines. This can be done by the classical Benjamini-Hochberg procedure for control of false discovery rate [52] or by an alternative approach published by Lister and co-workers [36, personal communication].

In the method by Lister and co-workers, a binomial distribution is used to calculate the probability of erroneously detecting k methylated cytosines at an unmethylated position in the genome if the read coverage at that position is n and the error rate is p:

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.2}$$

The error rate, p, is determined from the frequency of sequenced cytosines in reference cytosine positions in the unmethylated Lambda DNA that is added to the genomic *P. falciparum* DNA before fragmentation. Since the Lambda DNA is unmethylated, no cytosines should be left after bisulfite treatment. The frequency of sequenced cytosines in the Lambda DNA is therefore an estimate of the error due to sequencing error and non-conversion of cytosines.

In order to keep the false discovery rate below, say 1 %, the probability $f(k; n, p)$ has to be less than the value M where:

$$M \times n_{uC} < 0.01 \times n_{mC}. \tag{3.3}$$

n_{uC} is the number of unmethylated cytosines and n_{mC} is the number of methylated cytosines in the entire genome. In other words, the probability of erroneously classifying an unmethylated cytosine as methylated multiplied by the number of unmethylated cytosines in the genome has to be less than one percent of the number of methylated cytosines in order to keep the false discovery rate below one percent.

Since the numbers of methylated and unmethylated cytosines are not known in advance, M has to be determined iteratively using the formula:

$$M = 0.01 \frac{n_{mC}}{n_{uC}} \tag{3.4}$$

In the first iteration, all reference cytosines where at least one methylated cytosine has been sequenced are counted as methylated and the remaining reference cytosines are counted as unmethylated. These values of n_{uC} and n_{mC} are used to calculate M. In the second iteration, for every reference cytosine, $f(k; n, p)$ is now compared to the value of M to determine if this cytosine is methylated or not. If $f(k; n, p) < M$, it is methylated. This yields new values of n_{uC} and n_{mC}, which in turn are used to update M. This procedure is iterated until M converges to a stable value and the final methylation calls can be made.

Consider the following hypothetical but realistic example where the error rate, p, has been determined to be 0.005 and the value M has converged to the value 0.00043 after a number of iterations. The probability, $f(k; n, p)$, for different combinations of read deapth, n, and number methylated cytosines, k, at a given position in the genome is shown in table 3.2.

From table 3.2 we see that if a reference cytosine is only covered by a single read, i.e. the read depth, n, is equal to 1, we will never be able to call it as methylated because the probability that this has happened by error is too large (0.005). It has to be less than M, which in this example is 0.00043, in order to be called as methylated. If the read depth is 2 and both reads are methylated at that position, it will be called as methylated. If the read depth is 20, it requires 3 methylated reads to call it as methylated and if the read depth is 30, it requires 4 methylated reads.

3.4 Methods

Genome sequence

The complete sequence of the *Plasmodium falciparum* genome was downloaded from PlasmoDB version 7.1 (genome version date 2010-06-01) [53]. The complete

n	k	$f(k;n,p)$
1	1	0.0050
2	1	0.010
2	2	0.000025
3	1	0.015
3	2	0.000075
3	3	0.00000013
4	1	0.020
4	2	0.00015
4	3	0.00000050
20	1	0.091
20	2	0.0043
20	3	0.00013
30	1	0.13
30	2	0.0095
30	3	0.00044
30	4	0.000015

Table 3.2. The table shows the binomial probability ($f(k;n,p)$) of erroneously observing k methylated cytosines out of n trials (read depth) if the error rate p is 0.005.

P. falciparum genome was publised in 2002 [10] and is being finished and systematically re-annotated by the Wellcome Trust Sanger Institute.

Simulated reads

To simulate the reads, we created an *in silico* bisulfite treated *P. falciparum* genome. First, the reverse complement of the genome was created, because the bisulfite treatment will render the two DNA strands different. Methylation was introduced at a random rate to the cytosines on both strands but at different rates in the different sequence contexts. In the CG context the simulated methylation rate was 0.25, in the CHG context it was 0.10 and in the CHH context it was 0.02. These rates are similar to those found in *Arabidopsis thaliana* [43]. All non-methylated cytosines were then converted to thymines to mimic the action of the bisulfite treatment.

Short reads of either 35 nt or 75 nt were then sampled randomly from the genome. In order to sample evenly from the entire genome, all chromosomes were concatenated but with a marker separating the genomes. Reads were then sampled at a uniform rate across the entire genome while discarding any reads containing the chromosome separation marker.

Errors were introduced in the sampled reads to mimic the sequencing error. The sequencing error from the Illumina GAII platform is typically reported to be in the order of 1–2 % but it varies a lot between experiments. The quality is typically best in the 5' end of the reads and decreases towards the 3' end of the reads. During pre-processing, the low quality part of the read will be removed so the part that is used in the alignment will have a lower error rate than the one reported here. It is very difficult to predict how the quality of the reads will be in the final experiment and how it will be distributed in the read, so in order to make things a little easier in the simulation, we decided to discard 30 % of the reads and use an error rate of 0.002 in the remaining reads.

Discarding 30 % of the reads also accounts for the reads that will be filtered out due to clonal amplification. A fraction of reads will still be filtered out because they share the same start position as part of the post-processing of reads. They share a common start position for random reasons but when the real experiment is conducted, it is not possible to distinguish these from the truly clonal reads.

We are planning to run a single lane on the Illumina GAII sequencing machine as a first trial. This generates around 40 000 000 reads. Discarding 30 % of these corresponds to obtaining 28 000 000 useful reads.

Aligning and post-processing reads

Reads were aligned to the reference genome of *P. falciparum* using both Bowtie [50] and Bismark [51]. Bismark is particularly suited for the analysis of bisulfite treated DNA. It uses Bowtie to perform the alignements and then extracts information about methylation status of the individual reads.

When Bowtie was used to align the reads, it was run in the default alignment mode allowing 2 mismatches in the seed region, which was the first 28 nucleotides of the read. The total sum of quality values of all mismatched positions were not allowed to exceed 70. Quality values were rounded to the nearest 10 and saturated at 30. Bowtie was run with the -m 1 and –all options which means that all valid alignments were returned for a given read but only if there was only one valid alignment to return. This way, only unique alignments were reported. Quality scores of the Illumina 1.3+ pipeline were selected with the –solexa1.3-quals option, and –chunkmbs 256 were set to assign enough memory to allow Bowtie to find all valid alignments for each read.

Bismark was run with default options except for the options passed on to Bowtie. The alignment options were either the default options as explained above, the seed region was extended to 40 nucleotides with the -l 40 option, or the seed region was extended to 40 nucleotides and only one mismatch was allowed with -l 40 -n 1. The –solexa1.3-quals and –chunkmbs 256 options were also set.

Clonal reads were removed after running Bismark using the script deduplicate_bismark_alignment_output.pl and methylation status was extracted using the script methylation_extractor. Both scripts were part of the Bismark package.

Custom scripts were used to count the clonal reads after aligning with Bowtie alone.

MethylC-seq library generation

Based on the work by Lister and co-workers [36, 42] and Illuminas standard protocol for library generation for Illumina GAII (Paired-End sample preparation kit), we deduced our own protocol for MethylC-seq library generation. Our protocol is outlined below, paying special attention to the steps that deviates from the standard Illumina protocol. Illuminas protocol for paired-end analysis was used as a starting point because Illumina only provided methylated adaptors for paired-end sequencing. The generated libraries can be used for either single-end sequencing (use Standard Cluster Generation Kit) or paired-end sequencing (use Paired-End Cluster Generation Kit).

Protocol for MethylC-seq library generation

1. **Prepare DNA**
   Purify 5 μg genomic DNA and add 25 ng unmethylated *c*1857 *Sam*7 Lambda DNA (Promega).

2. **Fragment the DNA**
   Fragment the DNA using sonication (with Bioruptor) to yield fragments around 100–150 bp. These fragments will be appropriate for single-end sequencing of 75 bp. If the library is going to be used for paired-end sequencing, the length of the fragments should be adjusted accordingly.

3. **Perform end-repair**
   Convert heterogeneous ends to blunt ends using the T4 DNA polymerase and Klenow enzyme provided by Illumina. Phosphorylate the 5' ends of the DNA by adding T4 PNK (Illumina). Use an alternative dNTP mixture lacking dCTP to avoid insertion of cytosines with unknown methylation status.

4. **Adenylate 3-ends**
   Adenylate the 3' ends of the blunt-ended DNA fragmetns using reagents from Illumina.

5. **Ligate adapters**
   Ligate methylated adapters from Illumina (ME-100-0010) to the DNA fragments containing a 3' A overhang.

6. **Purify and size select ligation products**
   Purify the ligation products by gel electrophoresis and band excision to removed unligated adapters and adapter dimers. The excised band should be

around 200–250 bp for single-end sequencing of 75 bp (each adapter is 60 bp).

7. Bisulfite conversion of DNA
Perform bisulfite treatment of the purified ligation product using the Qiagen EpiTect Bisulfite Kit. The bisulfite treatment should be performed twice to achieve the highest possible cytosine to uracil conversion rate.

8. Amplification of bisulfite converted DNA
Enrich the bisulfite converted DNA by PCR amplification. Use the uracil-insensitive *PfuTurbo* Cx Hotstart DNA polymerase (Agilent Technologies) because it is insensitive to the uracils. As a result the uracils will be replaced by thymines after the PCR amplification. Use PCR primers from Illumina containing the adapters required for cluster generation in the flow cell. PCR reaction mixture:

- 2.5 U uracil-insensitive *PfuTurbo* Cx Hotstart DNA polymerase
- 5 μL 10X *PfuTurbo* reaction buffer
- 25 μM dNTPs
- 1 μL PCR Primer PE 1.0
- 1 μL PCR Primer PE 2.0
- Adjust the final volume to 50 μL

Run the PCR reaction for 10–12 cycles. The number of cycles might have to be adjusted to get enough DNA for sequencing but should be kept as low as possible to avoid clonal amplification.

9. Purify the PCR product
Purify the PCR product by gel electrophoresis and band excision.

10. Validate the DNA library
Check the size distribution of the DNA library, either by gel electrophoresis or by the Agilent Bioanalyzer.

11. Library quantitation
Measure the final DNA concentration.

3.5 Results

Before sequencing the bisulfite treated genome of *Plasmodium falciparum*, we simulated the experiment. The simulated data gave us some indication of what results we could expect and guided the decision of which sequencing strategy to use.

	35 nt reads	75 nt reads
Total reads	28 000 000	28 000 000
Not aligned	67 285	438 040
Not unique	8 107 061	2 212 856
Uniquely aligned reads	19 825 654	25 349 104
Clonal reads	13 063 937	8 898 166
Unique and non-clonal	6 761 717	16 450 938
Coverage per strand (X)	5.1	26.5

Table 3.3. The influence of read length on alignment results.

Simulated reads

We created an *in silico* bisulfite treated genome with a simulated methylation rate of 0.25 in the CG context, 0,10 in the CHG context and 0.02 in the CHH context. These rates are similar to those found in *Arabidopsis thaliana* [43]. Short reads were sampled randomly from the genome and sequencing errors were introduced at a rate of 0.002 in the reads. We generated 28 000 000 reads corresponding to 70 % of the reads obtained from running a single lane on the the Illumina GAII sequencing machine (see methods, section 3.4, for details).

Read length

To study the effect of read length, we sampled reads of 35 and 75 nucleotides. We used Bowtie to align the reads to both strands of a genome where all cytosines were converted to thymines. Reads were only accepted if they had a unique valid alignment. The result is shown in table 3.3. When using reads of 35 nt, only 67 285 could not be aligned compared to 438 040 when using 75 nt reads. However, 8 107 061 reads of 35 nt had more than one valid alignment in the genome and were discarded compared to only 2 212 856 reads of 75 nt. As a result much more 75 nt reads than 35 nt reads were successfully aligned to a unique position in the genome (25 349 104 versus 19 825 654 reads). In addition, the 35 nt reads produced much more apparently clonal reads compared to the 75 nt reads. In total 6 761 717 reads of 35 nt and 16 450 938 reads of 75 nt were useful resulting in a per strand coverage of 5.1 X and 26.5 X respectively.

Conversion of cytosines

In the previous paragraph, reads were aligned to a genome where cytosines were converted to thymines. The purpose was to avoid too many mismatches between the genome and the reads because all non-methylated cytosines had been converted to thymines in the reads. However, mismatches still occurred at the methylated

	35 nt reads	75 nt reads
Total reads	28 000 000	28 000 000
Not aligned	7 723	17 916
Not unique	8 291 331	2 372 224
Uniquely aligned reads	19 700 946	25 609 860
Clonal reads	13 169 203	8 741 333
Unique and non-clonal	6 531 743	16 868 527
Coverage per strand (X)	4.9	27.1

Table 3.4. The effect of converting cytosines to thymines in the reads.

cytosines because they were not converted in the reads. This could create a bias against heavily methylated regions. To avoid this, we also converted cytosines to thymines in the reads before aligning them against both strands of a cytosine to thymine converted reference genome. The result is shown in table 3.4. The conversion increased the number of reads that could be aligned. Only 17 916 reads of 75 nt could not be aligned compared to 438 040 in table 3.3. However, as the conversion also decreased the information content of the reads, the number of reads that did not align to a unique position in the genome increased. The outcome was a slight decrease in coverage for the 35 nt reads but a slight increase in coverage for the 75 nt reads.

Paired-end reads

The reads analyzes so far were single-end reads, where only a single stretch of DNA was sequenced in each read. Another sequencing technique involves paired-end reads, where a read is sequenced from both ends. The two ends that are being sequenced will usually be on opposite strands and are separated by a stretch of DNA called the linker. The advantage of using this technique is a higher coverage as the number of sequenced nucleotides is doubled but also the ability to reach into genome areas that are otherwise hard to sequence. Repeats are an example of such an area, where it can be difficult to align the reads because there are several possible valid alignments for each read. If one end of a paired-end read can be aligned to a unique position and the length of the linker is known, then the position of the other end of the read can also be determined with high reliability. We simulated the alignment of paired end reads where the length of the linker was normally distributed with a mean length of 400 nucleotides and a standard deviation of 5. The paired-end reads were aligned against both strands of a cytosines to thymine converted genome. The result is shown in table 3.5.

As expected, the coverage was much higher for both 35 nt reads and 75 nt reads when using paired-end reads. This was not only because each read contained

	35 nt reads	75 nt reads
Total reads	28 000 000	28 000 000
Not aligned	1 076 166	1 176 492
Not unique	2 081 323	866 317
Uniquely aligned reads	24 842 511	25 957 191
Clonal reads	443 272	428 222
Unique and non-clonal	24 399 239	25 528 969
Coverage per strand (X)	36.7	82.2

Table 3.5. The effect of paired-end reads.

the double amount of information but also because more reads could be uniquely aligned and the number of clonal reads decreased drastically (since both ends had to be identical before it was regarded as clonal).

The cost of sequencing increases with the length of the reads and the number of lanes to be used on the flow cell. One flow cell contains 8 lanes but the simulations were based on using a single lane as this is cheaper than using an entire flow cell. From the simulations it seems that we can get adequate coverage by using single-end sequencing with a read length of 75 nt. This gives around 27 X coverage per strand in the simulations. This will be adequate to allow us to determine the methylation frequency at each methylated cytosine in the genome with reasonable accuracy.

Aligning with Bismark

To simulate alignment and extraction of information about methylated residues we used Bismark to align the reads against the genome of *Plasmodium falciparum*. Only 75 nt reads were aligned. As described in details in section 3.3, Bismark used Bowtie to align the reads against the genome. The result is shown in table 3.6. When using Bismark, we were able to align more reads than we were with Bowtie, because Bismark only required that a read had a unique best alignment whereas with Bowtie, we only accepted reads that were fully unique.

Bismark was run with different options for the seed length and the number of allowed mismatches in the seed. The results of using the different options did not differ much in terms of the final coverage obtained but the running time decreased markedly, from more than 63 hours with the default options to less than 7 hours with the more stringent option of only allowing one mismatch in a 40 nucleotide seed region.

	Default	Faster	Fastest
Bowtie options			
- mismatches in seed	2	2	1
- length of seed	28	40	40
Total reads	28 000 000	28 000 000	28 000 000
Not aligned	13 888	13 198	65 337
Not unique	1 407 406	1 407 431	1 404 723
Uniquely aligned reads	26 578 706	26 579 371	26 529 940
Clonal reads	6 598 283	6 605 058	6 582 658
Unique and non-clonal	19 980 423	19 974 313	19 947 282
Coverage per strand (X)	32.2	32.1	32.1
OT and OB reads	26 564 194	26 564 901	26 515 025
CTOT and CTOB reads	14 512	14 470	14 915
Error	1.09×10^{-3}	1.09×10^{-3}	1.12×10^{-3}
Running time	63 h 47 min	29 h 58 min	6 h 34 min

Table 3.6. Aligning the reads with Bismark.

Estimating the alignment error

The simulated reads were only sampled from the original top (OT) and original bottom (OB) strands (see figure 3.1). When performing the alignment, Bismark classified the successfully aligned reads as belonging to one of the four possible strands: OT, OB, CTOT and CTOB (see figure 3.1). Since the reads were only sampled from the OT and OB strands, the number of reads reported to the CTOT and CTOB strands could be used to estimate the amount of erroneous alignments. Assuming that the same number of errors were made among the OT and OB strands as among the CTOT and CTOB strands, the total number of errors was calculated as twice the number of reads reported to align to the CTOT and CTOB strands. The error rate was consistently around 10^{-3} (table 3.6).

Methylation status

Bismark also provided information about the methylation status of the cytosines in the reads. In all the alignments performed, the methylation frequency in the CG context in the reads were 0.251. In the CHG context it were 0.100 and in the CHH context it were 0.021. These methylation frequencies are equal to those used to simulate methylation in the reads.

3.6 Discussion

We simulated reads generated by high-throughput sequencing of a bisulfite converted *Plasmodium falciparum* genome that we imagined were methylated. Both single-end and paired-end reads of 35 and 75 nt were generated. The reads were aligned to the reference genome of *P. falciparum*.

The best coverage was obtained with paired-end reads of 75 nt when the reads were aligned with Bowtie. But since reasonable coverage (27X per strand) was also obtained with single-end reads of 75 nt and these are cheaper than paired-end reads, we have decided to use single-end reads of 75 nt in the real experiments.

When single-end reads of 75 nt were aligned to the reference genome using Bismark, a slightly higher coverage (32X per strand) were achieved because of small differences in the alignment strategy. The alignment error rate was determined to be in the order of 10^{-3}.

The running time of Bismark differed markedly between the default mode (64 hours) and the more stringent alignment mode where fever mismatches were allowed (7 hours). Despite this difference in running time, the performance in terms of achieved coverage did not differ much. Therefore, the more stringent mode is probably to be preferred in most cases, because it allows fewer mismatches in the alignments in addition to being almost 10 times faster than the default mode. If it is a problem to get enough coverage or if certain areas are hard to align, one can consider trying the less stringent modes.

Reads were only sampled from the OT and OB strands during the simulation. When reads of all four types (OT, OB, CTOT and CTOB) are generated, it will most likely have the effect that fewer reads appear to be clonal due to random sharing of the start position. This is because the reads are now distributed over four types and not just two types. This will result in a higher coverage as less reads are discarded.

It will be exciting to see what the study of DNA methylation in *P. falciparum* can teach us about this parasite. Due to its complex life cycle, extensive gene regulation is likely to take place and epigenetic factors such as DNA methylation could play an important role in this.

Chapter 4

Novel drug targets in malaria parasites

4.1 Introduction

Several drugs for treating malaria are already on the market and have been for many years. Chloroquine was the first drug that was used systematically to combat malaria, but it is virtually useless today due to the spread of drug-resistant parasites. Other common drugs include sulfadoxine-pyrimethamine, mefloquine, amodiaquine and quinine but they are all loosing their efficacy as a result of resistant parasites appearing and spreading around the world [54]. The most effective drugs available today are artemisinin and its derivatives (artesunate and artemether), which are used in combination therapies as the preferred first-line treatment for all falciparum malaria in malaria endemic countries [55]. However, recent reports from western Cambodia of artemisinin resistant *Plasmodium falciparum* parasites highlights the need for new alternatives in the ongoing struggle to combat this deadly parasite [56, 57].

Recently, the ApiAP2 family of putative transcription factors in *Plasmodium falciparum* and other apicomplexan species were discovered. It did not escape our attention that these regulatory proteins could potentially act as drug targets for new antimalarials [58]. When the crystal structure of one of these proteins, PF14_0633, became available, we initiated a project with the aim of testing this protein as a drug target.

The project was done in collaboration with Kasper Jensen, Irene Kouskoumvekaki and Gianni Panagiotou from the Computational Chemical Biology group

at the Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark; Manuel Llinas and Erandi K De Silva at the Department of Molecular Biology and the Lewis-Sigler Institute for Integrative Genomics, Princeton University; and Thomas Lavstsen and Thor Grundtvig Theander from the Centre for Medical Parasitology, Department of International Health, Immunology and Microbiology, University of Copenhagen.

Kasper Jensen performed the *in silico* molecular docking experiments at the Technical University of Denmark and the *in vitro* binding assays were performed by Manuel Llinas and Erandi K De Silva at Princeton University.

4.2 Background

The ApiAP2 family

The Apicomplexan Apetala2 (ApiAP2) family of putative transcription factors were discovered in 2005 by Balaji and co-workers [59]. Until then, only one other transcription factor had been identified in *Plasmodium falciparum*, PfMyb1 [60], and comparative genomics studies had failed to reveal any specific transcription factors with DNA binding domains similar to those known from other eukaryotic species [59].

The ApiAP2 proteins are found in all apicomplexan species studied to date and contain one or more Apetala2 (AP2) domains [59,61]. This domain is known from plant transcription factors, e.g. the AP2/ERF DNA-binding proteins, which comprises the second largest group of transcription factors in *Arabidopsis thaliana* [62].

In *P. falciparum* 26 ApiAP2 proteins have been identified and 22 of them were shown to be expressed in specific developmental stages duing the intraerythrocytic developmental cycle, suggesting that they could be involved in the regulation of stage-specific gene expression [59].

PF14_0633

The X-ray crystal structure of the AP2 domain in the plasmodial ApiAP2 protein PF14_0633 revealed that the AP2 domain contains three anti-parallel beta sheets supported by an alpha helix. DNA binding requires the formation of AP2 dimers in which the alpha helices of the two domains swap place to support the beta sheets of the other subunit (see figure 4.1). The beta sheets of the AP2 domains wrap into the major groove of the DNA double-helix and contain all the residues that are in contact with the bound DNA. These residues are Asn-72, Arg-74, Arg-88 and Ser-90 [63].

Using protein-binding microarrays, De Silva and co-workers showed that PF14_0633 binds the consensus DNA sequence TGCATGCA, with the core CATGC being most important [61]. Gene Ontology analysis showed that genes containing at least one instance of the binding site are most enriched for "cytoadherence to the

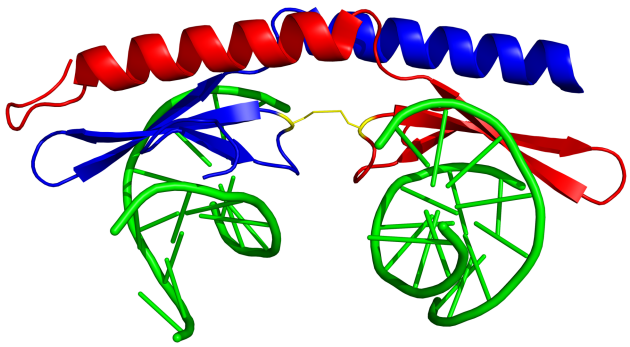


Figure 4.1. Two PF14_0633 AP2 domains (blue and red) in complex with double-stranded DNA (green). The AP2 domains are forming a dimer, which is stabilized by a disulfide bond (yellow) between Cys-76 of the two subunits.

microvasculature". Among the possible target genes were members of the UpsB and UpsC *var* gene subfamilies. Furthermore, the consensus sequence CATGCA was found to correspond to the SPE1 site, shown to be associated with subtelomeric UpsB promoters [64]. The expression of a nuclear factor binding to the SPE1 site was shown to correlate with the expression of subtelomeric *var* genes. In addition, the PF14_0633 binding motif was highly similar to a motif associated with genes whose expression peaked within a short window of the intraerythrocytic development cycle [65]. These data suggest that PF14_0633 are involved in the regulation of *var* gene expression. The PF14_0633 binding motif was also found upstream of another ApiAP2 protein, PFF0200c, suggesting a regulatory cascade of ApiAP2 transcription factors [61].

4.3 Methods

Protein-protein interaction network

Protein-protein interactions from the large scale yeast two-hybrid study by La Count and co-workers were downloaded from the publishers website [66]. Custom scripts were used to query the interactions.

Molecular docking

The crystal structure of PF14_0633 with PDB ID: 3IGM were retrieved from the Protein Data Bank [67]. The Tres Cantos data set was downloaded from the publishers website [68] and the DrugBank data set was downloaded from the DrugBank version 2.5 [69].

Molecular docking was performed using AutoDock with the lamarckian genetic algorithm used as the search algorithm [70]. All compounds were initially docked with a maximum of 25 000 energy evaluations and 20 repeats. The 1 000 best hits in each data set were docked again using a maximum of 250 000 energy evaluations and 100 repeats.

Gel-shift assays

Gel-shift assays were performed using the LightShift Chemiluminescent EMSA kit (Pierce). Briefly, 100 ng puried protein was incubated with varying concentrations of compound inhibitor in 1X binding buffer at room temperature for 10 minutes. 10 fmol of biotinylated probe was then added along with 50 ng poly(dI-dC), 5 mM $MgCl_2$, 1 mM EDTA, 50 mM KCl in binding buffer. Reactions were incubated at room temperature for a further 20 minutes before being separated on 6 % non-denaturing acrylamide gels run in 0.5X TBE. Probes were transferred to Hybond nylon membrane and visualized using the Chemiluminescent Nucleic Acid Detection Module (Pierce) according to the manufacturers instructions.

4.4 Results

Protein-protein interaction network of PF14_0633

Except for an AT-hook in PF14_0633, the ApiAP2 proteins in *P. falciparum* have not been found to contain any other known functional domains from the Pfam repository [61]. This has lead to the speculation that they exert their functions through protein-protein interactions.

To investigate this we created a protein-protein interaction network for PF14_0633 (figure 4.2). We used data from a yeast two-hybrid study, which is the only large scale study of protein-protein interactions in *P. falciparum* [66].

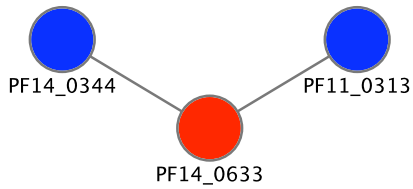


Figure 4.2. Protein-protein interaction network of PF14_0633.

The network was very small and contained only two interaction partners for PF14_0633. One of them, PF11_0313, were the 60S ribosomal protein P0, thought to be part of the large ribosomal subunit [71].

The other, PF14_0344, were the translocon component PTEX150. This protein is part of the *Plasmodium falciparum* translocon of exported proteins (PTEX), which is located in the vacuole membrane and mediates export of proteins containing the PEXEL motif [72].

The same two interaction partners were found when querying the STRING database [73] and the MINT database [74].

We could not immediately draw any conclusion about the function of PF14_0633 based on these two interaction partners. It is possible that they are false positives but the interactions could also have functions we do not yet understand.

Molecular docking of small compounds to PF14_0633

In 2010, a large scale study of antimalarial activity of nearly two million compounds in GlaxoSmithKline's chemical library were published. Of the 1 986 056 compounds tested, 13 533 inhibited *P. falciparum* growth by at least 80 % at 2 μM concentration [68]. The chemical structure of the 13 533 compounds, also referred to as the Tres Cantos data set, were made publicly available.

DrugBank is a publicly available database of drug and drug target information [69]. Although the majority of drugs in DrugBank are not known to have antimalarial activity, it might well be that some of them have. If an already approved drug can be used to target malaria, a lot of developmental costs will be saved because the drug has already been extensively tested.

The 13 533 compounds in the Tres Cantos data set and 4 603 compounds from DrugBank were docked to PF14_0633 using AutoDock. The resulting free energy of binding of each compound is shown in figure 4.3 and figure 4.4. The sense strand of the DNA sequence known to bind PF14_0633 were used as a positive control of docking, and the free energy of binding was plotted as a function of the distance

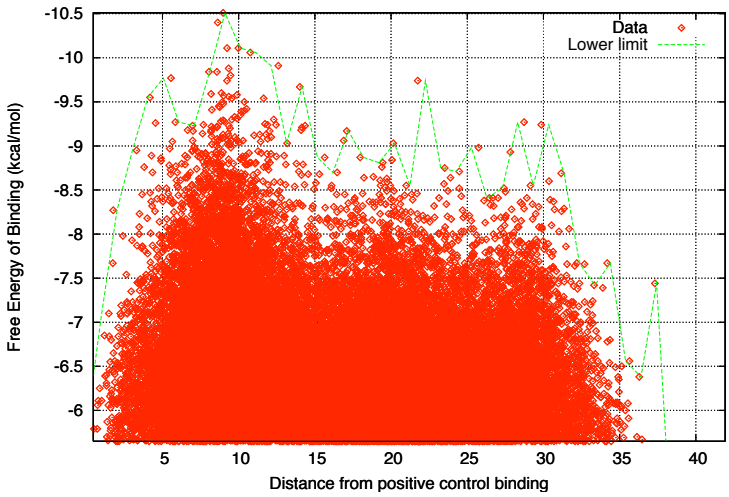


Figure 4.3. The free energy of binding of Tres Cantos compounds to PF14_0633.

to the geometric center of the bound DNA. The energy plots show that several compounds bound strongly within a distance of 5–10 Å from the positive control. The peak in free energy at this distance suggests that a binding pocket might exist within 5–10 Å of the DNA binding site.

Figure 4.5 shows the result of docking TCMDC-124220 from the Tres Cantos data set to PF14_0633. The figure reveals a binding pocket at the interface between the two ApiAP2 domains. Furthermore, it can be seen that TCMDC-124220 extends out of the binding pocket and possibly interferes with the residues Asn-72 and Arg-74 of one of the ApiAP2 domains. Both of these residues are involved in DNA binding and TCMDC-124220 could be interfering with the DNA-binding capabilities of PF14_0633.

We were interested in identifying compounds that could interfere with the DNA binding ability of PF14_0633. Hence, we selected the best binders within a short distance of the positive control. In addition, we screened the binding data for compounds with a geometric center no more than 5 Å from the center of at least two of the DNA binding residues in PF14_0633, and selected those with the lowest free energy of binding. The final list of compounds selected for *in vitro* testing are shown in table 4.1.

Unfortunately, not all the compounds we selected were commercially available or in the case of DB02633, it was only available in solution cross-linked to agarose beads. In some of these cases, we were able to find available compounds with a structure very similar to the ones we had selected. We accepted compounds with

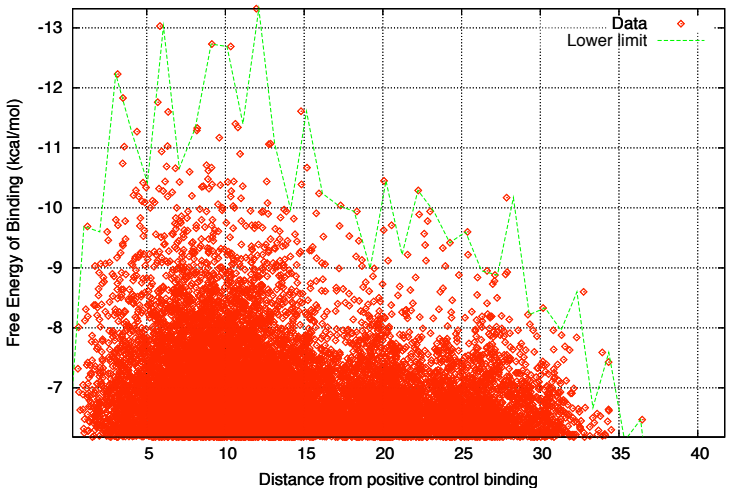


Figure 4.4. The free energy of binding of DrugBank compounds to PF14_0633.

a vendor listed in PubChem [75] and with a Tanimoto score of at least 0.9 to the original compound. The Tanimoto score is a measure of the similarity between two compounds (or their fingerprints) and is on the scale 0.0–1.0, where a score of 1.0 means that the fingerprints are identical. Those compounds were also docked to PF14_0633 and the best ones were included in the list for *in vitro* testing.

The *in silico* docking experiments and identification of interesting compounds for *in vitro testing* were performed by Kasper Jensen.

In vitro testing of protein-compound binding

The compounds shown in table 4.1 were tested for binding to PF14_0633 using gel-shift assays. The compounds were incubated at different concentrations with the purified PF14_0633 protein. The results are shown in figure 4.6 and figure 4.7. The blue arrows indicate binding between the compound and PF14_0633.

Most of the tested compounds were able to bind PF14_0633. DB00562 showed binding at only 125 nM while DB02633 bound at 0.2 μM. CID5751169, CID5750730, CID4541005, CID1365835, AG-690/09705007, Procion blue MX-R and DB04640 all showed binding at 2.5 μM. Only TCMDC-123924, DB04409 and DB01219 did not show any binding activity at the tested concentrations.

The gel-shift assays were performed by Manuel Llinas and Erandi K De Silva at Princeton University.

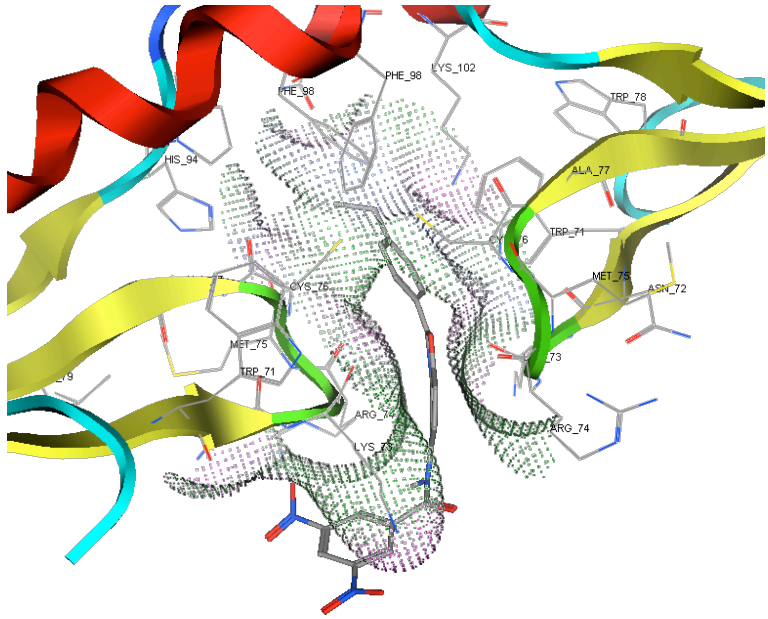


Figure 4.5. TCMDC-124220 docked to PF14_0633.

4.5 Discussion

The study of the protein-protein interaction network for PF14_0633 revealed two interaction partners (PF14_0344 and PF11_0313). Although we could not immediately draw any conclusion about the function of PF14_0633 based on these interaction partners, it is possible that they are involved in functions we do not yet understand.

However, the complete protein-protein interaction network that the interactions were extracted from covers only 25 % of all predicted proteins in *Plasmodium falciparum* [66]. Hence, it is possible that PF14_0633 may have other interaction partners and exerts its function through them.

We performed *in silico* molecular docking on the Tres Cantos data set and on a data set of compounds from DrugBank. Many compounds docked with a low free energy and in proximity to the DNA binding residues of PF14_0633. The binding between the compounds and PF14_0633 was confirmed *in vitro* using gel-shift assays. 9 out of 12 tested compounds were able to bind PF14_0633 at concentrations of 2.5 μM or lower.

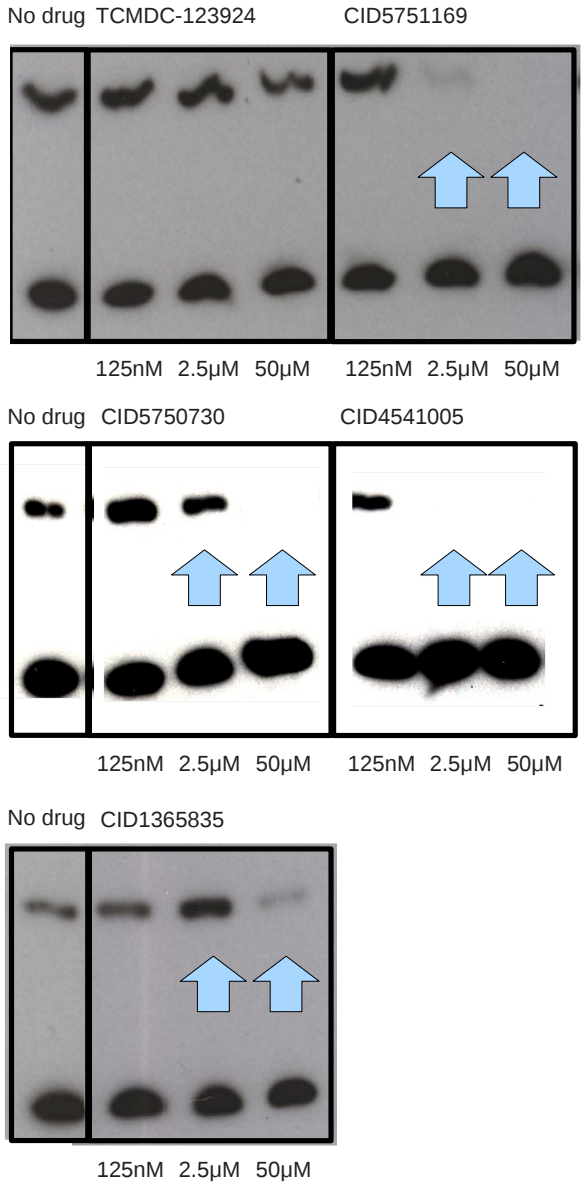


Figure 4.6. Gel-shift assays of five compounds selected based on the Tres Cantos data set. The compounds were incubated at different concentrations (shown below each lane) with 100 ng purified PF14_0633 protein. The blue arrows indicate evidence of binding between the compound and the protein.

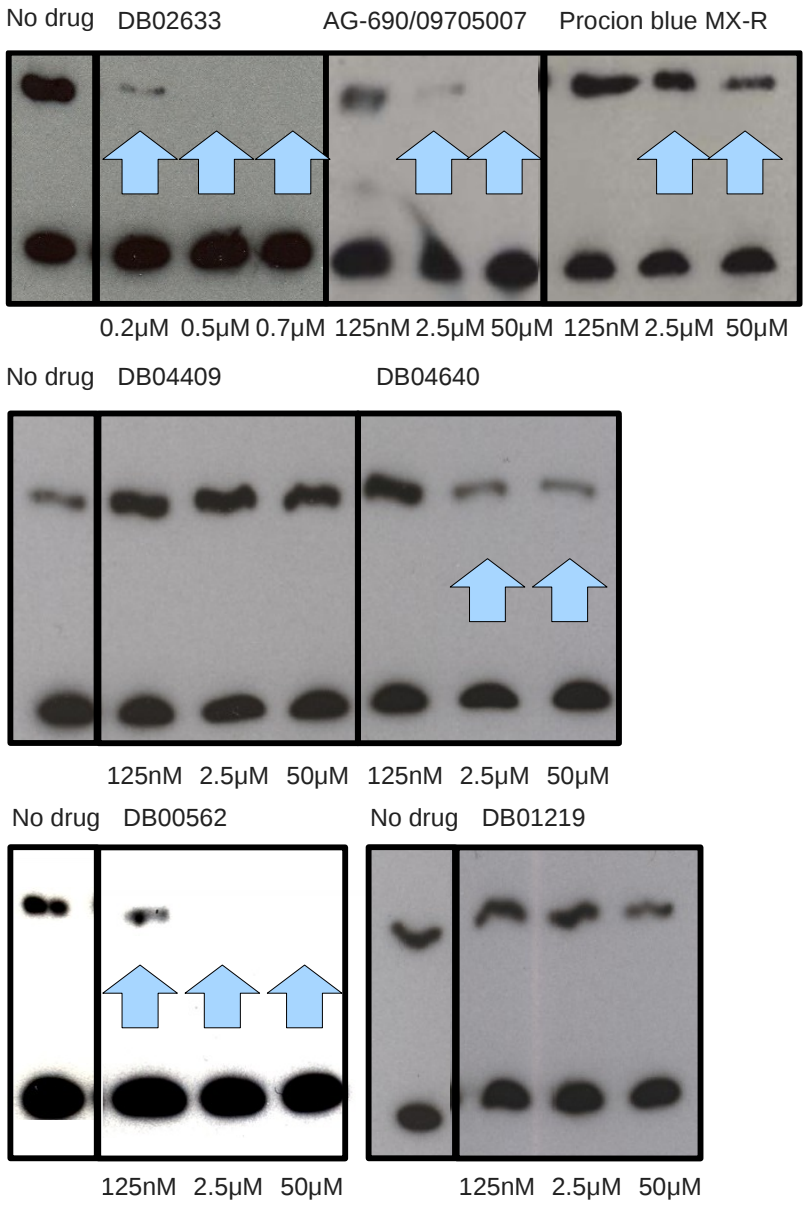


Figure 4.7. Gel-shift assays of seven compounds selected based on the DrugBank data set. The compounds were incubated at different concentrations (shown below each lane) with 100 ng purified PF14_0633 protein. The blue arrows indicate evidence of binding between the compound and the protein.

Tested compound	Selected compound	Tanimoto	Therapeutic area
TCMDC-123924	TCMDC-123924	1.00[1]	-
CID5751169	TCMDC-124220	0.90[1]	-
CID5750730	TCMDC-124220	0.93[1]	-
CID4541005	TCMDC-124220	0.92[1]	-
CID1365835	TCMDC-124220	0.91[1]	-
DB02633	DB02633	1.00[2]	HIV-1
AG-690/09705007	DB02633	0.92[2]	HIV-1
Procion blue MX-R	DB02633	0.92[2]	HIV-1
DB04409	DB04409	1.00[2]	Tumours
DB04640	DB04640	1.00[2]	Parasitic/bacterial infections
DB00562	DB00562	1.00[2]	High blood pressure/edema
DB01219	DB01219	1.00[2]	Muscle relaxant agent

Table 4.1. The compounds selected for *in vitro* testing. The tested compounds as well as the compounds originally selected are shown as well as the Tanimoto score between the two compounds. For the DrugBank compounds, the current therapeutic area according to DrugBank is also shown. 1: Tanimoto score based on CACTVS keys, 2: Tanimoto score based on MACCS keys.

The Tres Cantos compound TCMDC-124220 was shown to inhibit parasite growth by at least 80 % at 2 μM concentration [68]. Furthermore, molecular docking predicted that it could bind PF14_0633. Although we were not able to obtain this compound from any vendor, we obtained four structurally similar compounds that were also predicted to bind PF14_0633. All of these compounds were able to bind PF14_0633 *in vitro* at 2.5 μM concentration. Hence, it is likely that TCMDC-124220 also binds PF14_0633.

Figure 4.5 shows the docking of TCMDC-124220 to PF14_0633. From the figure it can be seen that one end of TCMDC-124220 extends out of the binding pocket between the two ApiAP2 domains and possibly interferes with the residues Asn-72 and Arg-74, both of which are involved in DNA binding. In this manner, TCMDC-124220 might be interfering with the DNA-binding capabilities of PF14_0633.

Whether this is the reason for the inhibition of parasite growth by TCMDC-124220 is hard to say, and more experiments are needed to test this hypothesis. The prediction that PF14_0633 regulates *var* genes and other genes involved in cytoadherence to the microvasculature does not seem to support this, as the importance of these genes should be somewhat limited when the parasites are grown under laboratory conditions. However, PF14_0633 was also predicted to regulate another ApiAP2 protein, PFF0200c, which in turn was predicted to regulate late-stage genes involved in the critical process of preparing the parasite for host cell rupture and re-invasion [61]. If this is the case, TCMDC-124220 might pre-

vent PF14_0633 from activating these downstream genes and thereby halt parasite growth.

The compound DB04640 which is used in experimental treatment of parasitic and bacterial infections, were able to bind PF14_0633 at 2.5 μM concentration. According to DrugBank, this compound is also known to bind lactate dehydrogenase. Interestingly, lactate dehydrogenase is also the target of the well-known antimalarial chloroquine [76].

The next step will be to confirm whether the compounds that are able to bind PF14_0633 interfere with its DNA binding capabilities. If they do, it will be very interesting to see the downstream effects of this. One possibility is that it kills the parasites. If the parasites are not killed, it will be interesting to study the downstream effect on gene expression. This will give us further insight into the function of PF14_0633 and gene regulation *P. falciparum*. Ultimately, these studies will show whether PF14_0633 is a suitable drug target in malaria parasites.

Part III

MHC Classification

Chapter 5

Classification of MHC-binding peptides

5.1 Introduction

Short peptides binding to major histocompatibility complex (MHC) proteins are important for correct activation of our immune system in response to various pathogens. This binding is a specific event since a particular MHC molecule will only bind certain peptides. The genetic variation within the MHC region is enormous, leading to a wide variety in the binding capacities of MHC proteins. The experimental determination of which peptides are bound by which MHC alleles has been an ongoing effort for many years [77, 78] and computational efforts to predict peptide binding has followed [79, 80].

The aim of this project was to develop a computational method that is able to distinguish several binding motifs in a mixture of peptides binding to different MHC molecules or to a single molecule but in different binding modes. Experimental evidence suggests, that at least some MHC molecules are able to bind peptides in different modes, or in other words, that they are characterized by two or more distinct binding motifs. This has been shown for the MHC class I allele HLA-A*0101 [81–83] and the MHC class II allele HLA-DR3 [84–87].

Being able to distinguish the different binding modes of MHC molecules should enable us to better predict peptide binding by MHC molecules. Some binding modes might lead to stable peptide binding while others could be more unstable and a precise definition of the binding motifs could help us understand these issues. This might in turn aid the development of peptide based vaccines, by focusing on

stable peptide binders and avoiding performing assays on unstable binders with a very low immunogenicity.

This project was done in collaboration with Massimo Andreatta and Morten Nielsen from the Immunological Bioinformatics group at the Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark.

5.2 Background

Antigen presentation by the major histocompatibility complex proteins

The proteins of the major histocompatibility complex (MHC), in humans referred to as human leukocyte antigen (HLA) proteins, are proteins involved in the presentation of peptides on the cell surface to the immune system.

Proteins from intracellular pathogens as well as the organisms own proteins are degraded by the proteasome to yield small peptide fragments. Some of these are transported into the endoplasmatic reticulum by the transporter associated with antigen presentation (TAP). Here they are presented to partially folded MHC class I molecules, which will complete folding upon successful binding of a short peptide. This process is facilitated by the tapasin protein. The MHC class I molecule in complex with the bound peptide is exported to the cell surface where it can be recognized by the T-cell receptor of $CD8^+$ cytotoxic T lymphocytes and thereby initiate an adaptive immune response. The presentation of peptides by MHC class I molecules happens on all nucleated cells [79, 88].

Extracellular antigens are taken up by the antigen presenting cells of the immune system (e.g. macrophages, B lymphocytes and dendritic cells). The proteins are being degraded by proteases in specialized compartments such as phagosomes, endosomes and lysosomes before they are transferred to the MHC-II containing compartment where they are loaded onto MHC class II molecules. The MHC class II molecules are folded and pre-loaded with a CLIP fragment, which is exchanged for the antigenic peptide. This process is facilitated by the chaperone DM. The MHC class II molecule in complex with the bound peptide is transported to the cell surface where they can be recognized by $CD4^+$ T helper cells. Cross-presentation—the presentation of extracellular anitigens by MHC class I proteins—also occurs [88, 89].

The binding of a peptide to an MHC molecule is a specific event; a particular MHC molecule will only bind certain peptides (only 0.5–2 % of the peptides within a given source protein will bind to a given MHC molecule [90, 91]), and this specificity is important for the specificity of the adaptive immune system as it is the peptide in complex with the MHC molecule that is recognized by the T-cell receptor.

Peptide motifs

A set of peptides binding to a particular MHC molecule can be summarized in a weight matrix describing the amino acid preference at each position in the peptide motif. The weight matrix contains the log-odds scores for each amino acid at each position in the motif:

$$W_{ia} = 2\log_2 \frac{p_{ia}}{q_a}, \tag{5.1}$$

where p_{ia} is the frequency of amino acid a at position i in the motif and q_a is the background frequency of amino acid a.

The peptides have to be aligned, which is straight forward if the peptides all have the same length. Peptides binding to MHC class I molecules are typically 9 amino acids long although they can vary a little. Peptides binding to MHC class II molecules are of different lengths and can be aligned using Gibbs sampling [92,93]. The part of the alignment that is important for binding specificity of the peptide to the MHC molecule is referred to as the core and corresponds to the binding motif.

The peptides known to bind a specific MHC molecule is most likely just a sample of all the peptides that can bind this MHC molecule. As a result, some amino acids that are allowed in certain positions in the motif, might not have been observed in these positions. This problem, referred to as *low counts*, can be somewhat corrected for by the inclusion of pseudocounts. This involves the addition of a small count to all the counts in the frequency matrix. The simplest would be to add one to all counts but it is also possible to calculate a more sophisticated pseudocount based on the observed frequencies and substitution frequencies from e.g. the Blosum substitution matrices (see methods, section 5.3) [16, 94].

Data sets of peptides known to bind a specific MHC molecule might be biased and contain an overrepresentation of some sequence signals. This could result in the erroneous impression that a particular amino acid might be more important at a given position than is actually the case. A number of methods can be used to account for redundancy in the sequences including homology reduction [95], clustering [92] and position based sequence weighting [96]. In homology reduction, sequences are removed from the data set in such a way that none of the remaining sequences are more similar to each other than a predefined threshold. In clustering, similar sequences are clustered together and assigned a weight reciprocal to the size of the cluster it belongs to. In position based sequence weighting, each sequence is assigned a weight depending on how similar it is to the other sequences in the data set.

Sequence logos

A peptide motif can be visualized in a convenient way using a sequence logo [97]. A sequence logo consists of a stack of letters for each position in the sequence. The

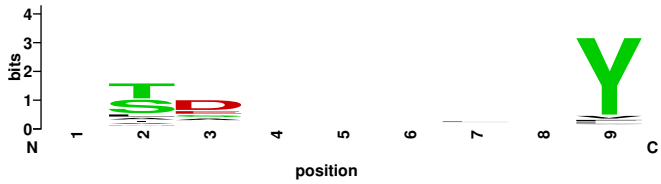


Figure 5.1. Sequence logo for the MHC class I allele HLA-A*0101. The logo was created from 300 peptides known to bind this protein. The logo shows conservation of position 2, 3 and 9 within the motif. Position 9 is most conserved (it has the highest information content) with a strong preference for tyrosine (Y) at this position.

height of the stack is a measure of the information content (i.e. the conservation) on that position and the height of each letter within the stack indicates the relative frequency of that amino acid at that position. The information content at a given position, i, is defined as the difference between the maximum possible entropy and the observed entropy at that position [97]:

$$I_i = \log_2(N) + \sum_a p_{ia} \log_2(p_{ia}). \tag{5.2}$$

Here, p_{ia} is the observed (possibly corrected) frequency of amino acid a at position i in the alignment. N is the total number of different amino acids, which is 20. The summation is over the 20 amino acids. An example sequence logo for the MHC class I allele HLA-A*0101 is shown in figure 5.1.

Akaike's information criterion

Akaike's information criterion (AIC) is a measure of the relative Kullback-Leibler distance between a model and the unknown true mechanism that generated the observed data [98]. It is given by:

$$AIC = -2\log(\mathcal{L}(\hat{\theta}|data)) + 2k, \tag{5.3}$$

where $\mathcal{L}(\hat{\theta}|data)$ is the likelihood of the estimated parameter values, $\hat{\theta}$, given the *data* at its maximum point and k is the number of estimable parameters [98, 99]. For a given set of candidate models, the best model is the one with the lowest AIC value. The term $2k$ is a penalty term that prevents overfitting by the inclusion of too many parameters in the model.

If there are too many parameters compared to the sample size, n, AIC may perform poorly. In this case a correction for small (or finite) sample size should

be added [100]:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}. \tag{5.4}$$

The extra term in equation 5.4 is merely an extra penalty on the number of parameters in relation to the sample size. If the sample is large compared to the number of parameters, this term will be negligible and AIC should perform well. A rule of thumb is that if $n/k < 40$ then AIC_c should be used [98].

Given a set of R models and AIC (or AIC_c) values for each model, it is possible to calculate the probability that each model is best in terms of Kullback-Leibler distance. This requires that there are only R models and that one of them must be the best model [98]. These probabilities are also referred to as the Akaike weights, w_i $(i = 1...R)$, for the models.

5.3 Methods

Data sets

Peptides known to bind specific MHC alleles were downloaded from the Immune Epitope Database (IEDB) [101].

A neural network was trained to learn several peptide motifs simultaneously. The network was trained with data from 1 000 peptides from each allele (both binders and non-binders). 100 000 random peptides were then scored with the neural network and the top 1 % were regarded as the best binders. These 1 000 peptides were then used as a data set. The test data set consisted of 2 000 peptides. This work was performed by Massimo Andreatta.

Simulated data were generated manually by creating count matrices with high counts on the desired positions. Random counts on the remaining positions were either chosen manually or copied from positions in existing motifs without any significant information content.

The first set of simulated data sets contained 40 data sets to be clustered. These data sets consisted of peptides generated from 14 different frequency matrices described in table 5.1. The frequencies in matrix number 2 were used as the background frequencies for matrix 3–14. These matrices then had one inflated amino acid on one position as shown in the table. Logos for the 14 weight matrices are shown in appendix D, (figure D.1 – figure D.14).

The 40 data sets were created by generating random peptides using the frequencies in the matrices. These peptides were then mixed as shown in table 5.2. Notice that data set 1–30 are balanced data sets containing 100 peptides from each matrix, data set 31–36 are balanced data sets containing 300 peptides from each matrix, while data set 37–40 are unbalanced data sets containing 100 peptides from one matrix and 300 from another.

Matrix	Description
1	flat, all amino acids have the same frequency
2	flat, but with small random variation in the frequencies
3	N on position 2: $\times 5$
4	N on position 2: $\times 10$
5	N on position 2: $\times 20$
6	N on position 2: $\times 40$
7	L on position 2: $\times 5$
8	L on position 2: $\times 10$
9	L on position 2: $\times 20$
10	L on position 2: $\times 40$
11	N on position 3: $\times 5$
12	N on position 3: $\times 10$
13	N on position 3: $\times 20$
14	N on position 3: $\times 40$

Table 5.1. Simulated matrix 1–14. Matrix number 2 was used as the background for matrix 3–14. These matrices had one amino acid on one position inflated compared to the background. The table shows that in matrix number 3, for example, N on position 2 is 5 times more frequent than the background frequency. N: Asparagine, L: Leucine.

The second set of simulated data sets contained 18 data sets to be clustered. These data sets consisted of peptides generated from 9 different frequency matrices described in table 5.3. These matrices had randomized background frequencies taken from existing matrices to get more realistic background frequencies. One or two positions in these matrices were inflated as described in the table. Logos for the 9 weight matrices are shown in appendix D, (figure D.15 – figure D.23).

The 18 data sets were created by generating random peptides using the frequencies in the matrices. These peptides were then mixed as shown in table 5.4. Notice that data set 101–106 are balanced data sets containing 100 peptides from each matrix, data set 107–112 are balanced data sets containing 300 peptides form each matrix, while data set 113–118 are unbalanced data sets containing 100 peptides form on matrix and 300 from another.

In the predictive performance approach to model selection, a measured affinity value is needed in order to calculate the correlation. For the simulated data, these affinities (a_S) were constructed using:

$$a_S = -\sum_{i=1}^{l} \log(f_{ia}), \tag{5.5}$$

where f_{ia} is the frequency of amino acid a at position i in the peptide and l is the

Data set	Description
1	matrix 1 (100)
2	matrix 2 (100)
3	matrix 2 (100) + matrix 3 (100)
4	matrix 2 (100) + matrix 4 (100)
5	matrix 2 (100) + matrix 5 (100)
6	matrix 2 (100) + matrix 6 (100)
7	matrix 2 (100) + matrix 3 (100) + matrix 07 (100)
8	matrix 2 (100) + matrix 4 (100) + matrix 08 (100)
9	matrix 2 (100) + matrix 5 (100) + matrix 09 (100)
10	matrix 2 (100) + matrix 6 (100) + matrix 10 (100)
11	matrix 2 (100) + matrix 3 (100) + matrix 11 (100)
12	matrix 2 (100) + matrix 4 (100) + matrix 12 (100)
13	matrix 2 (100) + matrix 5 (100) + matrix 13 (100)
14	matrix 2 (100) + matrix 6 (100) + matrix 14 (100)
15	matrix 3 (100)
16	matrix 4 (100)
17	matrix 5 (100)
18	matrix 6 (100)
19	matrix 3 (100) + matrix 7 (100)
20	matrix 4 (100) + matrix 8 (100)
21	matrix 5 (100) + matrix 9 (100)
22	matrix 6 (100) + matrix 10 (100)
23	matrix 3 (100) + matrix 11 (100)
24	matrix 4 (100) + matrix 12 (100)
25	matrix 5 (100) + matrix 13 (100)
26	matrix 6 (100) + matrix 14 (100)
27	matrix 3 (100) + matrix 7 (100) + matrix 11 (100)
28	matrix 4 (100) + matrix 8 (100) + matrix 12 (100)
29	matrix 5 (100) + matrix 9 (100) + matrix 13 (100)
30	matrix 6 (100) + matrix 10 (100) + matrix 14 (100)
31	matrix 1 (300)
32	matrix 2 (300)
33	matrix 3 (300) + matrix 7 (300) + matrix 11 (300)
34	matrix 4 (300) + matrix 8 (300) + matrix 12 (300)
35	matrix 5 (300) + matrix 9 (300) + matrix 13 (300)
36	matrix 6 (300) + matrix 10 (300) + matrix 14 (300)
37	matrix 3 (100) + matrix 11 (300)
38	matrix 4 (100) + matrix 12 (300)
39	matrix 5 (100) + matrix 13 (300)
40	matrix 6 (100) + matrix 14 (300)

Table 5.2. Simulated data set 1–40. The table shows which matrices were used to generate the peptides for each of the 40 simulated data sets. The number of peptides from each matrix is shown in parenthesis.

Matrix	Description
101	strong signal on pos 2 og 3
102	strong signal on pos 2 og 3 (differs from matrix101)
103	strong signal on pos 6 og 7
104	strong signal on pos 2
105	strong signal on pos 2 (differs from matrix104)
106	strong signal on pos 6
107	medium signal on pos 2
108	medium signal on pos 2 (differs from matrix107)
109	medium signal on pos 6

Table 5.3. Simulated matrix 101–109.

Data set	Description
101	matrix 101 (100) + matrix 102 (100)
102	matrix 101 (100) + matrix 103 (100)
103	matrix 104 (100) + matrix 105 (100)
104	matrix 104 (100) + matrix 106 (100)
105	matrix 107 (100) + matrix 108 (100)
106	matrix 107 (100) + matrix 109 (100)
107	matrix 101 (300) + matrix 102 (300)
108	matrix 101 (300) + matrix 103 (300)
109	matrix 104 (300) + matrix 105 (300)
110	matrix 104 (300) + matrix 106 (300)
111	matrix 107 (300) + matrix 108 (300)
112	matrix 107 (300) + matrix 109 (300)
113	matrix 101 (300) + matrix 102 (100)
114	matrix 101 (300) + matrix 103 (100)
115	matrix 104 (300) + matrix 105 (100)
116	matrix 104 (300) + matrix 106 (100)
117	matrix 107 (300) + matrix 108 (100)
118	matrix 107 (300) + matrix 109 (100)

Table 5.4. Simulated data set 101–118. The table shows which matrices were used to generate the peptides for each of the 18 simulated data sets. The number of peptides from each matrix is shown in parenthesis.

length of the peptide. This yielded a_S values in the range 23 – 28 for the binders, which corresponds to very strong binders.

Clustering of peptides

Peptides were clustered using two different clustering algorithms. The first and simplest one, referred to as *clustering algorithm 1*, randomly assigns the peptides to different clusters. Then, it iteratively updates the clusters based on how well the peptides score against the weight matrix of each cluster. The algorithm is as follows:

Step 1: Assign each peptide to a random cluster.

Step 2: Calculate a weight matrix for each cluster based on the peptides assigned to that cluster.

Step 3: Score all peptides against all weight matrices and assign each peptide to the cluster where it gets the highest score.

Step 4: Evaluate whether the stop criterion has been met. If not, go to step 2.

The input to the program is a list of peptides and the output is the assignment of the peptides to the different clusters. We started with the easiest case of clustering MHC class I peptides that were all of the same length (9 amino acids) so no alignment of the peptides were needed. The number of clusters were also fixed. The stop criteria were either that no sequences changed cluster or that a maximum of 10 000 iterations had been reached.

The other clustering algorithm, referred to as *clustering algorithm 2*, is similar to the above but uses a Monte Carlo Gibbs sampling approach to escape from local minima in the energy landscape. The algorithm is as follows:

Step 1: Assign each peptide to a random cluster.

Step 2: Select a random peptide and remove it from its cluster (C_0).

Step 3: Update the weight matrix of C_0.

Step 4: Select a new random cluster (C_n).

Step 5: Score the peptide against both C_0 and C_n to get the scores S_0 and S_n respectively.

Step 6: Assign the peptide to C_n with probability P given below; otherwise assign it to C_0:

$$P = \min\left(1, \exp\left(\frac{\Delta E}{T}\right)\right), \tag{5.6}$$

where $\Delta E = S_n - S_0$ and T is the temperature, which is lowered in regular steps during the iterations.

Step 7: Evaluate whether the stop criterion has been met. If not, go to step 2.

Again, the algorithm requires pre-aligned peptides and the number of clusters is also fixed. The stop criterion was that a maximum number of iterations had been reached. This was set to 100 times the number of peptides times the number of temperature steps (10). The algorithm was implemented by Massimo Andreatta.

Weight matrices

Weight matrices were constructed using equation 5.1. When pseudocounts were used, the effective amino acid frequencies were calculated according to [16]:

$$p_{ia} = \frac{\alpha \times f_{ia} + \beta \times g_{ia}}{\alpha + \beta}. \tag{5.7}$$

Here, α is the effective number of sequences minus one in the alignment, β is the weight on pseudocounts, f_{ia} is the observed frequency of amino acid a at position i and g_{ia} is the pseudocount of amino acid a at position i. If no sequence weighting is performed, the effective number of sequences is equal to the number of sequences.

Pseudocounts were calculated using [94]:

$$g_{ia} = \sum_b f_{ib} \times q(a|b), \tag{5.8}$$

where f_{ib} is the observed frequency of amino acid b at position i and $q(a|b)$ is the substitution frequency for amino acid a, conditional on the observation of amino acid b obtained from the Blosum62 substitution matrix [102].

Sequence weighting was performed using the weighting scheme proposed by Henikoff and Henikoff [96]. The weight on peptide k is given by:

$$w_k = \sum_i \frac{1}{r_i \times s_{ia}}, \tag{5.9}$$

where r_i is the number of different amino acids at position i in the motif and s_{ia} is the number of occurrences of amino acid a at position i in the motif. When this scheme was used for sequence weighting, the effective number of sequences, α, was calculated as the mean number of different amino acids on each position in the alignment:

$$\alpha = \frac{1}{L} \sum_i r_i, \tag{5.10}$$

where L is the length of the sequence motif.

Sequence logos

Sequence logos were created using WebLogo [103].

Akaike weights

Akaike's information criterion (AIC and AIC_c) was calculated using equation 5.3 and equation 5.4 respectively. Akaike weights, also referred to as model probabilities, were calculated as described by Burnham and Anderson [98]. Briefly, given a set of R models, the difference in AIC (or AIC_c) between each model, i ($i = 1...R$), and the best model, AIC_{min}, i.e. the model with the lowest AIC, was calculated as:

$$\Delta_i = AIC_i - AIC_{min}. \tag{5.11}$$

The relative likelihood of each model given the data was then calculated as:

$$\mathcal{L}(model_i|data) \propto \exp\left(-\frac{1}{2}\Delta_i\right), \tag{5.12}$$

where $\propto$ means proportional to. The Akaike weights, w_i, were then calculated by normalizing the relative likelihoods:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta_r)}. \tag{5.13}$$

The Akaike weights are positive values summing to one and are therefore also interpreted as the model probabilities [98].

Multidimensional small sample AIC

Fujikoshi and Satoh's small sample AIC for multivariate linear regression [104]:

$$CAIC = AIC + 2\frac{K(k+1+p)}{n-k-1-p}, \tag{5.14}$$

where $K = kp+p(p+1)/2$, k is the number of regressors, n the number of samples and p the number of regressions.

Burnham and Anderson's hypothesized multidimensional small sample AIC [98]:

$$AIC_c = AIC + 2\frac{K(K+v)}{np-K-v}, \tag{5.15}$$

where v is the number of distinct parameters in the variance-covariance matrix and falls in the interval $1 \leq v \leq p(p+1)/2$, K is the total number of parameters, n the number of samples and p the number of dimensions.

Predictive performance

The predictive performance of a model was evaluated using an independent test set of peptides containing both binders and non-binders to each of the alleles that

were clustered. Any peptides included in the clustering were excluded from the test set. The peptides in the test set had a measured binding affinity against one or more of the clustered alleles. If a peptide was present multiple times, i.e. if its binding affinity had been measured against several of the alleles, the one with the best binding affinity was selected for the test set. The binding affinities were transformed to be on a scale between 0 and 1 using the formula [105]:

$$y = 1 - \frac{\log(a)}{\log(50\,000)}, \tag{5.16}$$

where a is the measured affinity and y is the transformed affinity. High binding peptides with a measured affinity stronger than 50 nM will have a value above 0.638 after this transformation and intermediate binders with a measured affinity stronger than 500 nM will have a value above 0.426.

All peptides in the test set were scored against all the clusters in a particular clustering using the log-odds matrices representing the clusters. Each peptide was assigned to the cluster where it got the highest log-odds score. The correlation between the log-odds scores and the transformed affinity values, y, was then calculated. Both Pearson's product-moment correlation coefficient, r_P, and Spearman's rank correlation coefficient, r_S, were calculated using R [106].

5.4 Results

Clustering of peptides

We implemented a clustering algorithm, *clustering algorithm 1* (see methods, section 5.3), which clustered the peptides into a predefined number of clusters.

We used clustering algorithm 1 to cluster a balanced data set of 500 peptides, with 100 peptides known to bind each of the 5 MHC class I alleles HLA-A*0201, HLA-A*0301, HLA-B*0702, HLA-B*1501 and HLA-B*4402. The outcome of clustering these sequences using both sequence weighting and pseudocounts in the weight matrix is shown in table 5.5. The algorithm terminated after 12 iterations. As can be seen from table 5.5, the assignment of peptides to the clusters appear to be more or less random. Furthermore, the result was not reproducible as another run of the program ended up with completely different clusters. We ran the program with different values of β (the weight on pseudocounts) and we also tried turning off sequence weighting and/or pseudocounts but it did not make any difference. The results still appeared to be completely random.

However, when the peptides were assigned to the correct clusters from the beginning, they more or less stayed in those clusters. See table 5.6 for an example.

We also tried to cluster the strong binders only (binding affinity < 50 nM) as we observed relatively less misclassifications among them. However, when they were randomly distributed from the beginning, the result still looked more or less random.

Cluster	HLA-A*0201	HLA-A*0301	HLA-B*0702	HLA-B*1501	HLA-B*4402
1	7	11	29	26	18
2	12	11	8	6	49
3	45	15	20	22	12
4	23	6	41	31	9
5	13	57	2	15	12

Table 5.5. The result of clustering peptides from the five MHC class I alleles HLA-A*0201, HLA-A*0301, HLA-B*0702, HLA-B*1501 and HLA-B*4402 into five clusters. The table shows how many peptides from each allele that ended up in each of the five clusters. The program terminated after 12 iterations. Pseudocounts and sequence weighting were used.

Cluster	HLA-A*0201	HLA-A*0301	HLA-B*0702	HLA-B*1501	HLA-B*4402
1	93	0	1	11	0
2	0	98	0	1	0
3	2	0	97	6	0
4	5	2	1	80	0
5	0	0	1	2	100

Table 5.6. The result of clustering peptides from the five MHC class I alleles HLA-A*0201, HLA-A*0301, HLA-B*0702, HLA-B*1501 and HLA-B*4402 into five clusters when they were assigned to the correct clusters in the first iteration. The table shows how many peptides from each allele that ended up in each of the five clusters. The program terminated after 4 iterations. Pseudocounts were not included in this example but sequence weighting was.

We modified the algorithm such that only one peptide was moved in each iteration (step 3 in clustering algorithm 1). In other words, the weight matrices were updated every time a peptide was moved to a new cluster instead of when all peptides had been moved. This did not improve the clustering either.

Based on these results we came to the conclusion that the algorithm probably got stuck in a local minimum, when trying to find the optimal clustering. Hence, we decided to use a Monte Carlo Gibbs sampling approach to get out of local minima.

Clustering of peptides using Gibbs sampling

We implemented *clustering algorithm 2* (see methods, section 5.3), which cluster the sequences into a predefined number of clusters and use Gibbs sampling to get

Cluster	HLA-A*0101	HLA-A*0301	HLA-B*4402
1	295	6	1
2	2	0	99
3	3	44	0

Table 5.7. The result of clustering peptides from the three alleles: HLA-A*0101, HLA-A*0301 and HLA-B*4402 into three clusters. The table shows how many peptides from each allele that ended up in each of the three clusters. Sequence weighting and pseudocounts were used.

out of local minima in the energy landscape. The algorithm was implemented by Massimo Andreatta.

We created a data set of peptides consisting of 300 peptides known to bind the HLA-A*0101 allele, 50 peptides known to bind the HLA-A*0301 allele and 100 peptides known to bind the HLA-B*4402 allele. All peptides were 9 amino acids long. The data set was clustered using clustering algorithm 2. The clustering was repeated ten times with parameters set such that each clustering resulted in a different number of clusters, starting with one cluster and ending with ten clusters. Sequence weighting and pseudocounts were used in this and all subsequent clusterings.

The result of the clustering that resulted in three clusters are shown in table 5.7. The table shows how the peptides binding the different alleles were distributed among the three clusters. Cluster 1 mainly contains peptides from the HLA-A*0101 allele, cluster 2 mainly contains peptides from the HLA-B*4402 allele and cluster 3 mainly contains peptides from the HLA-A*0301 allele. If the three clusters each are said to represent the allele from which it contain the most peptides, the algorithm is able to classify 97 % of the peptides to their original allele. The remaining peptides are not necessarily classified wrongly because they might actually also bind the other alleles, whether experimentally tested or not.

The problem is to select the correct number of clusters, since the number of clusters is controlled by the parameters of the clustering algorithm. We tried two different approaches described below. The first approach is based on Akaike's information criterion (AIC) as a basis for model selection. The second approach is based on the predictive performance of the resulting weight matrices.

Model selection with Akaike's information criterion

Each result of the clustering was considered to be a model that described the data. The model consisted of a number of clusters represented by the frequency matrices. As a result of running the clustering algorithm ten times yielding from one to ten clusters, we had ten models which were considered to cover the model space.

The likelihood of each model was calculated by scoring each peptide, o, against the raw frequency matrix of the cluster, c, it was assigned to. The observed, unadjusted frequency of each amino acid at each position is the maximum likelihood estimate of the probability of observing that particular amino acid at that position. The frequencies for each position in a peptide was multiplied to get the overall probability of that peptide, and all the peptide probabilities were then multiplied to get the overall model probability. For convenience and to avoid underflow problems, the logarithm of the frequencies were summed instead of multiplying the frequencies themselves:

$$\log \mathcal{L}(\hat{\theta}|data) = \sum_{o=1}^{O} \sum_{i=1}^{l} \log f_{ia}^{oc}, \tag{5.17}$$

where f_{ia}^{oc} is the frequency of amino acid a at position i in peptide o belonging to cluster c. The total number of peptides is O and l is the length of the peptide.

The number of estimable parameters in a model was calculated as:

$$k = cl(m-1), \tag{5.18}$$

where c is the number of clusters, l is the length of the peptides and m is the number of amino acids. Since the number of amino acids is fixed and so is the length of the peptides for this study, k is given by:

$$k = cl(m-1) = c \times 9 \times (20-1) = 171c. \tag{5.19}$$

The number of samples, n, were set equal to the number of amino acids in all peptides in the data set.

AIC and AIC_c was calculated for all the considered models and the model probabilities (or Akaike weights), w, were then calculated according to equation 5.13. The result of clustering the three alleles HLA-A*0101, HLA-A*0301 and HLA-B*4402 is shown in table 5.8. The model that is assigned the highest probability when using AIC is the one with seven clusters. However, since $n/k < 40$, AIC should be corrected for small sample size and it is more correct to use AIC_c, which assigns the highest probability to the model with only four clusters.

Figure 5.2 shows the logos of the three alleles that were clustered. If we compare these with the logos of the four clusters that were suggested by the AIC_c criterion for model selection (figure 5.3) we see that they are very similar except that HLA-A*0101 have been split into two clusters (figure 5.3a and figure 5.3b). The two HLA-A*0101 clusters are very similar and does not correspond to the two different binding modes of HLA-A*0101 specific binding peptides that have been observed [81–83]. According to these studies, the two binding modes are characterized by a preference for tyrosine (Y) at their C terminus and either serine (S),

c	n	k	$\mathcal{L}(\hat{\theta}\|data)$	AIC	AIC_c	w	w_c
1	4 050	171	−10 451.6	21 245.3	21 260.3	0.000000	0.000000
2	4 050	342	−9 878.3	20 440.7	20 503.7	0.000000	0.000000
3	4 050	513	−9 575.6	20 177.2	20 326.2	0.000000	0.000000
4	4 050	684	−9 277.0	19 921.9	20 199.9	0.000000	0.999999
5	4 050	855	−9 029.7	19 769.3	20 227.3	0.000000	0.000001
6	4 050	1 026	−8 778.9	19 609.8	20 306.8	0.000000	0.000000
7	4 050	1 197	−8 568.6	19 531.2	20 536.2	1.000000	0.000000
8	4 050	1 368	−8 439.1	19 614.3	21 011.3	0.000000	0.000000
9	4 050	1 539	−8 302.8	19 683.6	21 571.6	0.000000	0.000000
10	4 050	1 710	−8 079.3	19 578.6	22 079.6	0.000000	0.000000

Table 5.8. Comparison of models and model probabilities for models with different number of clusters. c: number of clusters in the model, n: sample size, k: estimable parameters, $\mathcal{L}(\hat{\theta}|data)$: likelihood, AIC: Akaike's information criterion, AIC_c: AIC corrected for small sample size, w: Akaike weight or model probability based on AIC, w_c Akaike weight or model probability based on AIC_c.

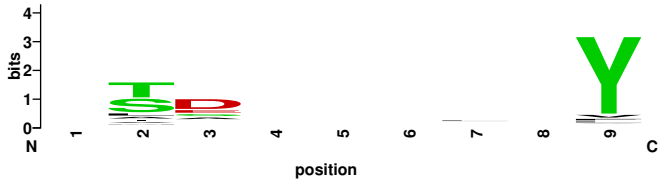
threonine (T) or glutamic acid (E) in position 2 or aspartic acid (D) or glutamic acid (E) in position 3. The two motifs we identified did not distinguish clearly between those two binding modes although the relative importance of position 3 differed a little.

Model selection using predictive performance

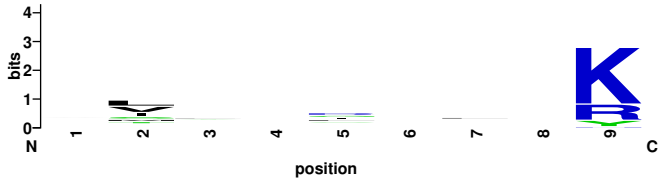
We selected an independent test set of peptides containing both binders and non-binders to each of the three alleles HLA-A*0101, HLA-A*0301 and HLA-B*4402 that were clustered above.

All peptides in the test set were scored against all the clusters in a particular clustering using the log-odds matrices representing the clusters. Each peptide was assigned to the cluster where it got the highest log-odds score. Finally, the correlation between the log-odds scores and the transformed affinity values, y, was calculated. We calculated both Pearson's product-moment correlation coefficient, r_P, and Spearman's rank correlation coefficient, r_S. The results for the clustering of the three alleles HLA-A*0101, HLA-A*0301 and HLA-B*4402 are shown in figure 5.4.

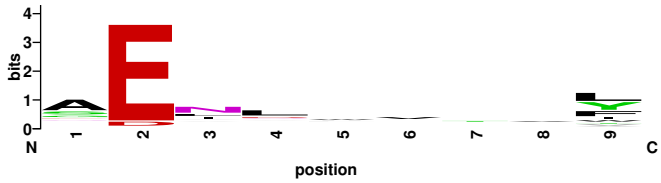
From figure 5.4 it can be seen that both the Pearson and the Spearman correlation peaked at 3 clusters suggesting that this was the optimal way to cluster the sequences.



(a) Logo of the allele HLA-A*0101



(b) Logo of the allele HLA-A*0301



(c) Logo of the allele HLA-B*4402

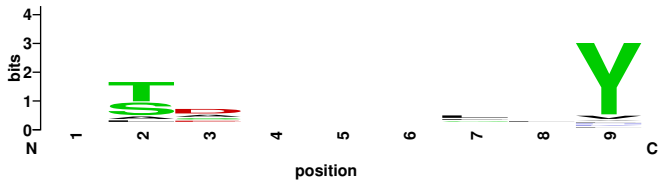
Figure 5.2. Logos of three MHC class I alleles.

Clustering balanced data sets

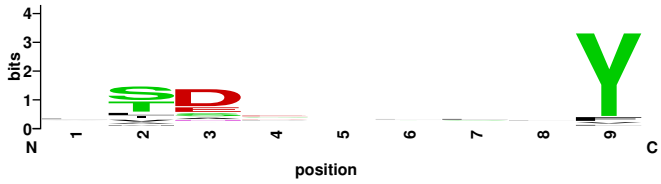
We created five data sets based on peptides belonging to known MHC motifs. The data sets contained peptides mixed from 1 to 5 different MHC motifs as shown in table 5.9. Each data set was clustered ten times resulting in one to ten clusters. We used both the AIC based approach and the predictive performance approach to model selection, and the optimal number of clusters suggested by each approach is shown in table 5.10. Again AIC suggested more clusters than there were alleles in each data set, whereas the other methods ended up with the original number of clusters except for the 5 allele case where AIC_c only suggested 4 clusters.

Clustering based on neural network data

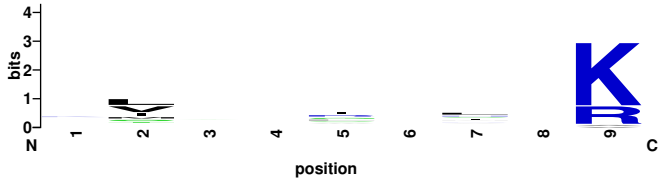
We created two artificial data sets using a neural network. For the first data set, the neural network was trained to learn two different peptide motifs (HLA-A*0101 and HLA-A*0201). For the second data set, three different peptide motifs were learned (HLA-A*0101, HLA-A*0201 and HLA-A*0301). 100 000 random peptides were then scored with the neural network and the top 1 % were regarded as good



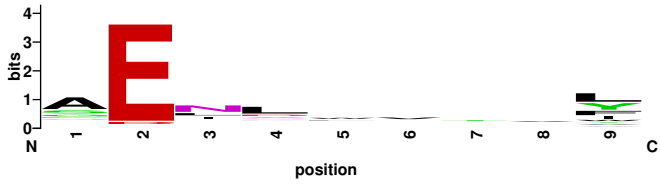
(a) Logo of cluster 1



(b) Logo of cluster 2



(c) Logo of cluster 3



(d) Logo of cluster 4

Figure 5.3. Logos of the four clusters suggested by AIC_c (see text).

Data set	Alleles	# peptides
1	HLA-A*0201	100
2	HLA-A*0301	100
	HLA-B*0702	100
3	HLA-A*0101	100
	HLA-A*0301	100
	HLA-B*4402	100
4	HLA-A*0201	100
	HLA-A*0301	100
	HLA-B*0702	100
	HLA-B*4402	100
5	HLA-A*0101	100
	HLA-A*0201	100
	HLA-A*0301	100
	HLA-B*0702	100
	HLA-B*4402	100

Table 5.9. Data sets used to test the model selection algorithms. The five data sets were created by mixing peptides from the listed motifs.

Data set	# alleles	AIC	AIC_c	r_P	r_S
1	1	2	1	1	1
2	2	4	2	2	2
3	3	7	3	3	3
4	4	8	4	4	4
5	5	9	4	5	5

Table 5.10. Result of clustering the five balanced data sets shown in table 5.9. The table shows the number of clusters suggested by using Akaike's information criterion or predictive performance for model selection. AIC: Akaike's information criterion, AIC_c: AIC corrected for small sample size, r_P: Pearson correlation, r_S: Spearman correlation.

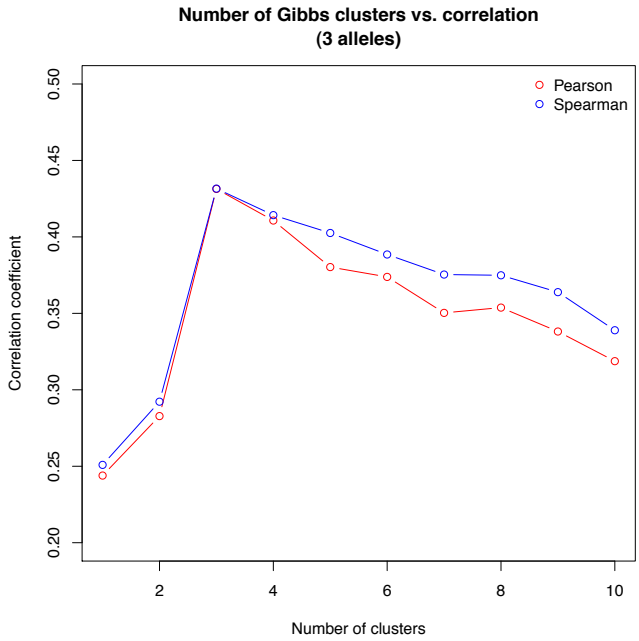


Figure 5.4. The result of using the predictive performance approach to select the optimal number of clusters after clustering peptides from the three alleles HLA-A*0101, HLA-A*0301 and HLA-B*4402.

binders to the two motifs. These 1 000 peptides were then clustered with clustering algorithm 2 and Akaike's information criterion and the correlation coefficients of the predictive performance was evaluated. The neural networks were developed by Massimo Andreatta for another project and we just used them here to generate the two data sets.

The result is shown in table 5.11. AIC og AIC_c suggests 9-10 clusters for these data, whereas the predictive performance suggests the correct number of clusters when the Pearson correlation is considered. The Spearman correlation suggests only 2 clusters for the case when 3 motifs were used, although the correlation coefficients for 2 and 3 clusters are very similar (figure 5.5).

Simulated data sets

The simulated data sets were created as described in the methods (section 5.3). Data set 1–40 were clustered and the number of clusters were evaluated using

Data set	Alleles	AIC	AIC_c	r_P	r_S
1	2	10	9	2	2
2	3	9	9	3	2

Table 5.11. The result of clustering two data sets based on a neural network. The table shows the number of clusters suggested by using Akaike's information criterion or predictive performance for model selection. AIC: Akaike's information criterion, AIC_c: AIC corrected for small sample size, r_P: Pearson correlation, r_S: Spearman correlation.

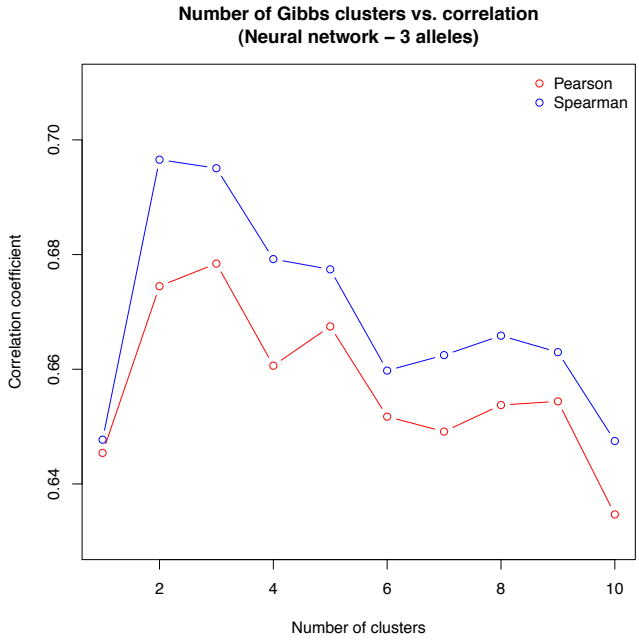


Figure 5.5. The result of using the predictive performance approach to select the optimal number of clusters in the neural network data for three alleles.

Akaike's information criterion and predictive performance (table 5.12). For the predictive performance approach, test data were created in the same way as the clustered data. 300 peptides were generated from each matrix.

Data set 101–118 were also clustered and the number of clusters were evaluated using Akaike's information criterion and predictive performance (table 5.13). For the predictive performance approach, the test data were created by generating random peptides from a flat distribution and then scoring them against the matrices. 3000 random peptides were created for each matrix.

The general observation from these data were that AIC tended to predict too many clusters while AIC_c predicted the expected number of clusters as long as there were only 100 peptides from each allele in the data set (data set 1–30 and 101–106). When there were more than 100 peptides from each allele in the data set (data set 31–40 and 107–118) AIC_c tended to predict too many clusters and the number of clusters seemed to correlate with the total size of the data set.

The predictive performance approach generally performed poorly on the simulated data sets. Typical correlation curves for these data are shown in figure 5.6. This was probably because of the artificial construction of the affinity values. Another possibility would be to use log-odds scores instead of the constructed affinity values. However, the method is based on the correlation between experimentally measured affinity values and the predicted log-odds scores. Any result that we could produce by using scores that are somehow related to the weight matrices would not validate the method. Therefore, it does not make sense to validate the predictive performance using simulated data sets.

In the calculations of AIC_c, the number of samples were equated with the number of amino acids in all peptides in the data set. This is one way to count the number of samples in multiple alignments, but it is not obvious that it is the correct way. In particular, the samples are not independent when counting the amino acids. First of all, the amino acids are linked together in sequences, making them dependent on each other, and the sequences also depend on each other through their evolutionary relationships. These dependencies makes it very hard to determine the effective sample size and the choice might influence the result of model selection with AIC_c as discussed by Posada and Buckley [107].

To see if our choice of sample size had an effect on the dependency of AIC_c on the size of the data set, we calculated AIC_c for all the data sets above with the sample size equal to the number of sequences in the data set (AIC_c^1). However, this did not work well and did not remove the size dependency of AIC_c (appendix D, table D.2 – table D.4).

Another way to consider the data is to consider each sequence as a sample in multiple dimensions with the number of dimensions equal to the length of the sequence. AIC is also valid for multidimensional data but AIC_c is not generally valid in this case. No generally valid AIC corrected for small sample size exists for the multidimensional case, however, Fujikoshi and Satoh have developed a small sam-

Data set	Alleles	AIC	AIC_c	r_P	r_S
1	1	2	1	-	-
2	1	2	1	9	9
3	2	5	2	2	2
4	2	4	2	1	2
5	2	4	2	1	1
6	2	4	2	1	1
7	3	8	3	1	2
8	3	8	3	1	1
9	3	6	3	2	2
10	3	7	3	1	1
11	3	8	3	1	1
12	3	7	3	1	1
13	3	7	3	2	2
14	3	7	3	2	2
15	1	2	1	1	1
16	1	3	1	1	1
17	1	2	1	1	1
18	1	2	1	1	2
19	2	5	2	1	1
20	2	5	2	1	1
21	2	5	2	3	2
22	2	5	2	2	1
23	2	4	2	1	1
24	2	5	2	2	2
25	2	5	2	2	2
26	2	4	2	2	1
27	3	6	3	1	1
28	3	8	3	1	1
29	3	6	3	2	2
30	3	6	2	3	2
31	1	6	3	-	-
32	1	7	3	2	2
33	3	10	10	1	4
34	3	10	10	3	2
35	3	10	8	3	3
36	3	10	10	3	7
37	2	8	3	8	1
38	2	10	4	1	1
39	2	9	4	3	3
40	2	9	3	2	2

Table 5.12. The result of clustering simulated data set 1–40. The table shows the number of clusters suggested by using Akaike's information criterion or predictive performance for model selection. AIC: Akaike's information criterion, AIC_c: AIC corrected for small sample size, r_P: Pearson correlation, r_S: Spearman correlation.

Data set	Alleles	AIC	AIC_c	r_P	r_S
101	2	2	2	1	1
102	2	2	2	1	1
103	2	2	2	1	1
104	2	3	2	1	1
105	2	3	2	1	1
106	2	3	2	1	1
107	2	8	4	1	1
108	2	8	4	1	1
109	2	9	5	1	1
110	2	7	4	1	1
111	2	9	5	1	1
112	2	9	5	1	1
113	2	7	3	1	1
114	2	6	3	1	1
115	2	7	3	1	1
116	2	7	2	2	2
117	2	6	4	1	1
118	2	6	3	2	2

Table 5.13. The result of clustering simulated data set 101–118. The table shows the number of clusters suggested by using Akaike's information criterion or predictive performance for model selection. AIC: Akaike's information criterion, AIC_c: AIC corrected for small sample size, r_P: Pearson correlation, r_S: Spearman correlation.

ple AIC for multivariate linear regression, termed $CAIC$, which is valid under the assumption that a general $p \times p$ variance-covariance matrix applies for the residual vector of each observation, where p is the number regressions (equation 5.14) [104]. In addition, Burnham and Anderson hypothesized a generally valid multidimensional small sample AIC, simple termed AIC_c (equation 5.15) [98].

We also calculated these measures of the multidimensional small sample AIC although $CAIC$ is not valid for our case as it is not a multiple linear regression. Burnham and Anderson's multidimensional small sample AIC depends on the number of parameters, v, in the variance-covariance matrix, which is in the interval $1 \leq v \leq p(p+1)/2$. Since we do not know this value, the small sample AIC was calcuated for v equal to each of the boundary conditions (1 (AIC_c^2) and $p(p+1)/2$ (AIC_c^4)) and the midpoint (($(p(p+1)/2 - 1)/2$ (AIC_c^3)). As a result, a total of four new estimates of the small sample AIC were calcualted for each data set (appendix D, table D.2 – table D.4). In general, these multidimensional small sample AIC values agreed with each other and with the univariate AIC_c (equation 5.4)

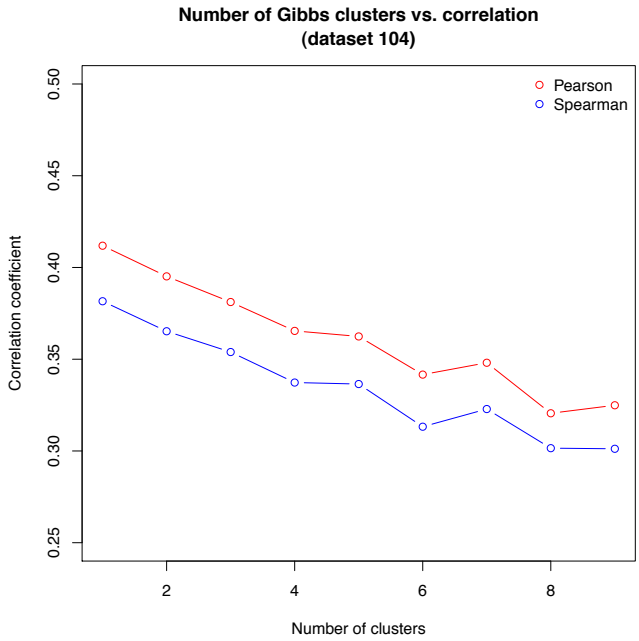


Figure 5.6. The result of using the predictive performance approach to select the optimal number of clusters in simulated data set 104. This is a typical example of how the correlation curves looked like for the simulated data sets.

although they made a few more mistakes than the univariate AIC_c. However, they depended on the size of the data set, the same way that the univariate AIC_c did.

When calculating AIC and AIC_c above, the observed amino acid frequencies were used to calculate the likelihood as these are the maximum likelihood estimates for the probabilities of observing a particular amino acid. We also calculated AIC (AIC^5) and AIC_c (AIC_c^5) using frequencies that had been adjusted for pseudo-counts and sequence weighting to calculate the likelihood (appendix D, table D.1 – table D.4). The result was that AIC usually estimated the correct number of clusters, while AIC_c estimated too few clusters. However, the dependency of the size of the data set was still present.

5.5 Discussion

We developed a clustering algorithm based on Gibbs sampling that were able to cluster mixed data sets of peptides belonging to different MHC class I alleles. The number of clusters had to be selected before running the algorithm and we tested two different approaches for selecting the optimal number of clusters after the clustering had been performed. One approach was based on Akaike's information criterion (AIC) while the other was based on the predictive performance measured on independent data.

Due to the relatively small sample sizes compared to the number of estimable parameters in these models, we used a version of AIC corrected for small sample size, AIC_c. AIC_c performed well on both real and simulated data sets as long as only 100 peptides from each allele were included in the data set. When more than 100 peptides from each allele were included, AIC_c predicted too many clusters. We tested different versions of AIC_c for multidimensional data and we also attempted different ways of counting the number of samples or calculating the likelihood used to calculate AIC, but the dependency on the size of the data set did not disappear.

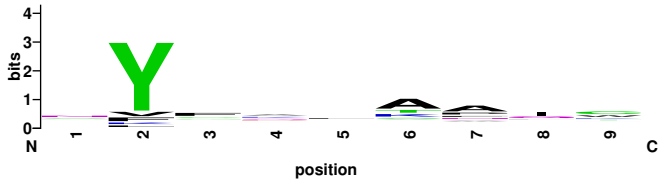
It is difficult to say why AIC has this dependency on the size of the data set but it was robust over the different ways we calculated AIC. We can not exclude that it is due to dependencies in the data. Another explanation could be that when more data are added, it becomes more likely that subsets of peptides share certain sub-patterns for random reasons, and that it is these sub-patterns that causes AIC to suggest a higher number of clusters.

It would need more investigation to confirm that this is actually the case, but as an example consider the clustering of the simulated data set number 115. It is a mixture of peptides from two matrices: 300 peptides from matrix 104 and 100 peptides from matrix 105. The clustering algorithm makes an almost perfect separation of the two data sets; only five mistakes are made when it is clustered into two clusters and four mistakes are made when it is clustered into three clusters. Nonetheless, AIC_c suggests that the optimal clustering is the one consisting of three clusters. Here, most of the 100 peptides from matrix 105 ends up in one cluster while the 300 peptides from matrix 104 are split into two clusters (see table 5.14).

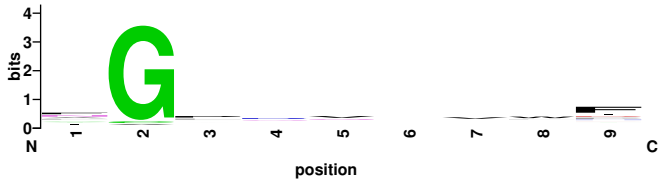
The logos for the three resulting clusters are shown in figure 5.7. When comparing the logos from cluster 1 and 3 it is obvious that, while the amino acid preferences for most positions are the same (strong preference for tyrosine (Y) at position 2 and no preference on positions 1, 3, 8 and 9), cluster 1 has a preference for alanine (A) on position 6 and 7, which is not seen in cluster 3, and cluster 3 has a preference for leucine (L) on position 4 and tryptophan (W) on position 5, which is not seen in cluster 1. These preferences, although weaker, are also visible in the logo for matrix 104 (figure D.18). But since there were no dependencies between the different positions when we simulated the data, it is due to random

Cluster	Matrix 104	Matrix 105
1	121	1
2	2	98
3	177	1

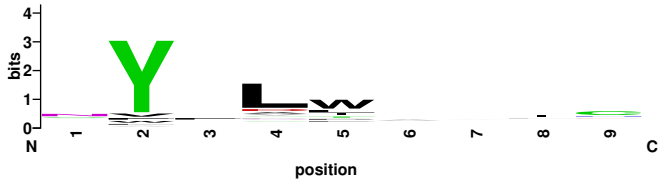
Table 5.14. The result of clustering peptides from matrix 104 and 105 into three clusters. The table shows how many peptides from each matrix that ended up in each of the three clusters.



(a) Logo of cluster 1



(b) Logo of cluster 2



(c) Logo of cluster 3

Figure 5.7. Logos of the clusters that results form clustering peptides from matrix 104 and 105 into three clusters.

reasons that they show up as sub-patterns in the peptide motifs.

The predictive performance, measured as the correlation between log-odds scores and measured binding affinities for an independent set of peptides, performed well on all the data sets that contained peptides with experimentally measured binding affinities. It did not work on the simulated data sets, but we think this is primarily because we were not able to create realistic artificial binding affinity for the peptides. The predictive performance did not show any dependency on

the size of the data set as *AIC* did, although this conclusion is based on very few observations, since we can not use those based on the simulated data sets. It would be beneficial to test the size dependency of both *AIC* and the predictive performance further on peptides with experimentally measured binding affinities to known MHC class I molecules.

The drawback of the predictive performance approach is that it requires that peptides with measured binding affinities against the MHC alleles in question are available. This is certainly not always the case and limits the usability of this method for detecting the correct number of clusters.

We were not able to detect any sub-motifs in the known MHC class I binding motifs. In one clustering, the HLA-A*0101 motif was split into two sub-motifs but they did not correspond to the previously observed sub-motifs for this allele [81–83].

The next step will be to develop the clustering algorithm to be able to deal with MHC class II motifs. Since these peptides are of different lengths, they will have to be simultaneously aligned and clustered. This can be achieved by a Gibbs sampling strategy, where the sequences are not only moved between clusters but also shifted back and forth in the alignment [93].

As mentioned in the introduction, it would be beneficial if we could detect different binding modes of MHC molecules. But the method would also be useful in the detection of antibody specificities. Blood serum from a patient with a response to a given disease will usually be poly-clonal. If the antibody specificities are investigated on a peptide chip, the result will be a mixture of peptides from antibodies with different specificities. Here our method could be used to identify the individual specificities in the blood serum. This could be useful e.g. in a study of cross-reactive antibodies.

Appendix

Appendix A

Paper I

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer

Thomas S. Rask[1,2]*, Daniel A. Hansen[1], Thor G. Theander[2], Anders Gorm Pedersen[1], Thomas Lavstsen[2]*

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark, **2** Centre for Medical Parasitology, Department of Medical Microbiology and Immunology, University of Copenhagen, Copehagen, Denmark

Abstract

The *var* gene encoded hyper-variable *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) family mediates cytoadhesion of infected erythrocytes to human endothelium. Antibodies blocking cytoadhesion are important mediators of malaria immunity acquired by endemic populations. The development of a PfEMP1 based vaccine mimicking natural acquired immunity depends on a thorough understanding of the evolved PfEMP1 diversity, balancing antigenic variation against conserved receptor binding affinities. This study redefines and reclassifies the domains of PfEMP1 from seven genomes. Analysis of domains in 399 different PfEMP1 sequences allowed identification of several novel domain classes, and a high degree of PfEMP1 domain compositional order, including conserved domain cassettes not always associated with the established group A–E division of PfEMP1. A novel iterative homology block (HB) detection method was applied, allowing identification of 628 conserved minimal PfEMP1 building blocks, describing on average 83% of a PfEMP1 sequence. Using the HBs, similarities between domain classes were determined, and Duffy binding-like (DBL) domain subclasses were found in many cases to be hybrids of major domain classes. Related to this, a recombination hotspot was uncovered between DBL subdomains S2 and S3. The VarDom server is introduced, from which information on domain classes and homology blocks can be retrieved, and new sequences can be classified. Several conserved sequence elements were found, including: (1) residues conserved in all DBL domains predicted to interact and hold together the three DBL subdomains, (2) potential integrin binding sites in DBLα domains, (3) an acylation motif conserved in group A *var* genes suggesting N-terminal N-myristoylation, (4) PfEMP1 inter-domain regions proposed to be elastic disordered structures, and (5) several conserved predicted phosphorylation sites. Ideally, this comprehensive categorization of PfEMP1 will provide a platform for future studies on *var*/PfEMP1 expression and function.

Citation: Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T (2010) *Plasmodium falciparum* Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer. PLoS Comput Biol 6(9): e1000933. doi:10.1371/journal.pcbi.1000933

Editor: Jonathan A. Eisen, University of California Davis, United States of America

Received April 3, 2010; **Accepted** August 16, 2010; **Published** September 16, 2010

Funding: This work was funded in part by a grant from the Foundation for the National Institutes of Health through the Grand Challenges in Global Health Initiative. This work also received support from the University of Copenhagen, Program of Excellence (Membrane Topology and Quaternary Structure of Key Parasite Proteins Involved in *Plasmodium falciparum* Malaria Pathogenesis and Immunity). Sequencing of the IT clone was funded by the European Union 6th Framework Program grant to the BioMalPar Consortium [grant number LSHP-LT-2004-503578]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rask@cbs.dtu.dk (TSR); thomasl@sund.ku.dk (TL)

Introduction

Plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1) mediates adhesion of infected erythrocytes (IE) to various host cells on the vascular lining, during the blood stage of malaria infection [1–2]. Naturally acquired protective antibodies in malaria-exposed individuals target PfEMP1, suggesting it is possible to develop PfEMP1 based vaccines [3–9].

The majority of the parasite's ~60 PfEMP1-encoding *var* genes are situated in subtelomeric regions close to other variant antigen-encoding genes such as the *rif* and *stevor* gene families, while the remaining ~40% are found centrally in the chromosomes. Based on sequence similarity, *var* 5′ UTR sequences can be divided into upstream sequence (UPS) classes A, B, C or E. These UPS classes correlate with chromosomal position of the genes, as well as domain complexity of the encoded PfEMP1 [10–11]. Subtelomeric UPSA and UPSB genes are oriented tail to tail (3′ to 3′), while central UPSC genes are oriented head to tail in a tandem repeat manner [12], which has lead to the definition of group A, B and C *var*/PfEMP1, and two intermediate groups B/A and B/C, that contain *var*/PfEMP1 with chromosomal position or domain composition different from that predicted from their UPS class. The hyper-variable *var* gene repertoire is to a large extent generated by frequent meiotic ectopic recombination in the mosquito abdomen, probably facilitated by alignment of *var* genes in the nuclear periphery [13–14]. There is also evidence suggesting that mitotic recombination occur, and that this allows further diversification of the *var* gene repertoire during human infection [15]. Comparison of the clones 3D7, IT4 and HB3 revealed only two *var* genes, *var1* and *var2csa*, that were conserved in all three genomes, and a semi-conserved gene, *var3*, found in IT4 and 3D7. The three conserved *var* genes are more than 75% identical over multiple domains, whereas most other PfEMP1 (even proteins with the same domain architecture) display less than 50% amino acid sequence identity between individual domains [16]. *Var2csa* is particularly unique as it has a unique UPSE, encodes unique Duffy binding-like (DBL) domains, as well as a distinct acidic terminal segment (ATS) [17].

Author Summary

About one million African children die from malaria every year. The severity of malaria infections in part depends on which type of the parasitic protein PfEMP1 is expressed on the surface of the infected red blood cells. Natural immunity to malaria is mediated through antibodies to PfEMP1. Therefore hopes for a malaria vaccine based on PfEMP1 proteins have been raised. However, the large sequence variation among PfEMP1 molecules has caused great difficulties in executing and interpreting studies on PfEMP1. Here, we present an extensive sequence analysis of all currently available PfEMP1 sequences and show that PfEMP1 variation is ordered and can be categorized at different levels. In this way, PfEMP1 belong to group A–E and are composed of up to four components, each component containing specific DBL or CIDR domain subclasses, which in some cases form entire conserved domain combinations. Finally, each PfEMP1 can be described in high detail as a combination of 628 homology blocks. This dissection of PfEMP1 diversity also enables predictions of several functional sequence motifs relevant to the fold of PfEMP1 proteins and their ability to bind human receptors. We therefore believe that this description of PfEMP1 diversity is necessary and helpful for the design and interpretation of future PfEMP1 studies.

Thus, parasite genomes appear to harbor essentially similar *var* repertoires, each reflecting the worldwide *var* diversity that has ensured the optimal survival of the parasite population. The clinical significance of the described *var* groups has been demonstrated in several studies, and indicates the existence of underlying functional differences in adhesion characteristics of the expressed PfEMP1 variants. This relationship is best illustrated by the malaria syndrome occurring in pregnant women, which is precipitated by the accumulation, in the placenta, of parasites expressing VAR2CSA that mediates binding to proteoglycans on syncytiotrophoblasts [17–21]. Several lines of evidence indicate that the relatively rapid development of immunity to severe childhood malaria is mediated through antibodies directed against a restricted semi-conserved subset of parasite antigens [22–23] that are associated with the development of severe disease [24–25]. In particular group A and to some extent group B *var* genes have been linked to disease severity in studies of expression of these variants in patients with symptomatic and asymptomatic infections [26–33]. A recent study has corroborated these findings and qualified which group A and B PfEMP1 variants may be associated with severe malaria disease, by demonstrating a sequential and ordered acquisition of antibodies to PfEMP1 domains in Tanzanian plasma donors [34].

In contrast to pregnancy malaria, it is still unclear which human receptor binding, if any particular, is linked to severe forms of childhood malaria. Parasite adhesion has been demonstrated to endothelial cells, immune system cells, uninfected erythrocytes and platelets. Several human cell receptors, including the extensively studied CD36 and intercellular adhesion molecule 1 (ICAM-1), have been implicated in adhesion, although no consensus on association between receptor binding and severe malaria has been reached (reviewed in [35]). PfEMP1 is responsible for parasite adhesion, as several single domains of the large multi-domain PfEMP1 molecules have been shown to bind human receptors. From N- to C-terminal, PfEMP1 has previously been described as composed of an N-terminal segment (NTS), Duffy binding-like (DBL) domains, Cys rich inter-domain regions (CIDR), C2 domains, one transmembrane region (TM) and the acidic terminal segment (ATS) (Figure 1A). Six major classes of DBL domains have been proposed based on amino acid sequence similarity: DBLα, β, γ, δ, ζ, and ε. DBL domains have been further characterized by definition of 10 semi-conserved homology blocks (HBa-j) interspersed by hyper-variable regions [36], and by definition of three structural subdomains (S1–3) [37] (Figure 1D). It has been shown that various DBLβ domains have affinity for ICAM-1 [38–40], whereas DBLδ adheres to platelet-endothelial cell adhesion molecule 1 (PECAM-1) and DBLα has been associated with binding to heparin sulfate (HS), blood group A antigen and complement receptor 1 (CR1) [41–42]. CR1 binding is associated with IE adhesion to uninfected erythrocytes, a phenomenon known as rosetting, which appears to be mediated to some degree by group A PfEMP1 [42–44].

CIDR domains have been divided into three classes: CIDRα, β, and γ [2,10,16,36], and described as consisting of three regions, those being the minimal CD36 binding region denoted M2, flanked by less conserved M1 and M3 regions [36,45]. Several CIDRα class domains have been found to mediate binding to the human CD36 receptor [1,45–46], however, such binding is limited to group B and C PfEMP1, indicating that group A variants have a distinct function [47]. Furthermore, CIDRα domains have been found to bind immunoglobulin M and PECAM-1 [41].

Although it is evident that the organization of PfEMP1 sequence diversity is of relevance for malaria pathogenicity, the vast sequence variation of the protein family continues to impede experimental procedures and interpretations. In order to better understand and determine the potential targets for a PfEMP1-based vaccine against severe malaria, it is therefore essential to establish a rigorous classification and solid reference frame of PfEMP1 diversity.

In this work, PfEMP1 repertoires from seven genomes are annotated with updated domain boundary definitions. The data includes four thoroughly sequenced *P. falciparum* genomes that have not previously been classified: DD2 from Indochina (9.55× coverage), RAJ116 from India (7.3× coverage), IGH-CR14 from India (10.19× coverage), and the Ghanaian isolate PFCLIN (8× coverage). Domain architectures of 399 PfEMP1 are aligned, revealing conserved domain architectural features. The homology block concept, first described by Smith *et al.* (2000) [36], is extended from DBL domains to the entire PfEMP1 by application of a novel iterative homology search technique, defining 628 homology blocks covering on average 83% of any PfEMP1 with only 4% self-overlap. The homology blocks describe relations between sequences in finer detail than domains, revealing that domain subclasses often consist of fragments from different domain super-classes, probably as a result of extensive recombination. Evidence for a recombination hotspot is also found. The definition of conserved blocks in PfEMP1 allows identification of conserved functional elements, such as predicted sites for post-translational modifications, which may significantly affect both substrate binding and immune evasion.

Results/Discussion

The *var* gene sequence analysis was based on two different bioinformatics approaches. First, phylogenetic trees were constructed using re-assessed PfEMP1 domain borders, with the aim of reclassifying and annotating the main PfEMP1 features UPS, NTS, DBL, CIDR and ATS. Secondly, a novel iterative homology detection method, defining a set of homology blocks, was used to describe domain similarities and to guide *var* gene recombination site and functional predictions.

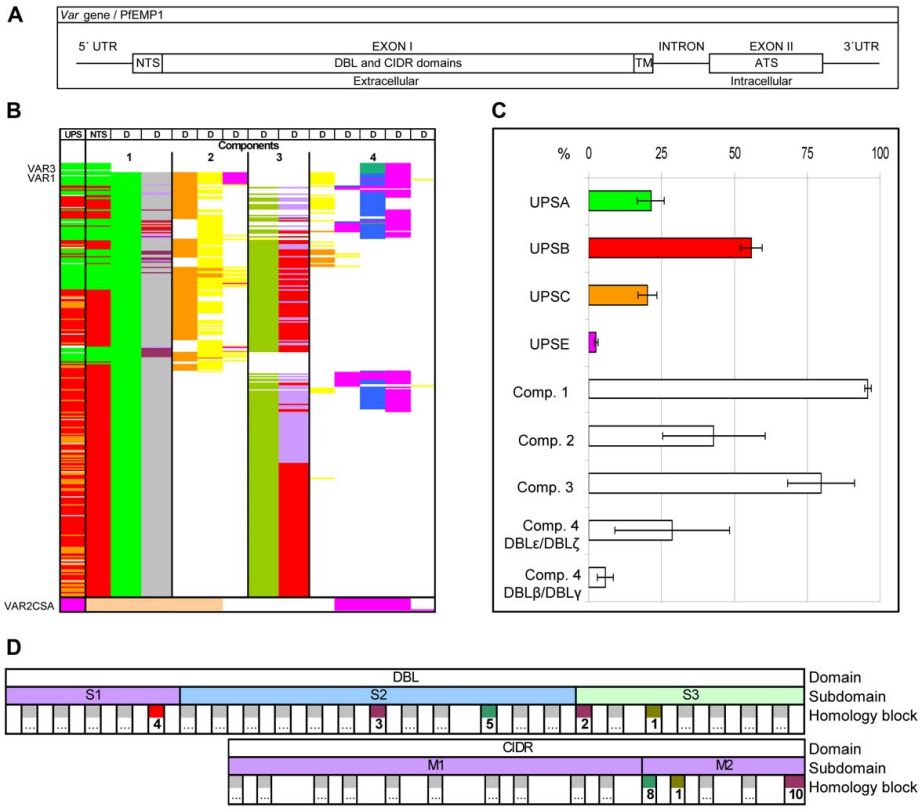


Figure 1. PfEMP1 annotation overview. (A) Schematic of the *var* gene locus. (B) 399 *var* exon1 annotated with UPS class and encoded major NTS, DBL and CIDR domain classes and their arrangement in four components. Color code for UPS column: Green: UPSA; Red: UPSB; Orange: UPSC; Pink: UPSE. Color code for NTS column: Green NTSA, Red: NTSB, Cream: NTSpam. Color code for DBL and CIDR domains (D columns): Bright Green: DBLα; Orange: DBLβ; Yellow: DBLγ; Olive green: DBLδ; Pink: DBLε; Blue: DBLζ; Blue stripes: DBLα of VAR3. Grey: CIDRα; Red: CIDRβ; Light purple: CIDRγ; Dark purple: CIDRδ. (C) Average distribution (% +/− 95% confidence intervals) of UPSA–E flanked and component 1–4 containing genes in the seven sequenced genomes 3D7, HB3, DD2, IT4, PFCLIN, RAJ116 and IGH. (D) Schematic presentation of DBL and CIDR subdomains and homology blocks. The numbered blocks represent the core homology blocks found in all DBL domains (HB2, 3, 4 and 5), all CIDR domains (HB8 and 10) or both domain types (HB1), further described in Figure 5.
doi:10.1371/journal.pcbi.1000933.g001

Grouping and componential composition of PfEMP1

In total 399 PfEMP1 sequences were annotated and their domains aligned. The alignments confirmed what recent studies of the DBL fold [48–49] and binding affinities [38] have implied; that the domain borders, by which PfEMP1 domain subclasses have been classified [36], needed revision. The redefined domain borders introduced by this study are specified in Text S1, and lead to two fundamental nomenclature changes: omitting the term "C2" from DBLβ domains, as also suggested in [39]; and the separation of M3 sequences from CIDR domains. Distance tree analysis of all DBL domains confirmed the expected phylogenetic grouping of DBL into six major classes (DBLα, β, γ, δ, ε, and ζ), as well as five smaller distinct classes (the four N-terminal DBL domains of VAR2CSA [10], and the DBLα of VAR3 which grouped in a separate cluster between DBLα and DBLζ). Five major CIDR domain classes were defined: CIDRα, β, γ, δ, and pam (Figure S1). The CIDRδ class has not previously been identified, probably due to the difference in sequence depth between this and previous CIDR classification (655 vs. 36) [36]. The inter-domain 2 (ID2) of VAR2CSA is partially homologous to CIDR domains [50], and was therefore included here as CIDRpam, although particularly different from other CIDR domains. NTS sequences were divided into three classes, NTSA, NTSB, and NTSpam (Figure S2L and Figure S3Y), while ATS sequences were divided into ATSA, ATSB, ATSpam, ATSvar1, and ATSvar3 (Figure S2M).

The 5′ upstream sequences of *var* genes were analyzed by two different methods: Markov clustering (MCL) [51–53], and neighbor joining (NJ) clustering (based on multiple sequence alignments). The two analyses yielded congruent trees, although additional subclusters could be identified in the NJ tree (Figure S2N and O). All previously suggested UPS subgroups [16] (UPSA1–2, UPSB1–4, UPSC1–2 and UPSE) could be identified, although with some modifications and four additional subgroups (UPSA3 and UPSB5–7).

Although the number of available *var* sequences varied between the seven studied genomes (39 to 63), the genomes contained similar *var* UPS distributions (Figure 1), and as expected, UPSE

flanked *var2csa*, NTSA and ATSA were exclusively encoded by UPSA flanked genes, whereas other NTS and ATS classes were found in UPSB and C flanked genes. The general observation that UPSA and UPSB genes are located head to head in the telomeres was also confirmed (data not shown), although only limited information on chromosomal location was available. Based on domain annotation of the extracellular part of PfEMP1 (Figure 1 and Figure S4), these could be described as consisting of four components: component 1 (present in ~95% of all PfEMP1) containing the N-terminal NTS-DBLα-CIDR domains, component 2 (present in ~43% of all PfEMP1) containing one to three DBLβ and DBLγ domains, component 3 (present in ~80% of all PfEMP1) containing DBLδ-CIDRβ/γ domains, and component 4 containing C-terminal domain combinations of DBLζ and DBLε domains (present in ~28% of all PfEMP1) or single DBLβ or DBLγ domains (present in ~8% of all PfEMP1). The complexity of domain structure followed the UPS classification, in agreement with established group A, B and C PfEMP1 nomenclature [10–11]. There was an overrepresentation of component 2 encoding genes in group A compared to group B or C *var* ($p<0.0001$; X^2 test of component 2 prevalence in group A–C), and component 4 was found in both group A and B but rarely C.

PfEMP1 inter-domain (ID) sequences were also aligned and classified. Most ID sequences were found to flank component 3, and characteristic for these sequences were long Pro-rich stretches, charged polyAsp/Glu stretches, and an amino acid composition biased towards Ala, Asp, Glu, Pro, Lys, and Val. The sequences downstream of component 3 could be classified, and were either of a M3A type if flanked by component 4, or M3AB if flanked by TM-ATS (Figure S3Z and Figure S4). Due to less functional constraints, ID sequences may have more relaxed requirements to the position of recombination break points, compared to within domains. The ID sequence variation supports the division of PfEMP1 into the four components, which suggest that the low-complexity ID sequence may act as recombination break points.

Inter-domain elasticity. The function of the ID sequences is unknown, although one possibility is that these regions confer elasticity to the PfEMP1 proteins, as suggested for similar sequence in the PEVK region of the human striated muscle protein titin (also known as connectin). The PEVK region of titin contains several PPAK domains, a 26–28 residue repeat consisting of low-complexity sequence biased towards Pro, Ala, Val, Lys, and Glu, and these domains are interspersed by polyGlu regions. The PEVK region length is correlated with elongation ability of sarcomeres in striated muscle [54], and the secondary structure has been found to be disordered [55].

The PfEMP1 ID regions are found in lengths up to ~200 residues, and the amino acid composition is very similar to the one found in titin PEVK. Hits to the Pfam PPAK domain definition [56] in four PfEMP1 supports the sequence similarity ($E<0.1$ in DD2var52, IT4var64, HB3var34 and PFCLINvar47). The acidic and basic residues can potentially form random structures based on polar interactions, mixed with Pro which introduces kinks in the protein backbone, together forming a structure with spring-like properties. Elasticity could enhance the ability of infected erythrocytes to adhere to endothelial cells by providing a smooth deceleration, as well as extend the time given to establish strong molecular interactions with targets. It is likely that the variant disordered structure of the inter-domain regions impede antibody targeting.

PfEMP1 groups contain specific subclasses of DBL and CIDR domains

The redefinition of domain borders, and the large increase in sequence data, called for a detailed subclassification of PfEMP1 domains. This was done by a distance tree analysis described in detail in Text S1, summarized for DBL and CIDR domains in Figure 2. The sequence diversity of the major DBL and CIDR domain classes differed both with respect to homogeneity (i.e. shared AA %-identity), and the degree to which subclasses could be distinguished. The previously observed division of DBLα into DBLα1 and DBLα0 [10–11] was confirmed, however a third distinct class of sequences, DBLα2, was also identified. Sequences of DBLα1 grouped relatively evenly into eight subclasses, including the particularly distinct DBLα1.3 of VAR3 (note description of nomenclature usage in Text S1), whereas the DBLα0 sequences spread more unevenly into 24 subclasses (Figure S2A,B and Figure S3I–K). The homology block analysis of VAR3 (described in the homology block section below) revealed that the N-terminal part of DBLα1.3 is similar to other DBLα domains, but interestingly, the C-terminal half of the domain is essentially a DBLζ3 domain. All DBLε and DBLζ domains grouped evenly into distinct subclasses, while DBLβ and DBLγ domains were divided into less distinct subclasses of varying sizes, and most (~90%) of DBLδ sequences could not be subclassified. The homogeneity of the six major classes differed with DBLβ domains being the most (45%) and DBLε the least (31%) homogenous classes. In particular subclasses DBLε1/2/11/13 were distinctively different from the majority of DBLε domains (Figure S2E and Figure S3N–Q). Similar to DBL domains, the level of subclassification of major CIDR domain types varied. Most members of CIDRα3.1 and CIDRβ subclasses could not be separated, whereas other CIDR domains grouped in evenly sized subclasses. The homogeneity of CIDR classes varied with CIDRα1 and CIDRδ domains exhibiting higher sequence similarities than the other CIDR classes. Sequence conservation logos for all large CIDR classes can be found in Figure S3A–H.

Annotation of the PfEMP1 using detailed DBL and CIDR subclassification (Figure S4) showed that most classes could be linked to a specific UPS class (Figure 2). When domain classes were found frequently in genes of more than one group, they were most often shared between group A and B or group B and C, but rarely A and C. These observations support the validity of the subclassification, and the notion that group A–C *var* genes predominantly recombine separately.

In conclusion, the phylogenetic domain analysis allowed classification of all PfEMP1 domains, and defined several novel domain classes. In addition, PfEMP1 domain variation was described in an unprecedented level of detail, by the allocation of the DBL and CIDR domains into subclasses. This classification is based on domain similarities averaged over the whole domains, opposed to local similarities which may vary across the length of the domains, as described in the homology block analysis below. The validity of the classification must be experimentally tested, but the association between domain and UPS class suggests, that at least some of the domain subclasses confer specialized cytoadhesion properties.

Identification of conserved PfEMP1 domain cassettes

Conserved domain compositional features in PfEMP1 molecules were studied in alignments of annotated PfEMP1 sequences. Alignments guided by conserved C-terminal and N-terminal domain architectures are given in Figure S4A and B, respectively. In particular the alignments were investigated to identify domain cassettes, which were defined as two or more consecutive domains belonging to particular subclasses and present in three or more of the 7 genomes (summarized in Figure 3).

The three conserved *var* genes *var1*, *var2csa* and *var3* (Figure 3, cassettes 1–3), all encoding unique DBL domains, were present in

Domain	DBLα																															
#Obs/ %ID / Avg length	365/42/420																															
Domain	DBLα0																								411	DBLα1							
Domain	DBLα0.1	DBLα0.2	DBLα0.3	DBLα0.4	DBLα0.5	DBLα0.6	DBLα0.7	DBLα0.8	DBLα0.9	DBLα0.10	DBLα0.11	DBLα0.12	DBLα0.13	DBLα0.14	DBLα0.15	DBLα0.16	DBLα0.17	DBLα0.18	DBLα0.19	DBLα0.20	DBLα0.21	DBLα0.22	DBLα0.23	DBLα0.24	DBLα2	DBLα1.1	DBLα1.2	DBLα1.3	DBLα1.4	DBLα1.5	DBLα1.6	DBLα1.7	DBLα1.8
Group																																
# Obs	40	3	11	12	22	13	7	16	24	9	11	11	8	4	15	13	8	13	6	6	6	5	5	3	13	12	13	6	12	12	11	9	6
# Genomes	7	3	5	5	7	6	3	6	7	4	4	6	4	3	6	5	3	5	5	3	3	3	3	3	6	5	6	3	6	6	6	6	3
%ID	57	71	63	58	59	60	62	55	58	64	59	58	64	63	57	60	61	55	59	60	73	58	58	62	57	71	61	79	60	63	59	61	62

Domain	DBLδ									DBLζ						DBLγ																	
#Obs/ %ID / Avg length	293/38/484									66/41/430						176/37/362																	
Domain	DBLδ1	DBLδ2	DBLδ3	DBLδ4	DBLδ5	DBLδ6	DBLδ7	DBLδ8	DBLδ9	DBLζ1	DBLζ2	DBLζ3	DBLζ4	DBLζ5	DBLζ6	DBLγ1	DBLγ2	DBLγ3	DBLγ4	DBLγ5	DBLγ6	DBLγ7	DBLγ8	DBLγ9	DBLγ10	DBLγ11	DBLγ12	DBLγ13	DBLγ14	DBLγ15	DBLγ16	DBLγ17	DBLγ18
Group																																
# Obs	##	5	5	9	13	4	4	3	3	6	10	14	12	9	15	8	14	5	8	13	13	8	9	9	13	25	10	11	6	7	7	6	4
# Genomes	7	4	5	6	6	4	3	3	3	6	5	6	6	5	6	4	5	5	4	6	6	4	7	5	6	7	5	5	4	4	5	3	4
%ID	38	59	52	48	54	61	49	60	68	75	51	46	59	56	59	70	47	51	50	54	56	61	71	41	51	46	54	48	60	68	47	54	45

Domain	DBLβ														DBLε														
#Obs/ %ID / Avg length	151/45/463									379	438	360	153/31/322																
Domain	DBLβ1	DBLβ10	DBLβ11	DBLβ12	DBLβ13	DBLβ2	DBLβ3	DBLβ4	DBLβ5	DBLβ6	DBLβ7	DBLβ8	DBLβ9	DBLpam1	DBLpam2	DBLpam3	DBLεpam4	DBLεpam5	DBLε10	DBLε1	DBLε2	DBLε3	DBLε4	DBLε5	DBLε6	DBLε7	DBLε8	DBLε9	DBLε11	DBLε12	DBLε13	DBLε14
Group																																
# Obs	10	6	7	12	4	7	31	5	30	12	12	10	5	11	11	10	10	10	10	11	12	14	13	8	14	9	8	11	8	6	5	4
# Genomes	6	3	4	6	3	3	6	4	6	4	4	5	3	7	7	7	7	7	7	7	5	7	4	6	5	5	5	4	3	4	4	3
%ID	63	53	95	56	61	60	47	52	51	52	57	49	55	78	77	87	90	84	58	70	56	59	46	76	53	57	83	58	50	48	55	48

Domain	CIDRα																												
	319/33/278																												
Domain	CIDRα1								CIDRα2													CIDRα3							
#Obs/ %ID / Avg length	56/51/251								102/38/265													129/44/302					261	254	251
Domain	CIDRα1.1	CIDRα1.2	CIDRα1.3	CIDRα1.4	CIDRα1.5	CIDRα1.6	CIDRα1.7	CIDRα1.8	CIDRα2.1	CIDRα2.10	CIDRα2.11	CIDRα2.2	CIDRα2.3	CIDRα2.4	CIDRα2.5	CIDRα2.6	CIDRα2.7	CIDRα2.8	CIDRα2.9	CIDRα3.1	CIDRα3.2	CIDRα3.3	CIDRα3.4	CIDRα3.5	CIDRα4	CIDRα5	CIDRα6			
Group																													
# Obs	9	7	5	10	6	8	8	3	16	5	5	15	12	12	8	8	8	7	6	63	36	8	19	3	9	13	10			
# Genomes	6	4	3	6	3	5	4	3	6	3	3	6	6	6	6	5	6	5	4	7	7	5	6	3	4	6	7			
%ID	71	94	98	60	61	61	67	56	48	57	52	48	55	50	59	55	53	52	55	49	50	56	53	85	47	48	49			

Domain	CIDRβ							CIDRγ															CIDRδ	
#Obs/ %ID / Avg length	204/41/256							99/37/261														214	22/60/256	
Domain	CIDRβ1	CIDRβ2	CIDRβ3	CIDRβ4	CIDRβ5	CIDRβ6	CIDRβ7	CIDRγ1	CIDRγ2	CIDRγ3	CIDRγ4	CIDRγ5	CIDRγ6	CIDRγ7	CIDRγ8	CIDRγ9	CIDRγ10	CIDRγ11	CIDRγ12	CIDRpam	CIDRδ1	CIDRδ2		
Group																								
# Obs	##	12	11	15	19	15	3	12	13	8	11	11	9	8	7	6	5	5	4	11	14	8		
# Genomes	7	5	6	7	6	7	3	5	6	5	6	6	6	6	6	4	4	3	3	7	6	5		
%ID	44	54	55	47	51	51	56	50	55	46	57	57	53	57	57	60	58	66	59	75	63	66		

Figure 2. DBL and CIDR domain class characteristics. Number of observations (#obs) of CIDR and DBL domain classes in 399 PfEMP1 (Figure S5), number of genomes represented in the classes (#genomes) (of the seven genomes 3D7, HB3, DD2, IT4, PFCLIN, RAJ116 and IGH), and the average shared sequence identity of major and minor subclasses (%ID). A color was added under domain classes where at least 25% of the observed domains were found in UPSA (green), UPSB (red), UPSC (yellow) or UPSE (pink).
doi:10.1371/journal.pcbi.1000933.g002

all seven *P. falciparum* genomes, except *var3* which was not present in HB3 and IGH. As previously reported, fragments of *var1* and *var2csa*, but not *var3*, were found in *P. reichenowi* [57]. Some variation in domain composition was observed within the three conserved gene families. Thus, in RAJ116, *var2csa* encoded an extra C-terminal DBLε domain, and in DD2, the *var1* gene encoded C-terminal domains different from the other *var1* genes. *Var1* of 3D7 and IT4 appeared to lack an exon2 sequence, whereas five *var1* genes had a premature stop codon at similar positions in their exon2. Domain pairs characteristic for *var1* (DBLγ1/15-DBLε1 and DBLζ1/2-DBLε) were found in other group A *var* genes (IT4var9, IGHvar32, DD2var23 and HB3var06). Taken together, this indicates that *var1* often is found as truncated gene, and that the particular functional properties of VAR1 may have moved to other PfEMP1 variants. Similarly, a VAR3 sequence corresponding to 80% of the exon1 as well as

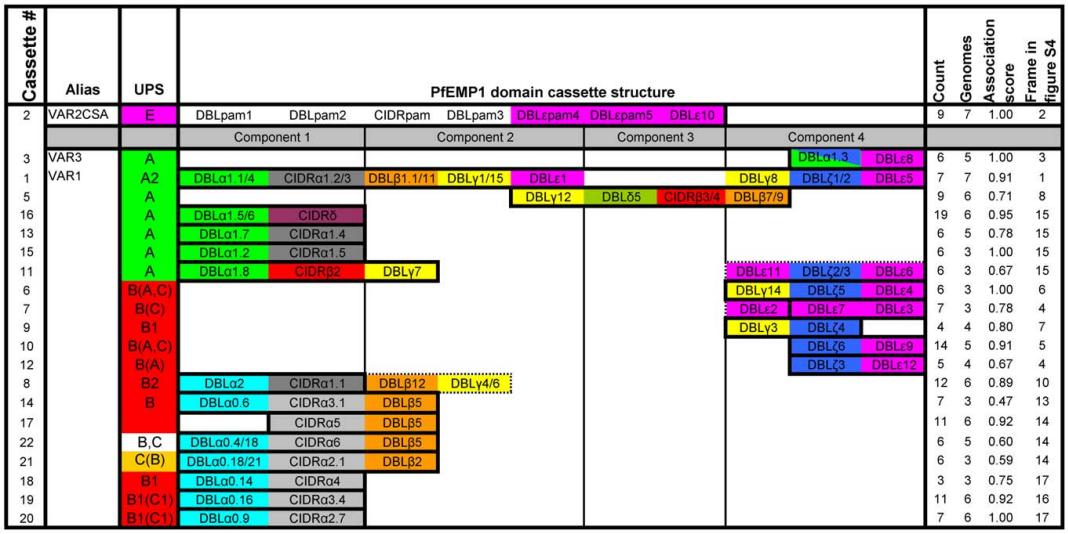


Figure 3. Overview of distinct PfEMP1 domain cassettes. A PfEMP1 domain cassette was defined as a *var* gene sequence encoding two or more DBL or CIDR domains with subclasses that could be predicted from each other. In a few cases domain cassettes (filled frames) could be expanded with additional domains but in limited number of genes or genomes (punctured frames). A cassette was given an association score calculated as the average of all domain pair associations of a domain cassette. Each domain pair association (A–B) was calculated by dividing the number of times the domain combination was observed in the dataset by the least number of times either A or B was found in the dataset. The association score does not include the UPS association. Associated UPS classes are colored according to the UPS class most often observed flanking the cassette. Less frequent flanking UPS classes are in brackets. The number of times a given domain cassette was observed (count) and the number of genomes in which it is present (genomes) within the seven genomes, 3D7, HB3, DD2, IT4, IGH, RAJ116 and PFCLIN are given. The frame number in Figure S4, detailing the genetic context of the domain cassette is also given.
doi:10.1371/journal.pcbi.1000933.g003

exon2 was found in the 3′ end of RAJ116var03, consistent with how DBLζ and DBLε domains are positioned in other PfEMP1. The domain composition variation within the three most conserved *var* genes highlight the importance of ectopic recombination of large single or multi domain elements for the generation of PfEMP1 diversity.

Among the novel domain composition phenomena, domain cassette 5 (Figure 3) was the most prominent. This four domain C-terminal cassette was found exclusively in ten group A PfEMP1, and in six of the seven *P. falciparum* genomes as well as in *P. reichenowi*.

Interestingly, nearly all DBLζ and DBLε domains were found in C-terminal domain cassettes (domain cassettes 1,3,6,7 and 9–12) and often occurred in genes encoding CIDRγ1/2/9 domains (approx. three of four CIDRγ1/2/9 domains flank DBLζ and DBLε domains). The unambiguous partition of DBLε subclasses and the positional and compositional similarities between different DBLε, could suggest that specialized functions reside in these structures.

In the PfEMP1 N-terminal, DBLα subclasses correlated well with the subclasses of their neighboring CIDR domain (Figure S4B). As expected, all group A PfEMP1 except VAR3 exclusively contained the domains DBLα1-CIDRα1/β2/δ/γ3, but furthermore, group A PfEMP1 appeared to be divided into those harboring either DBLα1.5/6/8-CIDRβ2/γ3/δ (includes cassettes 11 and 16 in Figure 3; Figure S4, frame 15) or DBLα1-CIDRα1. Within group B and C PfEMP1 two major groups were observed, those encoding DBLα0 domains associated with CIDRα2, and those encoding DBLα0 domains associated with CIDRα3. In addition, eight distinct CIDRα containing cassettes were found, including domain cassette 8 which is particularly noteworthy, as it is associated with UPSB2 (7 of 12 domain cassette 8 encoding genes are flanked by 7 of 11 UPSB2) and contains DBLα2, which formed a separate cluster from DBLα0 and α1 in the DBLα tree. Domain cassette 8 may be expanded further in a less well defined form with two domains (DBLβ12-DBLγ4 or DBLγ6) (Figure S4A, frame 10).

Several more elusive domain architectural constraints were observed, which may crystallize into domain cassettes if higher sequence depth is acquired. These included the group A specific domain combinations DBLα1.4-CIDRα1.6/7-DBLβ3, which both could represent the core of what have been proposed as VAR4 (represented by PFD1235w; Figure S4A, frame 9) as well as DBLβ7-DBLγ-DBLγ (Figure S4A, frame 9).

The present description of PfEMP1 diversity was based on analysis of seven near complete genome sequences: four Asian, two African [58], and one Central American isolate. None of the described domain architectural constraints were found exclusively in the African or Asian isolates, which strongly imply that there is no basic difference between the PfEMP1 repertoires of *P. falciparum* around the world. However, more *P. falciparum* genome sequences are desirable to gain a better resolution of conserved domain cassettes.

In general there were no correlation between occurrences of N-terminal and C-terminal domain cassettes, and whereas group A PfEMP1 shared no N-terminal domain cassettes with group B or C PfEMP1, C-terminal domain cassettes were more often shared among PfEMP1 groups. The three conserved *var* genes have already attracted warranted attention, but while the binding specificity of VAR2CSA and its relevance in pregnancy malaria is well established, no function or clinical importance has been assigned to VAR1 and VAR3. Several studies have aimed to

define the PfEMP1 molecular background for severe malaria in children. Most *ex vivo* studies [27,29–31] have relied on relating phenotypic or clinical data to the phylogeny of partial DBLα tags amplified from parasite cDNA, or direct quantitative PCR measurements of group A, B and C *var* genes. Although these approaches target some of the best conserved PfEMP1 phenomena, both methods disregard the structures unlinked to the PfEMP1 N-terminal, and fail to reflect some of the most evident of the conserved N-terminal domain cassettes. Nevertheless, the consensus drawn from these studies and *in vitro* studies of model parasite lines [28] emphasize the importance of group A PfEMP1 in severe malaria, and interestingly, often the particularly distinct group A domain cassette 5 [9,28,34].

Although several of the domain classes and PfEMP1 structural constraints presented here are vaguely defined and by themselves difficult to rank according to clinical relevance, the PfEMP1 diversity described by groups, components, domain classes and cassettes offers an operational model for design and interpretations of future experimental studies.

PfEMP1 homology blocks

DBL domains consist of hyper-variable and conserved regions, as previously described [2,36,59], and in a comparison of DBL similarity, Smith *et al.* (2000) were able to define a set of ten homology blocks with an average length of 21 amino acids, conserved in all DBL domain classes [36]. To describe in detail these frequent shifts in conservation level across PfEMP1, an iterative method was developed that automatically defines a set of homology blocks in a set of unaligned protein sequences. The method is especially appropriate for the frequently recombining *var* genes, as the short homology blocks are less inclined to group unrelated sequences which may be forced together in longer domain alignments.

The term homology block (HB) refers to a sequence profile defined from a multiple sequence alignment, here described by a hidden Markov model (HMM) [60]. Sequences with similarity above a threshold to the sequence profile are termed members, hits or occurrences of the homology block, and the members of a homology block can be defined in a sequence by searching with the HMM.

Starting from a full sequence database, homology blocks were one after one first defined and then excluded from the database. Each homology block was defined to be the sequence profile with the highest number of occurrences in the database, i.e. the most conserved sequence, with boundaries optimized to match this criterion. Sequence similarity was assessed with HMM log-odds scores, and a significance threshold of $S \geq 9.97$ bits was used for all homology blocks, to ensure that each member of a homology block was at least 1000 times more likely to be related to the sequence profile, than to a random sequence with amino acid frequencies as in the database. Thus, a set of homology blocks was defined, where each homology block comprises all related sequence stretches in the database. The method is described in detail in Text S2.

The analysis was performed on a database with 311 PfEMP1 sequences containing information on the entire molecule or a full exon1. Twenty DBL containing paralogs were also included to enable estimates of evolutionary relationships. The minimal length of the homology blocks was set to seven amino acids, as this is approximately the length required to reach the sequence similarity significance threshold. Sequences with less than five homologs in the database were not included in the homology block set, since PfEMP1 from more than seven *P. falciparum* genomes were in the dataset, and the main interest was to determine sequence features conserved in most of these genomes.

Characteristics for the resulting 628 homology blocks are shown in Figure 4. On average 83.5% of a PfEMP1 sequence was

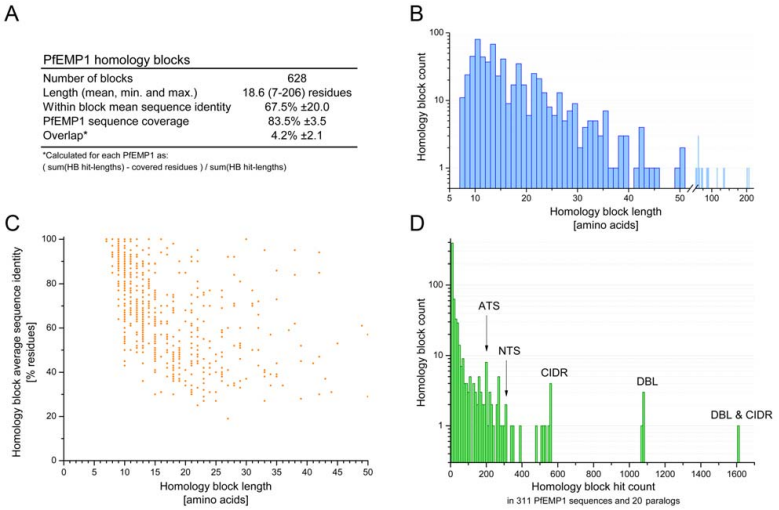


Figure 4. Characteristics for 628 PfEMP1 homology blocks. (A) Length corresponds to the alignment length of the multiple sequence alignment defining the HB. Sequence identity in the table is given as mean and SD for the distribution of all homology block avg. pairwise identities. HB coverage and overlap were calculated per PfEMP1 and mean and SD are given for these distributions. (B) Length distribution for HBs. The most frequent length was 10 residues. (C) Scatter plot showing avg. pairwise sequence identity for HBs of differing length. (D) Histogram showing number of HBs with same prevalences in the database. The bin size of the histogram is 10 hits. One HB was found with a prevalence of 1605 hits in the PfEMP1 database, representing a HB present in nearly all DBL and CIDR domains. Similarly, a number of homology blocks were found specifically in each of the domains DBL, CIDR, NTS and ATS. Most homology blocks had between 5 and 15 hits.
doi:10.1371/journal.pcbi.1000933.g004

described by homology blocks, and the remaining fragments were either shorter than seven residues, or had fewer than five homologous sequences. Overlap between homology blocks were mainly concentrated in areas with low complexity sequence, such as the inter-domain regions, and amounted to an average of 4.2% of HB occurrences in a PfEMP1 sequence (Figure 4A). The most frequent HB length was 10 residues, while the average was 19 residues (Figure 4B). HB average sequence identity was between 19–100%, and as might be expected for the shortest sequences, only similarities with high identity could be detected within the significance threshold (Figure 4C). The analyzed PfEMP1 sequences contained 311 NTS, 199 ATS, 1043 DBL and 552 CIDR domains, while the paralogs contained 30 DBL domains. One homology block occurred 1605 times in the database and was found in all DBL and CIDR domains, except 20 (not present in DBLpam2, DBLε7 and DBLε12), while four other blocks were found in all DBL domains and six blocks were strongly correlated with CIDR (Figure 4D). The homology blocks were numbered according to their frequency in the database, with the most frequent being HB number one.

88 PfEMP1 were not in the HB definition sequence set, and when the 628 defined homology blocks were predicted in these proteins, 82.5% (SD ±4.9%) of each PfEMP1 were on average covered by HBs, similar to the coverage in the definition sequences (Figure 4A), showing that the homology blocks describe universal PfEMP1 sequence features.

The VarDom server was developed to provide an interactive graphical interface to analyze information on domain classes, homology blocks and their distribution in PfEMP1 sequences. Alignments and other related files can be downloaded, and it is possible to submit new sequences to annotate them with domains and homology blocks, to classify them and relate them to other sequence groups in the seven genomes: http://www.cbs.dtu.dk/services/VarDom/ In the following, the HB distribution in PfEMP1 is presented, and several references are made to specific homology blocks. These blocks as well as the sequences they occur in can be inspected using the VarDom server.

Homology blocks describe the conserved core of DBL and CIDR domains

The five most prevalent homology blocks in PfEMP1 (HB1–5) were present in nearly all DBL domains. The relative positions of these five HBs in DBL domains were conserved (Figure 5A), and within the HBs several amino acid positions were strongly conserved in all DBL domains. Figure 5B shows occurrences of HB1–5 in DBL1 (a.k.a. F1) of the paralog PfEBA-175 and DBLpam3 (previously DBL3X) of VAR2CSA. The DBL structure consists of subdomain 1 (S1) with mixed helix-sheet structure, and two helix bundles (S2 and S3) [37,49]. Disulfide bonds between conserved Cys residues mainly serve to hold together each individual subdomain, demanding other types of interactions to hold a stable domain structure [37,50,61–62]. HB1, which was also found in CIDR domains, described a complete α-helix with one side conserved, giving a pattern of conserved residues spaced by 3 residues for each helix-turn (Figure 5A). The conserved side of HB1 faced HB2, which was found to be the most conserved sequence in DBL domains, with a mean amino acid sequence identity of 56%. HB2 was part of a longer helical structure and interfacing with HB1, HB3 from the other helix bundle, and HB4 which formed the non-surface exposed part of S1 (Figure 5B and C). All these interactions probably constitute the main selection pressure, keeping HB2 relatively conserved. HB3 in S2 corresponded to HB2 in S3, with interactions to HB2, HB5 and HB4, and with mean sequence identity of 47% it was found to be the second most conserved part of DBL domains. HB5 was mainly conserved on one side of the helix like HB1, suggesting for both that they may be frequently exposed on the surface of PfEMP1.

Side chains in conserved amino acid positions were mainly directed towards other conserved parts, although some were pointing outwards probably to interact with other less conserved domain parts (Figure 5B and C). Functions for some of the conserved amino acids in HB1–5 were identical in both structures (Figure 5A and D), where they formed polar and hydrophobic interactions between the three subdomains. Besides from the conserved polar interactions shown, the conserved Pro on position 4 in HB4, which introduced a kink in the β-sheet structure of S1, was in a position allowing it to interact hydrophobically with the also conserved Trp on position 8 in HB2. It may thus contribute to hold the β-sheet in place. In general the conserved positions of HB1–5 described a set of residues, which in the known DBL domain structures interact to hold together the three DBL subdomains, so they can be said to constitute the conserved core structures and interactions of DBL domains.

HB1–5 were found among the 10 homology blocks defined by Smith *et al.* (2000) [36], where HB4 = HBb, HB3 = HBd, HB5 = HBf, HB2 = HBh and HB1 = HBj. The remaining homology blocks, defined in that paper, were not found to be conserved in all DBL classes, based on the chosen similarity significance threshold.

Homology blocks specific for all CIDR domains were also found, and they were present in the most conserved part of CIDR, the designated minimal CD36 binding region or M2 [36,45], for which the structure is known [48] (Figure 5E and F). HB8, HB1 and HB10 were found to correspond to helix 1, 2 and 3 respectively in the three-helix bundle, and the similarity of this bundle to subdomain 3 of DBL was confirmed by the presence of HB1 in all CIDR and DBL domains. The conservation of these three helices suggests that this structure is common to all CIDR domains. Interestingly, four HBs (HB12, 7, 9 and 6) situated in subdomain S3 of DBLα and DBLδ domains, were exclusively found flanking all CIDR domains, strongly supporting the link between CIDR and DBL domains.

Side chains of conserved residues in HB1, 8 and 10 were mainly directed towards the center of the CIDR three-helix bundle (Figure 5F), where they interacted to keep the structure together. Some parts of the structure have not been solved, including the C-terminal end of HB8 with several conserved basic residues and a Cys likely to form a disulfide bridge to position 1 in HB1. A few conserved residues in HB8 were directed away from the helix bundle core. Among these were the basic position 24 and possibly also 28 as the distance fits with a helix turn. These residues may thus be involved in interactions with surrounding parts of the PfEMP1 such as the helix-loop of CIDR, or even substrate binding, and they may be target for the cross-reactive antibodies inhibiting CD36 binding described by Mo *et al.* (2008) [63].

Alignment of DBL homology blocks

Just as PfEMP1 can be represented as strings of amino acid symbols or strings of domain names, they can be represented at an intermediate level as a string of homology blocks. To study similarities between DBL domains, the homology block sequences of 1043 DBL domains, consisting of 378 different HBs, were studied (Figure 6). Occurrences of the same homology block were vertically aligned (Figure 6, center), and rows in the alignment were sorted according to a NJ-tree (Figure 6, left) built based on differences in HB composition of the DBL sequences. The five core homology blocks divides DBL domains into six regions, and

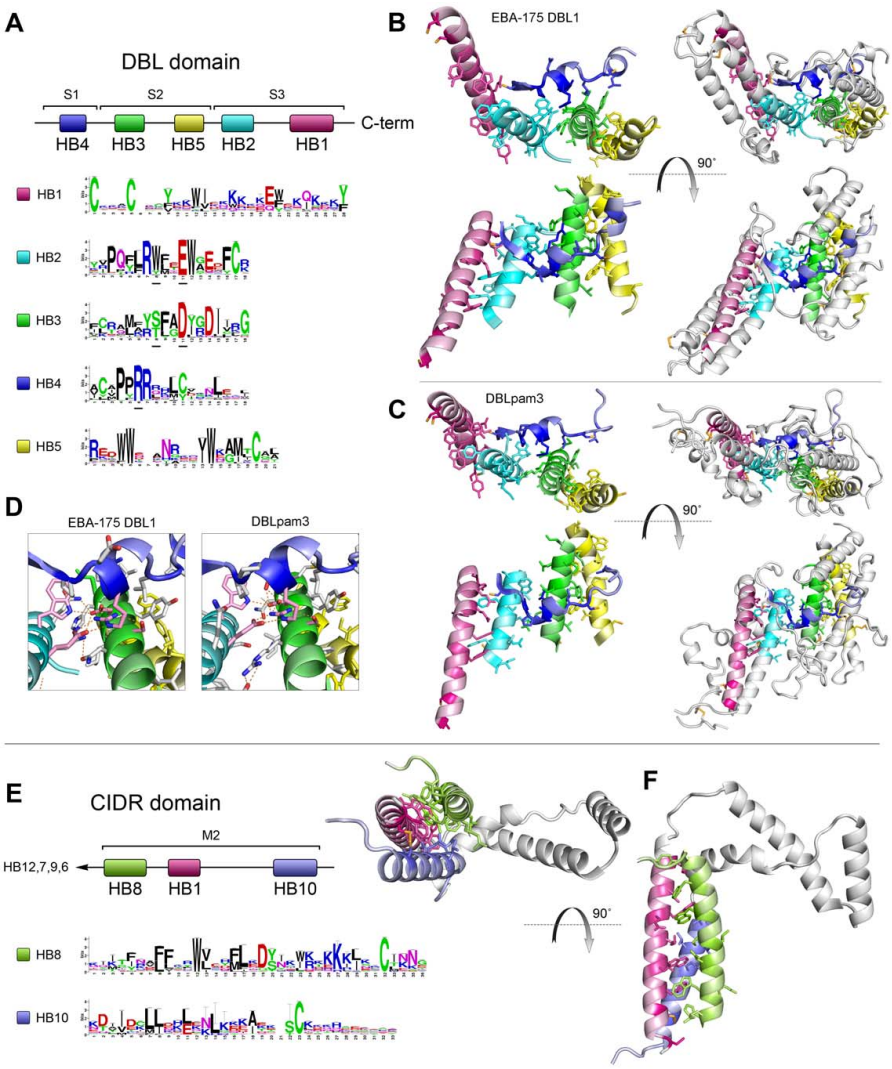


Figure 5. Conserved domain cores. (**A–D**) Five most conserved PfEMP1 homology blocks form DBL-core structure. (A) Schematic showing relative positions in DBL domains of HB one to five (S1–3 indicate subdomains) and sequence conservation logos for each homology block alignment. The height of each position in the logos indicate the amino acid conservation level, and the height of the individual amino acids reflect their relative frequencies on the position and thus their contribution to the conservation. A small sample bias correction has been subtracted in the logos, on alignment positions containing few (<40) amino acids, and error bar height is 2× the correction. Polar amino acids are green, neutrally charged are purple, basic are blue, acidic are red, and hydrophobic amino acids are black. HB numbering is based on level of conservation in PfEMP1 and related sequences. (B) HBs shown on PfEBA-175 DBL1 structure, and (C) on VAR2CSA DBLpam3 structure. Side chains are shown for conserved positions with conservation level higher than 50% of maximum, corresponding to 2.16 bits. DBL areas which are not part of HB1–5 are shown as lightgray in rightmost column, while left side shows only HB1–5, color coding as in panel A. Coloring intensity in the structure is proportional to conservation level in the HBs. (D) Polar interactions between conserved positions in EBA-175 and DBLpam3. The conserved pink residues are underlined in Figure 5A. (**E–F**) Conserved sequence blocks in CIDR domains. Relative homology block positions, and sequence logos (E). HB12, 7, 9 and 6 are all strongly correlated with CIDR domains. (F) HBs shown on the structure of the M2 part of MC179 CIDRα domain. Disulfide bridges are shown in orange.
doi:10.1371/journal.pcbi.1000933.g005

sequence conservation logos are shown for representative homology blocks in each region (Figure 6, top and bottom).

Many of the domain classes derived from trees based on amino acid alignments (Figure 2 and Figure S2), were also found by the tree based purely on the absence or presence of homology blocks (Figure 6), and these groups can thus be described by a specific homology block combination. Most major classes formed monophyletic groups, with the exception of DBLγ and DBLε,

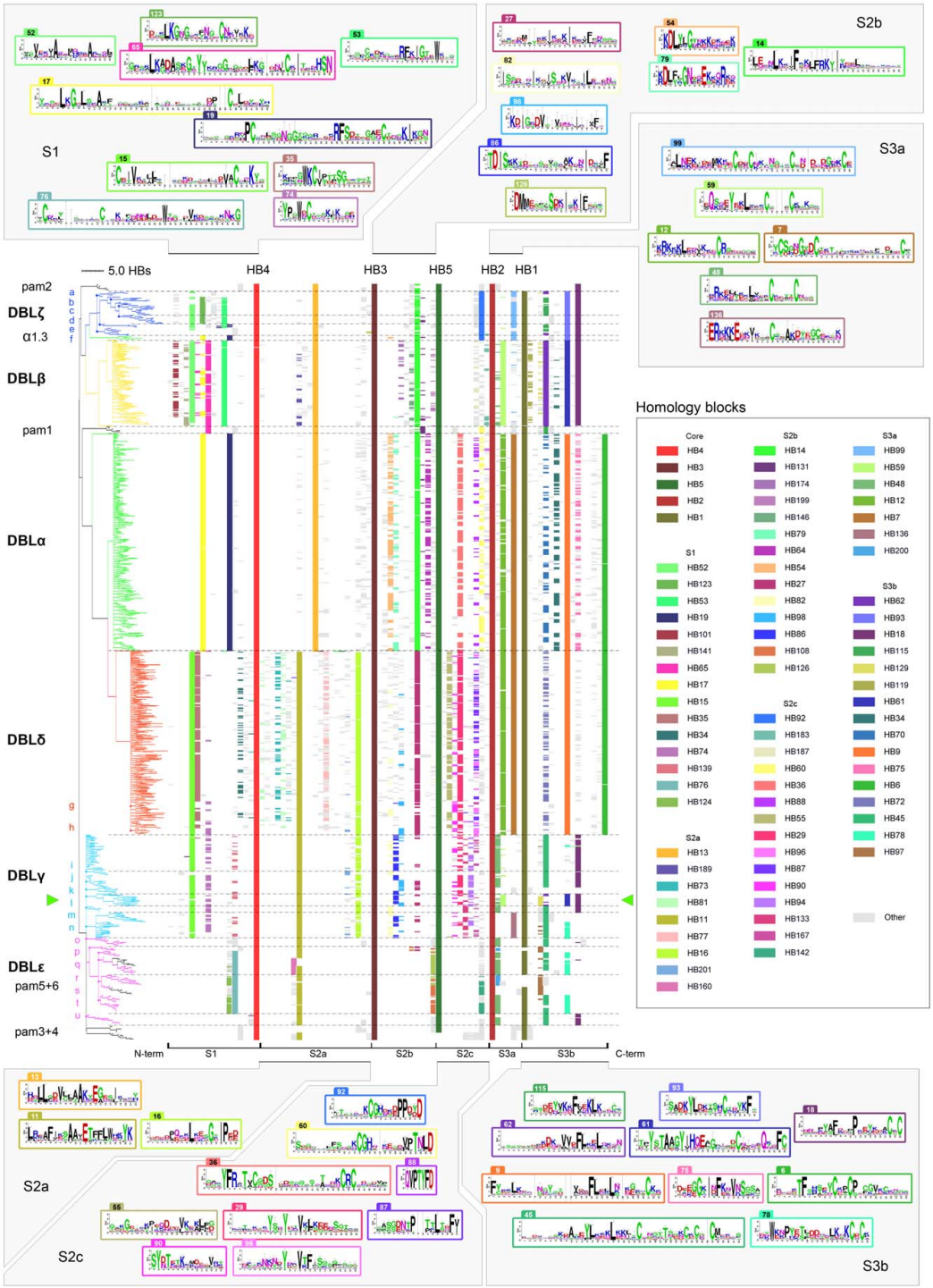

S1
S2b
S3a
5.0 HBs
HB4
HB3
HB5
HB2
HB1
pam2
DBLζ
α1.3
DBLβ
pam1
DBLα
DBLδ
DBLγ
DBLε
pam5+6
pam3+4
N-term
S1
S2a
S2b
S2c
S3a
S3b
C-term
Homology blocks
Core
HB4
HB3
HB5
HB2
HB1
S1
HB52
HB123
HB53
HB19
HB101
HB141
HB65
HB17
HB15
HB35
HB34
HB74
HB139
HB76
HB124
S2a
HB13
HB189
HB73
HB81
HB11
HB77
HB16
HB201
HB160
S2b
HB14
HB131
HB174
HB199
HB146
HB79
HB64
HB54
HB27
HB82
HB98
HB86
HB108
HB126
S2c
HB92
HB183
HB187
HB60
HB36
HB88
HB55
HB29
HB96
HB87
HB90
HB94
HB133
HB167
HB142
S3a
HB99
HB59
HB48
HB12
HB7
HB136
HB200
S3b
HB62
HB93
HB18
HB115
HB129
HB119
HB61
HB34
HB70
HB9
HB75
HB6
HB72
HB45
HB78
HB97
Other

Figure 6. DBL homology block alignment. HBs in 1043 DBL sequences aligned, and sorted by NJ-clustering based on differences in HB composition. Tree distances show the number of different HBs in the DBL domains. The sequences are divided into 6 segments by the conserved core HB1–5 (Figure 5), and the corresponding subdomain parts are noted below the alignment. Only the 80 most frequent of 378 HBs are colored. Sequence conservation logos as described in Figure 5 are shown for selected HBs, where number tabs indicate the HB number. Logos are when possible placed in order of appearance in the alignment. Letters next to the tree identifies groups marked by dots in the tree, matching domain subclassification based on amino acid alignments: (**a**) ζ3, (**b**) ζ5, (**c**) ζ6, (**d**) ζ4, (**e**) ζ1, (**f**) ζ2, (**g**) δ5, (**h**) δ4/8/9, (**i**) γ7, (**j**) γ11/15, (**k**) γ1, (**l**) γ2/9, (**m**) γ8, (**n**) γ5/6/12/16/17, (**o**) ε2, (**p**) ε7, (**q**) ε4, (**r**) pam6/ε3, (**s**) pam5/ε5/ε12, (**t**) ε6/9, (**u**) ε1/11/13. The green pointers mark products of recombination between DBLγ and DBLβ domains, with break point around HB2. Additional information for all HBs can be found by querying the VarDom server with the HB numbers, as given in the legend or on the logos. Labeled homology block alignments can be found in Figure S7.
doi:10.1371/journal.pcbi.1000933.g006

which formed one big cluster with several well-defined subgroups. Minor subgroups were mainly found in DBLζ, γ and ε (Figure 6, tree group a–u), and many correlated well with domain classes based on amino acid alignments. Most subgroups of DBLα, β, and δ were too subtle to be distinguished. The DBLα0-DBLα1 division was not clearly found, although HB36 may approximately describe the difference, by being present in 205 of 230 DBLα0 domains, and in none of the 61 DBLα1. HB36 was absent in all cys2 sequences but present in all cys4 sequences, thus describing the division between group 1–3 and 4–6 in the DBLα sequence tag classification [64].

Domain subclasses (Figure 2) could often be described by subclass specific homology blocks. For instance DBLζ4 was described by HB283 and HB284. Other subclasses were characterized by HBs shared exclusively with other major domain classes, examples being DBLζ1, which shared HB19 with DBLα (Figure 6e S1, blue), and DBLγ2/9 domains, which were characterized by having a DBLβ S3 subdomain (Figure 6 S3, green pointers). Similarly, the S3 subdomain of VAR1 DBLε1 was very similar to the one present in a number of DBLγ sequences (Figure 6 S3, tree group u). Cassettes could also be identified, exemplified by HB331, which occurred exclusively in the N-terminal of DBLβ domains in domain cassette 5 (Figure 3).

DBLα1.3 of VAR3 contained HB17 and HB19 which were characteristic for DBLα domains (Figure 6 S1), but S2c and S3 in DBLα1.3 were very characteristic for DBLζ, sharing several DBLζ specific homology blocks: HB92, HB99, HB592, HB93, and HB18. Thus, homology block analysis of VAR3 suggests that DBLα1.3 is a DBLα-ζ hybrid, and it will be interesting to see if the function of this domain is similar to any of the two combined classes alone. The finding of DBLζ elements in VAR3 associates this PfEMP1 with the domain combination DBLζ-DBLε, often found in component 4 cassettes (Figure 3, Component 4), which could imply functional analogies between VAR3 and these cassettes.

DBLpam1 and 2 shared homology blocks with DBLα/β/ζ, while DBLpam4 and to a high degree DBLpam5 and 6 shared blocks with DBLγ/δ/ε (Figure 6). Interestingly, DBLpam1 contained HB65 (Figure 6 S1, pink), a sequence that was mainly found in DBLβ. However, in the C-terminal end DBLpam1 shared HB60 with DBLα (Figure 6 S2c, yellow) and HB115 with DBLζ1/5/6 (Figure 6 S3b, green, tree group b, c and e). Thus, DBLpam1 appeared to contain elements from all of DBLα, β and ζ. The shared homology blocks, as well as the fact that the hybrid domains DBLα1.3 and DBLpam1 appears to be functional, suggests a more recent common ancestry and possibly related functions of DBLα, β, ζ, pam1 and pam2 domains.

Similarities between major DBL classes also varied considerably across the length of the domains (Figure 6), and a major homology break point, where similarities differed on each side, was observed for many sequences around HB2, the most conserved DBL homology block.

In the N-terminal, a clear division was found between DBLα/β/ζ and DBLγ/δ/ε, best defined by HB11 and HB13, respectively (Figure 6, S2a). At this end of DBL domains, only the core homology blocks HB1–5 occurred in both groups, indicating low levels of recombination between these groups, and possibly different functions. Within these groups, DBLζ had high similarity to DBLβ, most significantly in the S1 subdomain, and DBLδ was very reminiscent of the DBLγ in the N-terminal, some sequences were even identical on the homology block level (Figure 6g, h, j, and k).

The C-terminal of DBL domains could also be divided into two major groups, consisting of the S3 subdomains of DBLα/δ and DBLζ/β/γ/ε, respectively (Figure 6, S3). DBLα and δ shared four homology blocks connecting to the downstream CIDR domains. S3 homology blocks in DBLζ and β were uniform and specific to each class, whereas DBLγ and ε S3 were more diverse (Figure 6, S3).

N- and C-terminal ends of several major DBL domain classes thus appear to have different sequence similarities, most likely reflecting that the sequences have been joined through recombination, often with a break point around HB2, and they therefore have different evolutionary histories. Phylogenetic classification based on whole domain sequence alignments will tend to be an average of such different histories.

Evolutionary relationships among DBL subdomain sequences suggest intra-DBL recombination break point. Identification of adjacent genetic regions with different evolutionary histories is a widely used method for detecting recombination break points in distantly related sequences [65–66]. To get a complete picture of evolutionary relations among subdomain sequences, with the aim to determine if recombination has occurred with break point between S2 and S3, phylogenetic trees based on amino acid alignments were built for the three DBL subdomains (Figure 7). Trees in Figure 7 are included as Figure S6 with labels and bootstrap values.

Relations among sequences of the S3 DBL subdomain clearly differed from those of S1 and S2 (Figure 7). DBLα and DBLδ S3 subdomains were found to be closely related, separated from the remaining sequences in all 1000 bootstraps, whereas in S2, DBLα was most closely related to DBLβ and ζ, supported by 99% of the bootstraps. Similarly, DBLγ and DBLε S3 subdomains were closely related, while S2 sequences of DBLγ were closely related to DBLδ, separated from DBLε by several highly supported branches. This strongly indicates that the evolutionary histories for S2 and S3 subdomains are different, as also suggested by the homology block analysis (Figure 6), and that recombination most likely has occurred with break point between these subdomains.

In agreement with the homology block analysis, the division between DBLα/β/ζ and DBLγ/δ/ε was well supported by bootstrap values in both S1 and S2, as was the separation of each of the domain classes DBLα, β, and ζ (Figure 7, S1 and S2). For S1 and S2, DBLδ and DBLγ sequences were clustered together with low bootstrap support for separation within this group, although a specific set of DBLδ sequences had particularly close relations to DBLγ (Figure 7, S1 and S2). The relationship was most pronounced in the S2 subdomain, where 46 DBLδ sequences represented in all seven genomes, and including all non-

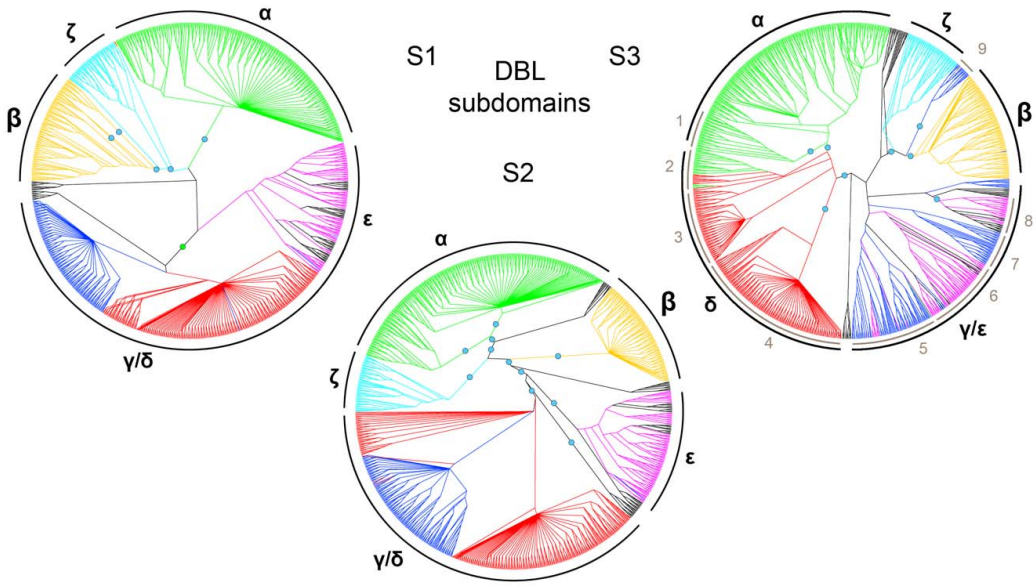


Figure 7. Evolutionary relatedness of DBL subdomain sequences. A cladogram is shown for each of the three DBL subdomains S1–3, where boundaries for the subdomains were chosen at the edges of HB4 and HB2, as shown in Figure 6. Colors indicate major DBL domain classes estimated from alignment of the whole domains: Green: DBLα; Orange: DBLβ; Blue: DBLγ; Red: DBLδ; Magenta: DBLε; Cyan: DBLζ. VAR2CSA sequences are black. Blue dots indicate major bipartitions supported by at least 50% of 1000 bootstraps. The green dot in S1 marks a bipartition with bootsrap value 0.39. Subdomain clade correlation with whole domain classes is indicated around the trees in black; Clades were split if supported by 50% of the bootstraps.
doi:10.1371/journal.pcbi.1000933.g007

DBLδ1 subclasses, were found closer to the DBLγ clade. The 3D7 genes containing these DBLδ sequences were MAL6P1.4, PF11_0521, PF13_0003 and PF11_0008. The latter *var* gene has been found to be the target for protective antibodies [9,34], and together with PF13_0003 contains cassette 5 (Figure 3).

DBLα1 S3 sequences flanked by CIDRα1 domains were well supported as a subgroup (Figure 7 S3-1). Interestingly, all those DBLα domains that were not followed by CIDRα (DBLα1.5/6/8 domains), had an S3 subdomain which clustered with DBLδ S3 sequences (Figure 7 S3-2), indicating recombination between DBLα and DBLδ. Similarly, DBLδ clusters were found for DBLδ domains followed by CIDRγ (Figure 7 S3-3), and CIDRβ (Figure 7 S3-4). These associations between S3 and CIDR indicate that the recombination break point occurs within the DBL domain when CIDR domains are exchanged, and further supports a functional dependency between CIDR and their upstream DBL domains.

DBLγ and DBLε S3 subdomains were found mixed in one cluster with low bootstrap support (Figure 7 S3-5, 6, 7, 8), although the subgroups were to some degree specific for either DBLγ or DBLε. One DBLγ clade was composed of S3 subdomains of DBLγ5/6/12/16/17 (Figure 7 S3-7), captured by HB136 (Figure 6n) and found in a set of 36 PfEMP1 nearly void of DBLε and ζ domains. Two small DBLγ subgroups, DBLγ1/15 of VAR1, and a group comprising DBLγ2/9 domains, were found separately, and the latter group was closely related to DBLβ S3 sequences (Figure 7-S3-9), as expected from the homology block alignment (Figure 6k and l). These DBLγ-β hybrid domains appeared in 16 PfEMP1, found in 6 of 7 genomes (not HB3), the 3D7 gene being PF07_0050.

DBLε S3 sequences were dichotomized with a bootstrap support of 80%. One clade contained all DBLpam5, two DBLpam6, as well as DBLε5/7/12 (Figure 7 S3-8). The S3-8 cluster was characterized well by HB97 (Figure 6p and s), which was also present in several paralogs, such as PFA0665w DBL2 and PFD1155w DBL2, indicating that HB97 describes an ancient conserved domain element, a notion supported by its presence in the conserved genes *var1* and *var2csa*. The presence of HB97 in paralogs and many DBLε domains, suggests that of all PfEMP1 domain classes, DBLε may bear the highest resemblance to a common ancestral DBL domain.

The subdomain sequence comparison thus corroborates observations on homology block and domain level. The relations found between S3 subdomain sequences differ markedly from relations between S1 and S2 sequences, which supports the theory of a recombination hotspot between subdomain S2 and S3. The homology block analysis further suggests that the break point often occurs around HB2.

The subdomains S1 and S2 of DBLγ and DBLδ domains appear to be closely related, whereas the S3 subdomain sequences are distantly related, indicating recombination with break point around HB2. Furthermore, HB2 recombination products have been identified with 5′ DBLγ and 3′ DBLβ/ε sequences, as well as with 5′ DBLα and 3′ DBLδ sequences.

The area around HB2 is a hotspot in the sense that recombination has occurred at this position more frequently than at other sites during the history of the *var* genes. It is however difficult to say if this area has an especially elevated recombination frequency, or if the high number of observed recombination events is purely due to functional selection, i.e. there has been recombination all over the gene, but mainly recombinants with break points near HB2 have been retained due to better functionality. Recombination between DBLβ and DBLγ appears

to be rare, judging from the fact that DBLγ-β hybrid domains are represented in 6 of 7 genomes (Figure 7-S3-9, Figure S6), and that these sequences form a cluster in the HB61 tree. This is suggestive of a common ancestral sequence dating back before geographic separation of the genomes. Recombination between DBLα and DBLδ with break point in the HB2 area, resulting in S3 and CIDR domain exchange, may be a more frequent event, judging from the fact that all four combinations of DBLα/δ-CIDRβ/γ occur, which are likely to be the product of at least two recombination events. Corroborating this, S3 subdomains followed by CIDR1β and CIDR2β clustered together, separate from a cluster of S3 sequences followed by CIDR1γ and CIDR2γ (Figure 7 S3-2). These sequence relations were also found in the phylogeny for HB7, indicating that the break point of these recombination events occurred upstream of HB7, and thus near HB2.

Frequent recombination around HB2 could suggest independent functions for S1+S2 and S3, as proposed for VAR2CSA domains where S3 generally was found to be less surface-exposed [50]. This may be particularly true for DBLγ/δ S1+S2 sequences, as they apparently can be combined successfully with very diverse downstream sequences, including DBLβ S3 subdomains and CIDR domains.

Recombination is also likely to occur between more closely related domains, e.g. within a domain class. This will probably occur more frequently due to higher sequence similarity, but will result in more subtle changes. DNA must be analyzed to detect such subtle changes optimally, and this could be done by studying the phylogenetic trees built for each homology block. This comprehensive task is however not within the scope of the current study. A recombination analysis has previously been performed on sequences encoding DBLpam3 domains [59], and interestingly the most significant recombination hotspot in this DBL class was also found near HB2.

Potential integrin binding of DBLα0 domains. Integrins are a family of cell surface membrane receptors, mediating binding to the extracellular matrix, as well as interacting with plasma proteins and counter receptors on other cells, thereby involving them in basic processes such as cell adhesion, cell migration and cell-cell communication. Integrins are heterodimers composed of two membrane anchored subunits, α and β of which the human genome encodes 18 and 8 variants respectively, combining into 24 known, human receptors [67]. Integrin subunit homologs are found in both complex and simple metazoan organisms including sponges and corals [68], and the wide distribution, both in species and across tissue types, makes the receptors an attractive target for pathogens, such as various bacteria, viruses, fungi, and parasites, which use these receptors for adhesion or internalization in the host [69–72]. Disintegrin domains in snake venom toxins, as well as ornatin from leech toxins, bind integrins to inhibit their function in platelet aggregation [73]. It has previously been shown that IE adhesion to human dermal microvascular endothelial cells (HDMEC) can be inhibited by anti-α_v antibodies (i.e. antibodies targeting the v variant of integrin α subunits), suggesting that IE can bind to $\alpha_v\beta_3$ integrins [74].

The amino acid trimer motif Arg-Gly-Asp (RGD) is commonly found in integrin binding proteins, including disintegrins, ornatin, and many extracellular matrix proteins. The RGD motif mediates binding to several integrin receptor variants, a binding which often can be out-competed by synthetic RGD peptides, confirming the surprising simplicity of this adhesive interaction [75]. RGD as well as other integrin binding motifs are often found in loops bounded by Cys residues, and the motif together with the flanking residues may determine the integrin type specificity [76–77].

The 3D7 proteome was searched for occurrences of the RGD motif, and a high number of motifs was found to be present in PfEMP1 (23 out of 244 motifs, $P = 5.8*10^{-6}$, cumulative binomial distribution with x = 23 motifs, p(RGD) = (244 motifs / 4099411 AA), n = 138055 AA). PfEMP1 domains from seven genomes were then searched, and significantly higher numbers of RGD motifs than what should be expected for random reasons (taking the skewed PfEMP1 amino acid distribution into account) were found in DBLα0 (56 motifs in 229 domains, $P = 5.2*10^{-14}$, cum. binom. distrib. with x = 56 motifs, $p(RGD) = 1.77*10^{-4}$, n = 98157 AA) and to a lesser degree in NTS (12 motifs in 311 domains, $P = 1.1*10^{-4}$, cum. binom. distrib. with x = 12 motifs, $p(RGD) = 1.77*10^{-4}$, n = 20511 AA). Only one motif was found per DBLα0 domain, and all seven genomes had RGD-containing DBLα0 domains. Interestingly all RGD motifs were evenly distributed in three fixed positions in DBLα0: (1) HB19 position 6–8, (2) HB12 position 14–16 and (3) HB7 position 15–17.

The three RGD sites in DBLα0 were predicted to be situated in loop regions by domain structure homology modeling (data not shown), and especially RGD position 2 and 3 were exposed on a loop in subdomain S3, between the helices covered by HB1 and HB2, held in place by several Cys residues.

PfEMP1 similarity to disintegrin and ornatin was found by searching 311 PfEMP1 against the Pfam domain database [56], resulting in six hits to the disintegrin domain, and five hits to ornatin (E<1 for all hits). 10 of these 11 hits were situated in DBLα0, overlapping the second RGD position mentioned above, and not all of the hit sequences contained an RGD motif.

The finding of two independent significant sequence features pointing towards integrin binding, and on top of this, the co-localization of these features in DBLα0, suggests that some DBLα0 domains are likely to mediate integrin binding, which may also be the phenomenon observed by Siano *et al.* (1998) [74].

In relation to this, pentamidine is an RGD analogue used for treatment of many pathogen-caused diseases including malaria [78], and it is possible that this drug may work partly as integrin antagonist, thus to some extent inhibiting IE binding to endothelial cells.

CIDR homology block alignment

158 homology blocks found in 552 CIDR domains were aligned and clustered by HB composition (Figure 8). CIDR domains could be divided into two major groups, CIDRβ/γ/δ containing HB22, and CIDRα with HB23 (Figure 8 M1). No significant homology block similarities were observed between CIDRα and CIDRβ/γ/δ, except the core homology blocks. The CIDRβ, γ and δ domain classes could each be distinguished by class-specific homology blocks, as could each of the CIDRα1 and CIDRα3 subclasses (Figure 8). HB148 described a distinct subgroup of CIDRγ sequences with high similarity to CIDRβ (Figure 8 M1, purple). HB148 was present in 32 PfEMP1 including amongst other PF11_0008 and MAL6P1.4 associated with severe disease [34] and IT4var60 expressed on rosetting IE [16]. Two other interesting CIDR homology blocks, HB450 and HB451, were strongly associated with the previously mentioned conserved domain cassette 8 (Figure 3).

In M2, which for some CIDRα has been proven to mediate CD36 binding [45–46], four types of sequences were found to fill the helix loop between the conserved core HBs (Figure 8 M2). CIDRα1 domains, which have been shown not to bind CD36 [47], shared HB121 in the M2 helix loop, which was markedly different from HB32 shared by the remaining CIDRα in this

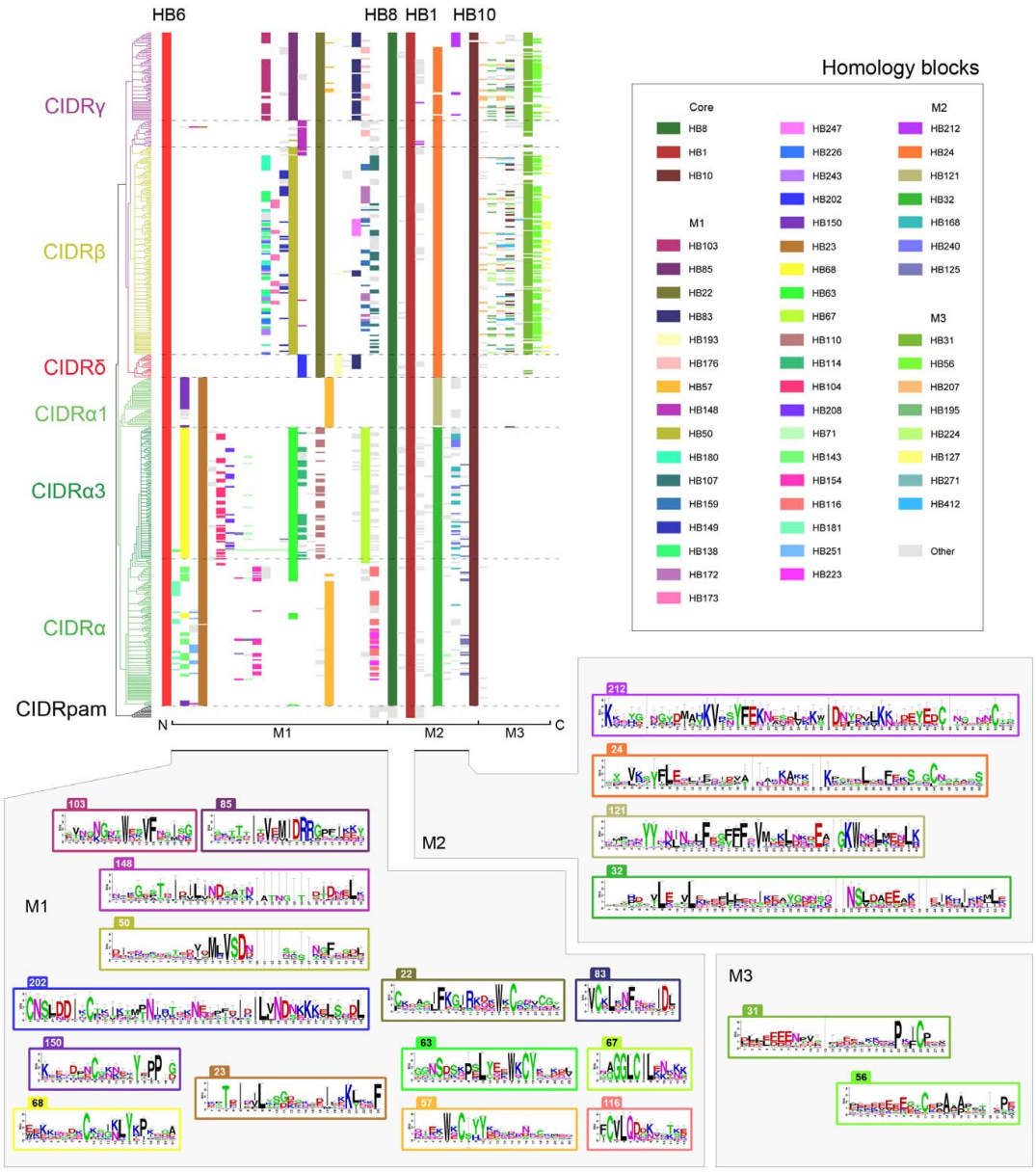


Figure 8. CIDR and M3 homology block alignment. Homology blocks in CIDR domains and M3 regions were aligned, and clustered based on differences in HB composition. The cladogram is colored according to amino acid level domain classification. Only the 54 most frequent HBs are colored, out of a total 158 HBs. Sequence conservation logos are shown for selected HBs in the regions M1–3. Core homology blocks HB1, 8 and 10 are described in Figure 5, while HB6 is the C-terminal of the upstream DBLα/δ domain (Figure 6). Alignments and logos for all HBs can be found by querying the VarDom server with the HB numbers.
doi:10.1371/journal.pcbi.1000933.g008

region. CIDRβ/δ/γ domains were characterized by HB24 in M2, except CIDRγ6/8 domains with a differing helix loop defined by HB212 (Figure 8 M2).

Using the VarDom server, two HBs were found in the helix-loop of the MC179 CIDRα structure: HB32 covering helix a and b (see logo in Figure 8 M2), and HB372 covering the small helix c,

a sequence which is mainly present in CIDRα2 domains (Figure 9). Though the structure appeared twisted in the crystal so helix a and b were slightly separated, it was found likely that semi-conserved HB32 hydrophobic positions 17, 18 and 21 in helix a, under monomeric circumstances interact with conserved HB32 hydrophobic positions 45, 48, 51 and 52 in helix b to keep the helices together (logo in Figure 8 M2; Figure 9, green residues). Similarly, the highly conserved HB32 positions 8 and 12 in helix a binds helix c through conserved hydrophobic interactions (Figure 9, green residues). The Asp-Ile-Glu (DIE) motif at HB32 position 44–46 supports CD36 binding, as binding ability has been found to be disrupted when the motif is substituted with the motif Gly-His-Arg [46]. This substitution of a conserved hydrophobic Ile with a charged His residue in helix b, is likely to result in a different conformation of these helices, emphasizing the importance of this helix pairing in CD36 binding. HB32 position 33–41 shows that in a subset of CIDRα (28% of the HB32 sequences), an insertion containing several acidic residues appears at the apex between helix a and b. In the majority of CIDRα, this apex contains a semi-conserved Tyr-Gly-Asn (YGN) motif on position 25 to 28 in HB32, which may also be surface-exposed in the monomeric structure. Phosphorylation sites are predicted in all HB32 sequences, and when present, the Tyr in YGN is also predicted as target for this modification. Phosphorylation is involved in CD36 binding, though only phosphorylation of the CD36 receptor has been shown [79–80].

A summary of homology block combinations specific for major DBL and CIDR classes can be found in Table S1. Most major classes can be distinguished by a few homology blocks, the exception being the mixed groups DBLγ and DBLε. Table S1 only shows combinations involving presence of homology blocks, and CIDRγ is hard to describe in this way, though it can easily be described by the presence of HB22, combined with the absence of HB50 and HB202 (Figure 8). These domain class specific homology blocks should be useful when analyzing functional differences, as well as for oligonucleotide array and recombinant protein design.

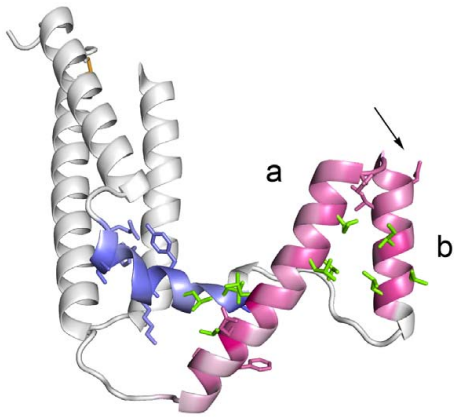


Figure 9. Helix-loop of MC179 CIDRα. HB32 (red) covering helix a and b, and HB372 (blue) covering helix c. Side chains conserved by more than 2.16 bits are shown. Green side chains are conserved hydrophobic residues. The arrow indicates Asn in the possibly surface exposed semi-conserved motif YGN at the apex of helix a and b. The conservation of residues in HB372 with 9 sequences has a high margin of error.
doi:10.1371/journal.pcbi.1000933.g009

PfEMP1 DBL domain relations to CIDR and paralog domains were also studied by means of the homology blocks, and the results are described in Text S3, including: PFA0665w containing distantly related DBL and ATS elements, PfDBLMSP with DBLε-like domains, paralog specific homology blocks, and support for the association between the CIDRpam and other CIDR domains.

NTS homology blocks

NTS homology blocks were aligned and sorted according to HB composition (Figure 10 NTS). Two homology blocks, HB20 and HB17, were found in the NTS of all PfEMP1 except VAR2CSA. HB20 described the pentameric motif [KR]xLx[EQD] known as the *Plasmodium* export element (PEXEL), which is required for protein transport to the host erythrocyte [81]. The motif constituted part of a longer motif with conserved positions every 3–4 amino acids, suggesting a conserved side of a structure predicted to be helical [36]. Even more highly conserved were the initial positions of HB17, the LkGxLxxA motif (Figure 10 NTS), which may be an extension of the PEXEL structure or of the downstream DBLα domain. NTSpam lacks the typical PEXEL motif despite of being present on the IE surface, which could be explained by a unique PEXEL motif in HB309 or HB65, both having three conserved hydrophobic positions with a basic and acidic residue conserved on each side of the middle position (HB309 position 7–15, HB65 position 5–9), like PEXEL in HB20 position 4–11.

Possible N-terminal N-myristoylation of group A PfEMP1 may anchor N-terminal in membrane and cause alternate transportation to IE membrane. HB155 and HB264 were found in the N-terminal of group A PfEMP1, containing the characteristic motif MGxxx[S/T] required for the lipid modification N-myristoylation (Figure 10 NTS). N-terminal N-myristoylation is the covalent attachment of a 14-carbon myristate group to N-terminal Gly through an amide bond, after removal of the start Met residue [82]. This reaction generally takes place in the cytoplasm during protein synthesis and entails transfer of the lipid chain from myristoyl-CoA, catalyzed by N-myristoyltransferase [83] (reviewed by Resh 2006 [84]). Myristate is able to insert hydrophobically into a lipid-bilayer, and thus create an unstable binding to a membrane [85]. Attachment of an N-myristoylated protein to the membrane can be stabilized by the presence of basic residues interacting with negatively charged membrane phospholipids [86], or by further acylation of the protein [87]. Two important roles for N-myristoylation are in membrane anchoring and protein trafficking [88].

N-myristoylation is conserved across eukaryotic species [89], and several experimentally confirmed N-terminally myristoylated proteins in *P. falciparum* share the common eukaryotic motif MGxxx[S/T] [90–94]. The myristoylation predictor NMT, which is trained on several eukaryotic species including protozoans [95–96], correctly predicts that the terminals of these five experimentally analyzed *P. falciparum* proteins are N-myristoylated. The two homology blocks, HB155 and HB264 were present in 41 PfEMP1 N-terminals (Figure 10 NTS) that were all predicted to be N-myristoylated by the NMT predictor. Prediction results for 311 PfEMP1 sequences are summarized in Figure 11A, which shows that the N-myristoylation motif was found predominantly in group A PfEMP1. Remarkably, all seven *P. falciparum* genomes had a set of PfEMP1 with conserved N-myristoylation motifs (Figure 11A).

N-myristoylation may act as a localization signal and affect trafficking of PfEMP1, like PfGRASP which is dependent on a functional myristoylation motif for localization to the golgi

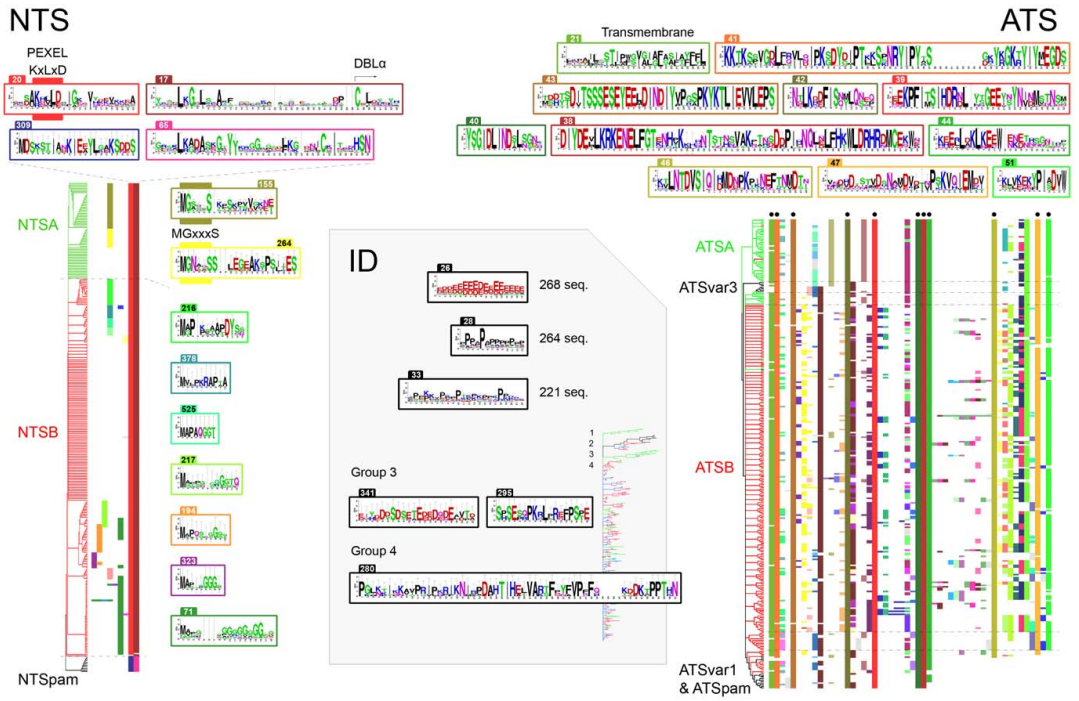


Figure 10. NTS, ID and ATS homology blocks. (**NTS**) Above the HB alignment, sequence conservation logos are shown for the two most conserved NTS homology blocks. The lower pair were found in NTS of VAR2CSA, and HB65 was also found in several DBLβ domains (Figure 6). The proposed PEXEL motif is noted above the HB20 logo, which together with several downstream positions was conserved in all PfEMP1 except VAR2CSA. On the right side of the alignment, logos covering the N-terminal methionine are shown. A conserved N-terminal N-myristoylation motif was found in NTSA HB155 and HB264. (**ATS**) Sequence logos for conserved ATS homology blocks marked by black dots in the alignment. The cladogram is colored according to ATS annotation based on amino acid alignment. Three conserved homology blocks were absent in VAR1 and VAR2CSA ATS. (**ID**) Inter-domain HBs were defined as HBs which occur with a frequency >50% outside other defined regions. Logos for three of the most conserved ID homology blocks are shown, with number of occurrences in the database with 311 PfEMP1 sequences. The phylogram is based on PfEMP1 differences in ID HB composition, where four interesting groups were distinguished: (1) VAR1, (2) VAR2CSA and PfEMP1 with C-terminal similarities to VAR2CSA defined by HB206, (3) group with UPSA flanked *var* including PFD1235w defined by HB295 and HB341 (4) UPSB flanked *var* defined by HB280. The tree is colored according to UPS type, where UPSA is green, UPSB is red, UPSC is blue and UPSE is black. Homology block sequence logos specific for group 3 and 4 in the phylogram are shown.
doi:10.1371/journal.pcbi.1000933.g010

apparatus through a brefeldin A independent pathway [94,97]. PfGRASP has a terminal sequence (MGAGQTK) which is very similar to IT4var08 (MGAGQST) and RAJ116var05 (MGASQSK), the latter getting the highest score of all PfEMP1 by the NMT predictor.

It is still unknown if the PEXEL motif is cleaved and acetylated in PfEMP1, like in some other exported proteins [98–99]. If NTS is not removed by PEXEL cleavage, then the N-myristoylated N-terminal can be translocated across the membrane [100–101], and exposed on the IE surface. The unstable membrane binding caused by N-terminal N-myristoylation could by itself play a major role in mediating adherence of IE to host cell membranes. The unspecific binding of several acylated PfEMP1 to any part of a host cell (e.g. endothelial cell) membrane, possibly combined with receptor binding mediated by other parts of PfEMP1, could together form a strong interaction. A mechanism known as myristoyl switching has been found in some acylated proteins, where ligand binding induces a conformational change, regulating if the fatty acid is hidden in a hydrophobic pocket within the protein or if it is exposed for membrane interactions [102].

Stable membrane anchoring is also possible, as the N-terminals of some PfEMP1 possess several basic residues that can act in synergy with the lipid chain to bind the membrane. Generally UPSA have a higher pI (i.e. are more basic) than other PfEMP1 N-terminals (Figure 11B). Other types of less site-specific acylation, such as S-acylation at some of the many Cys residues, may also help tether the protein to the membrane [87].

The potentially affected group A PfEMP1 have been associated with severe malaria [9,28]. Considering the implications for vaccine design, it should therefore be thoroughly investigated if any of the PfEMP1 variants are indeed myristoylated *in vivo*.

Inter-domain homology blocks

52 homology blocks had more than 50% of their occurrences in inter-domain regions, i.e. outside defined domains. Three of the most frequent inter-domain homology blocks are shown in Figure 10-ID. The 52 inter-domain homology blocks were mainly low complexity sequences, occurring in repeats and overlapping each other. To determine the distribution of these homology

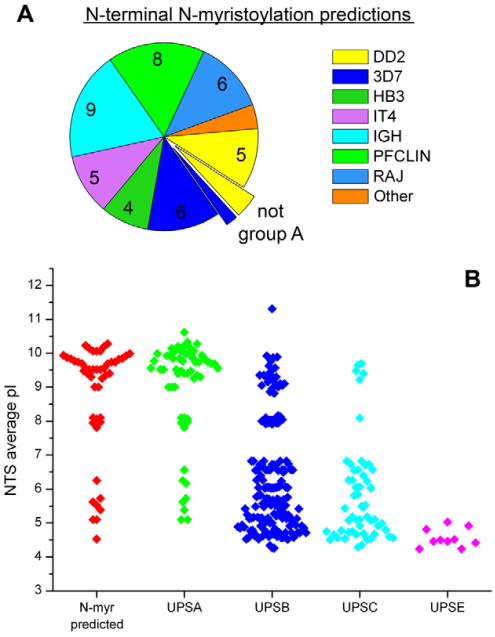


Figure 11. N-terminal N-myristoylation predictions. (A) 48 positive NMT predictions in 311 PfEMP1 N-terminals. All except three were group A PfEMP1. According to the predictions, the post-translational modification was well conserved in all seven genomes. (B) Average pI of NTS in 311 PfEMP1. Three groups (basic, neutral and acidic) can be clearly distinguished.
doi:10.1371/journal.pcbi.1000933.g011

blocks in PfEMP1 sequences, a NJ-tree was constructed based on ID homology block composition of the PfEMP1 (Figure 10 ID). In general the homology blocks were uniformly scattered amongst PfEMP1 sequences, although four groups were distinguished with representatives in at least 6 of the 7 genomes. VAR1 and VAR2CSA had unique conserved inter-domain sequences with low amounts of the low-complexity sequence found in many other PfEMP1, and therefore, they formed separate groups (Figure 10 ID, tree group 1 and 2). Interestingly, one cluster was defined by two unique inter-domain homology blocks, HB341 and HB295 (Figure 10 ID, group 3). This cluster of 11 group A PfEMP1 with similar DBLβ/γ containing domain composition (part of frame 9 in Figure S4A) captured all occurrences of double DBLβ domains, was represented in 6 of 7 genomes (not RAJ116), and the 3D7 genes were PFD1235w and PF11_0521, which have been linked to severe malaria and ICAM-1 binding respectively [28,38]. The fourth distinct group was defined by HB280 (Figure 10 ID, group 4), conserved in 5 of 7 genomes (not 3D7 and HB3) and comprised 11 proteins, including among others the ICAM-1 binding associated IT4var14 (A4var) [40]. All members in the fourth group lacked other ID HBs, most were flanked by UPSB1, and 10 of 11 had the same C-terminal domain combination ending with DBLγ-DBLζ4 (Figure 3, cassette 9; Figure S4A, frame 7). The conservation of an ID region together with the semi-conserved domain architecture and UPS sequences, suggests a more recent common ancestor for genes in these groups. It will be interesting to see if the members of these groups share receptor-binding properties.

The Cys-containing M3 regions (M3A and M3AB) were found to be positionally linked to the upstream CIDR domain, while the amino acid composition correlated more highly with the downstream domain architecture. Two homology blocks were able to capture most occurrences of the two Cys-residues found after CIDRβ and γ, despite of the surrounding low-complexity sequence, seeing that a few other positions besides the Cys were conserved (Figure 8 M3).

ATS homology blocks

Homology blocks of the conserved ATS were aligned and sorted according to domain composition, to describe variation in the intracellular part of PfEMP1 (Figure 10 ATS). ATS starts N-terminally with the transmembrane region, which was captured by HB21. The intron splice site between exon 1 and exon 2 lies immediately downstream of the transmembrane part, so the short basic stretch which follows transmembrane regions, and interacts with the negatively charged membrane phospholipids, was found in the following HB41. ATSA, which is associated with UPSA, was distinguished as sequences where HB69 and HB112 occurred simultaneously.

ATSvar1, ATSB17, and the ATS of VAR2CSA, were characterized by lacking the final three homology blocks conserved in all other ATS (Figure 10 ATS, HB46/47/51, Figure S7D).

ATSB17 was found in six group C PfEMP1, distributed in six genomes (not IT4), and containing several DBLβ/γ domains. The two *var2csa* genes in the HB3 genome had an ATSB14 more similar to the ATS of non-VAR2CSA PfEMP1, however these were truncated before the final three homology blocks. Other VAR2CSA ATS had normal length but contained unique sequences instead of the three conserved homology blocks. The five *var1* genes possessing an exon 2, were all flanked by a 3′UTR encoding the three missing homology blocks. Compared to a common ATS, ATSvar1 was missing ~150 AA, ATSB17 was lacking ~100 AA, whereas the ATS of VAR2CSA was missing or differed from the final 100–130 AA.

The finding that VAR1 and VAR2CSA both have a shortened ATS, could suggest that ATSvar1 is functional despite of truncation, and question the hypothesis that VAR1 exclusively exists as a pseudogene.

The final three ATS homology blocks could be a non-essential functional element in PfEMP1, for example acting as signal peptide during transport to the erythrocyte membrane, which would result in differences for VAR1, VAR2CSA, and ATSB17 PfEMP1, compared to other PfEMP1.

Conserved homology block residues may comprise phosphorylation sites

Phosphorylation occurs mainly at three types of residues: Ser, Thr and Tyr, and all three residues were markedly conserved in several homology blocks. Phosphorylation is a common modification of proteins expressed during the erythrocyte stages, and has been associated with differences in IE adhesion properties [103]. Ser/Thr phosphorylation of the PfEMP1 ATS was recently shown to alter its association with parasite-encoded knob-associated His-rich protein (KAHRP), and to regulate cytoadherence of IE [104].

Judging from phosphorylation site predictions and conservation levels in the homology blocks, some examples of conserved potential phosphorylation sites were, in DBL domains (Figure 6): HB19 position 28 (DBLα S1), HB82 position 11 (DBLα S2b), HB36 position 8 (DBLα0 S2c), and Tyr in HB29 (DBLδ and γ S2c). In CIDR one of many examples is the mentioned YGN motif in CIDRα HB32 (Figure 8). Several sites of all three types are conserved in the ATS HB41, HB43, and HB69 (Figure 10 ATS).

Phosphorylation sites have been predicted for all PfEMP1 sequences, and the conservation of these can be inspected for each homology block on the VarDom server.

It will be interesting to see if some of these sites are surface-exposed and thus accessible to kinases, as the introduction of large, negatively charged phosphate groups could result in conformational changes, or contribute to charged binding surfaces, and thus result in functional and antigenic variation.

Overall PfEMP1 homology block architecture

Homology block sequences of full-length PfEMP1 were aligned, to determine HB associations with specific positions in the whole proteins, as well as to find groups of PfEMP1 with similar HB compositions. Sequences were sorted according to NJ-clustering based on Manhattan distances between feature vectors consisting of exon 1 HB counts. The homology block alignment shown in Figure 12 gives a detailed overview of the diversity and structure in the PfEMP1 family. A labeled version of the alignment and the tree can be found in Figure S7E and Figure S8, respectively.

The differences between UPSA, B, and C flanked *var* genes were not clear enough to form separate clades in the tree, though homology blocks specific for UPSA-flanked *var* were observed in both ends of the alignment (Figure 12a comp.1 and Figure 12b ATS). The three conserved genes were all clearly distinguished (Figure 12,clade A, B, and E), as well as many small PfEMP1 groups, generally with low bootstrap support, as expected from uncorrelated domains in N- and C-terminal (Figure S4).

A list of homology blocks specific for each of the four components are summarized in Table S2. These specific homology blocks may be helpful for functional analysis of the PfEMP1, as well as for genotyping purposes.

Conclusion

The reclassification of PfEMP1 domains by alignment and distance tree analysis introduced a few larger and several smaller new subclasses. Although the classification is a result of a phylogenetic approximation of the different evolutionary histories of the domain sequence blocks, identification of conserved PfEMP1 domain architectures was possible. These structures represent a novel perspective on the PfEMP1 architecture. DBL and CIDR domains appear to be inherited in conserved domain structures that to a large extent fall within four major components.

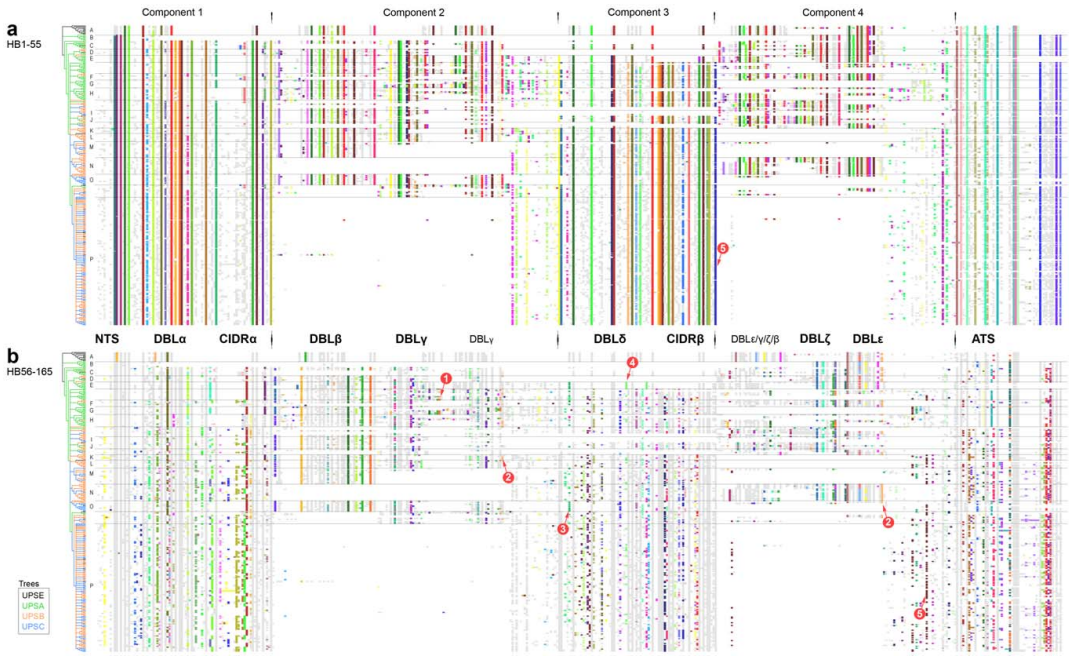


Figure 12. PfEMP1 homology block alignment. (**a**) and (**b**) are the same alignment, with HB1–55 colored in (a), and HB56–165 colored in (b). The sequences are sorted according to HB composition, and the tree is colored according to UPS class. The division of PfEMP1 into four components is indicated at the top of the figure. Between (a) and (b) is noted the most prevalent major domain class for that area in the alignment. The five core homology blocks should be distinguishable in (a), as well as less frequent homology blocks especially in (b). The alignment with all details can be found in Figure S7E, and the labeled tree in Figure S8. Alignment features (red arrows): (**1**) DBLγ-β hybrid domains; (**2**) The light orange column is HB78, present in both DBLγ and DBLε (Figure 6n, r, and t) and associated with C-terminal of comp. 2 and 4; (**3**) HB74 in DBLγ-like DBLδ domains, as in Figure 6g, h and Figure 7-S1, S2; (**4**) HB82 in DBLγ8 of VAR1, also found in DBLδ domains; and (**5**) M3 homology blocks. Notable clades in the tree: (**A**) VAR2CSA; (**B**) VAR3; (**C**) bootstrap 28%, 4 genomes, UPSA3, includes IT4var60 (rosetting); (**D**) bootstrap 25%, 3 genomes, incl. PFL0020w and PF08_0141; (**E**) VAR1; (**F**) 6 genomes, incl. MAL6P1.4; (**G**) 5 genomes, incl. PFD1235w and PF11_0521 (ICAM-1); (**H**) 5 genomes, incl. PF11_0008 and PF13_0003; (**I**) 4 genomes, incl. PF07_0050 and IT4var31 (CD36, ICAM-1); (**J**) 4 genomes, incl. IT4var14 (CD36, ICAM-1); (**K**) bootstrap 27%, 5 genomes, UPSB2, incl. PF08_0140 and IT4var06; (**L**) bootstrap 26%, 3 genomes, incl. IT4var16 (CD36, ICAM-1) and IT4var27 (rosetting); (**M**) bootstrap 18%, all genomes incl. MAL6P1.252 and PFL1950w; (**N**) bootstrap 68%, 5 genomes, UPSB; (**O**) bootstrap 49%, 5 genomes, UPSC1, incl. IT4var01 (rosetting) and TM284S2var1 (rosetting, IgG); and (**P**) Comp.1-Comp.3-ATS architecture (a.k.a. Type 1 *var*), UPSB and UPSC.
doi:10.1371/journal.pcbi.1000933.g012

These conserved domain structures although large and complex may well represent functional units of the whole PfEMP1 molecule.

Apart from the known conserved *var* genes, *var1*, *var2csa*, and *var3*, 18 domain cassettes and several less well-defined structural phenomena were observed for the seven sequenced genomes. The established division of group A, B and C was confirmed although importantly, N- and C-terminal conserved domain structures occurred independently of each other, with distinct C-terminal DBLε-containing structures transcending the three conserved genes, as well as group A, B, and C.

Homology blocks covering on average 83.5% of a PfEMP1 sequence were defined, describing the PfEMP1 family on a more detailed level than domains, yet more simplified than the amino acid level. Local similarities between domain classes were thus described, and homology blocks specific for PfEMP1 domain classes, components, and cassettes, were found. The HB analysis also revealed a recombination hotspot between subdomain S2 and S3 in DBL domains, which has helped shape the antigen repertoire. Thus, several DBL domains are hybrids of different major classes - an observation important for functional studies as well as antibody cross-reactivity and vaccine design.

Several conserved elements were described by the homology blocks, including: (1) DBL domain core interactions conserved in all DBL domains, holding the subdomains together, (2) an acylation motif found to be conserved in group A *var* genes, suggesting N-terminal N-myristoylation of a subset of PfEMP1, (3) conserved residues predicted to be phosphorylation sites, and (4) PfEMP1 inter-domain regions, which are proposed to be elastic disordered structures.

The novel iterative homology block detection method is potentially applicable to any protein dataset, and would be especially suitable for compositional analysis of other frequently recombining gene families.

The VarDom server was introduced, where all presented information on domain classes and homology blocks can be retrieved, and new sequences can be classified and related to other PfEMP1 proteins in the seven genomes. Ideally, the server will allow better interpretation and facilitate the development of new approaches in PfEMP1 research. For example analysis of *var* expression data from microarrays and short high through-put sequence reads or the design of recombinant proteins for immunizations or functional studies could all benefit from this detailed account of PfEMP1 diversity and ultimately aid the development of PfEMP1 based malaria interventions.

Methods

Datasets

Annotated *var* genes and *var* gene containing contigs were retrieved using BLAST, from NCBI nucleotide database and from genome assemblies of *P. falciparum* clones 3D7, HB3, DD2, IT4/FCR3, PFCLIN, RAJ116, IGH and *P. Reichenowi* clone PREICH at PlasmoDB, Broad and Sanger Institute servers, querying 3D7 *var* sequences. For all *var* genes with intact N-terminal segments, 2000 bp 5′ UTRs were also retrieved where possible. In total 399 annotated genes and open reading frames spanning over the length of at least two DBL/CIDR domains were kept for the sequence alignment and distance tree analysis, whereas the homology block dataset consisted of the 311 full length or exon1 sequences, as well as 20 DBL-containing paralogs from *Plasmodium falciparum*, *vivax*, *yoelii* and *knowlesi*. For meaningful interpretations, the first approach required sequence lengths spanning at least two PfEMP1 features, whereas the latter, was based on whole or exon1 sequences to avoid generating false homology block break-points. Nucleotide sequences of all *var* genes analyzed in this study are available in Dataset S1.

Domain alignment and phylogeny

Large phylogenies comprising all DBL or CIDR sequences were inferred by multiple sequence alignment using MUSCLE (version 3.7) followed by application of the neighbor-joining algorithm implemented in MEGA (version 4.0.2) [105]. Major domain classes were deduced and named according to previously defined classes [10].

Major domain-class sequences were further subclassified through a recursive process involving: (1) re-alignment of sequences, (2) construction of a maximum likelihood tree, and (3) split of sequences into two clusters at a tree bipartition validated by at least 50% of the bootstraps. If a suitable bipartition was found, the process would be repeated for each of the two formed clusters. If the sequences on the other hand were not divided, they were all assigned to the same subclass and given a number. In addition to bootstrap support, two other properties were used to evaluate bipartitions and determine if and where the trees should be split: the number of genomes represented in each cluster, and the within-cluster average distance (WCAD), which was used as a measure for the relatedness of clustered sequences. See Text S1 for details on domain border and distance tree cluster definitions.

Multiple sequence alignments of PfEMP1 domains were performed with AQUA [106], which optimizes alignments generated by MUSCLE (version 3.7) [107] and MAFFT (version 6.611b) [108], using refinement and evaluation implemented in RASCAL (version 1.34) [109] and NORMD (version 1.3) [110], respectively. Maximum likelihood trees were built using the multithreaded version (pthreads) of RAxML (version 7.2.5) [111–112]. The gamma model for substitution rate heterogeneity was used together with the WAG [113] amino acid substitution model with empirically determined amino acid frequencies. WAG and JTT [114] were found to be the most likely substitution models by fitting of models implemented in RAxML to fixed trees built from the different domain alignments and subsequent ML comparison. Within-cluster average distances were based on distances calculated using the JTT model implemented in Protdist from the PHYLIP package (version 3.69) [115].

Upstream sequences

Sequences were aligned with MAFFT (version 6.240) using the L-INS-i algorithm for multiple sequence alignment [108]. A neighbor-joining tree was generated and bootstrapped using Clustalw (version 2.0.9 for tree construction and version 1.83 for bootstrapping because version 2.0.9 crashed during bootstrap) [116].

Sequences were clustered using the Markov clustering algorithm (version 08-312) [51–53]. The Markov clustering algorithm is a graph-theoretical clustering method, which uses an all-against-all pairwise sequence alignment as input, generated with the blastn algorithm implemented in blastall (version 2.2.18) [117]. The inflation parameter of the Markov Cluster Algorithm was varied in steps of 0.2 from 1.2 to 5.0, and resource scheme 7 (most accurate) was used. A distinct clustering was generated for each value of the inflation parameter, and all the clusters were summarized in a consensus clustering. Briefly, each clustering was converted to a multifurcating tree with a branch representing each cluster. A consensus tree representing the consensus clustering was then constructed, using the majority rule consensus method (include all bipartitions with a frequency larger than 0.5) [118], with the extension that less frequent bipartitions were also included as long

as they continued to resolve the tree and did not contradict more frequent groups. Based on the results of the two clustering methods, a consensus annotation of the 5′ upstream sequences of the *var* genes was reached (Figure S5).

Trees were rendered and edited using Dendroscope (version 2.3) [119].

Homology block alignment and trees

The iterative homology search procedure used for defining the set of 628 homology blocks is described in Text S2.

Alignment of homology blocks was performed with a python implementation of the Smith-Waterman algorithm with linear (non-affine) gap penalty and a substitution matrix of the identity type [120].

To estimate trees based on homology block composition, homology block feature vectors were constructed for each sequence, either binary (DBL, CIDR, ATS, ID and NTS trees) or with counts (PfEMP1 tree), and accordingly distances were calculated as either Hamming or Manhattan distances. Trees were constructed as extended 50% majority rule consensus trees, based on 1000 neighbor joining bootstrap trees, built from distance matrix using ordinary neighbor joining implemented in Clearcut (version 1.0.8) [121].

Sequence logos were generated using WebLogo (version 2.8) [122], where small sample (<40 amino acids) bias is compensated for by subtraction of an error estimate on each position, the error bars are 2 times the estimated error.

Prediction of phosphorylation sites and N-terminal N-myristoylation

Phosphorylation sites were predicted using NetPhos 2.0 [123]. N-terminal N-myristoylation was predicted with the NMT myristoylation predictor which is trained for several eukaryotic species including protozoans [95–96].

Supporting Information

Dataset S1 *Var* gene sequences. *Var* gene cDNA encoding the PfEMP1 analyzed in this study. Sequence names in this fasta-file are the same as used everywhere else in this study, as well as on the VarDom server.
Found at: doi:10.1371/journal.pcbi.1000933.s001 (2.84 MB TXT)

Figure S1 Major DBL and CIDR domain classes. (A) NJ tree based on amino acid alignment of 1242 DBL sequences. Blue dots mark branches dividing DBL domains into six major groups and four N-terminal VAR2CSA DBL classes. (B) NJ tree based on amino acid alignment of 655 CIDR sequences. Blue dots mark branches dividing CIDR domains into four major groups as well as the CIDRα1 and CIDRpam subclasses. Leaf names are omitted from the figure to improve graphical presentation.
Found at: doi:10.1371/journal.pcbi.1000933.s002 (2.49 MB PNG)

Figure S2 Trees showing subclassification of all major PfEMP1 domain classes. ML trees based on amino acid alignments of each of the following domain classes are shown in panels **A–M**: DBLα0, α1, β, δ, ε, γ, ζ; CIDRα, β, γ, δ; NTS; ATS. Sequence names as well as start and stop position of the domains are given in the trees, followed by classification of the domain. Panel **N** and **O**: Assignment of sequences to UPS groups by Markov clustering (N) and neighbor joining (O). The UPS groups were named as indicated by the text color. The background colors show the group membership assigned by Kraemer *et al.* 2007 [16]. Sequences found upstream of domain cassette 8 (Figure 3) are marked with black squares. (**N**) The branch labels show the fraction of Markov clusters with this group present. (**O**) The branch labels show the bootstrap values as fractions of 1000 bootstraps. Monophyletic subgroups with a bootstrap support above 0.7 and containing sequences from at least four different strains of *P. falciparum* are highlighted with thick red branches. Some subgroups were further expanded (without bootstrap support) to form larger monophyletic groups: UPSA2 and UPSB3 are expanded to include additional sequences annotated to UPSA2 and UPSB3 respectively by Kraemer *et al.* 2007 [16], UPSB2 is expanded to include two genes with same domain architecture, and UPSC1 is expanded to include three sequences that fall between UPSC1 and UPSC2 but within the larger monophyletic group comprising all UPSC sequences. The sequences are shown with thick black branches. The additional sequences included by this expansion are denoted with an asterisk in the annotation in Figure S4 and S5. UPSA3 and UPSB1 are groups that contain all the sequences not assigned to any other subgroup in UPSA and UPSB respectively. ND: Not Determined.
Found at: doi:10.1371/journal.pcbi.1000933.s003 (1.11 MB ZIP)

Figure S3 PfEMP1 domain class logos. Sequence conservation logos for major PfEMP1 domain classes (panel A–Z): CIDRα, α1, α2, α3, β, δ, γ, pam; DBLα0, α1 (without α1.3), α1.3, β, δ, ε (without ε1, ε2, ε11, ε13, εpam), ε1, ε2, ε11, ε13, εpam4, εpam5, γ, pam1, pam2, pam3, ζ; NTSA, NTSB, and M3AB.
Found at: doi:10.1371/journal.pcbi.1000933.s004 (2.42 MB ZIP)

Figure S4 Annotated PfEMP1 sequences aligned according to C-terminal (A) and N-terminal (B) domain compositions. Gene names, parasite genome, 5′ UPS classes, PfEMP1 domain annotation (D = domain, ID = Inter Domain) and origin of sequence data (if sequence is not previously reported as *var* gene) are given. Sequences which partially contain unexpected identical sequence stretches to other sequences suggesting an incorrect contig assembly are noted "HBD" followed by the name of the potentially redundant sequence. Red arrows indicate component 1–4. Frames indicate clusters of correlated domain classes. 1:VAR1; 2: VAR2CSA; 3: VAR3; 4: DBLζ and DBLε domain combinations of component 4; 5: Cassette 10; 6: Cassette 6; 7: Cassette 9; 8: Cassette 5; 9: Other Group A PfEMP1 all containing component 2; 10: Cassette 8; 11: Group B and C genes containing component 2; 12: Group B and C PfEMP1 with no component 2 or 4; 13: Cassette 14; 14: Cassette 17,21 and 22; 15: DBLα1-CIDR subclass correlations including cassette 11,13,15 and 16; 16: DBLα0 subclasses associated with CIDRα3 subclasses; 17: DBLα0 subclasses associated with CIDRα2 subclasses. N-terminal segment (NTS), Duffy binding-like (DBL), Cys-rich inter-domain region (CIDR) and acidic terminal segment (ATS) are named according to the distance tree classification. Inter domains are annotated as either short if <32 AA (green) or long if >31 (yellow) and "A" or "B" if encoding M3A or M3AB.
Found at: doi:10.1371/journal.pcbi.1000933.s005 (0.15 MB PDF)

Figure S5 Schematic representation of annotated *var* genes sorted by genome origin. Gene names, 5′UTR class, domain architecture and origin of sequence data (if sequence is not previously reported as *var* gene) is given. Sequences are noted "F" (Fragment) in comments if predicted not to span a full length exon1, and "HBD" if incorrect contig assembly is suspected followed by the name of the sequence which partially contains unexpected identical sequence stretches. N-terminal segment (NTS), Duffy binding-like (DBL), Cys-rich inter-domain region (CIDR) and acidic terminal segment (ATS) are named according to the distance tree classification.

Found at: doi:10.1371/journal.pcbi.1000933.s006 (0.05 MB PDF)

Figure S6 Phylogenetic trees for DBL subdomains S1, S2 and S3, as in Figure 7 but with labels. Edge values are fractions of 1000 bootstraps, and each subdomain is given as: protein name, start position, end position, and the domain class the subdomain is a part of.
Found at: doi:10.1371/journal.pcbi.1000933.s007 (0.22 MB PDF)

Figure S7 Homology block alignments. Homology block alignments for (panel A–E): DBL, CIDR, NTS, ATS, and whole PfEMP1, with details of Figure 6, Figure 8, Figure 10 and Figure 12.
Found at: doi:10.1371/journal.pcbi.1000933.s008 (0.82 MB ZIP)

Figure S8 Tree in Figure 12 with labels. Bootstrap values are given as fractions of 1000 bootstraps.
Found at: doi:10.1371/journal.pcbi.1000933.s009 (0.33 MB PDF)

Table S1 Examples of HB combinations specific for DBL and CIDR domain classes. Domain counts and number of matches of the HB combination are given for the sequence set with 311 PfEMP1 sequences. The domain combination (17, 19) signifies a sequence where both HB17 and HB19 are present. These homology blocks are suggested for use in oligonucleotide array design, as well as for functional analysis of the domain types. The list is not exhaustive, and can be supplemented using Figure 6 and Figure 8, as well as the VarDom server.
Found at: doi:10.1371/journal.pcbi.1000933.s010 (0.11 MB PDF)

Table S2 Homology blocks specific for component 1–4 (Figure 12). Homology block numbers are given in parenthesis, and number of occurrences in the component with 311 sequences, is given next to the number of occurrences elsewhere. These homology blocks are suggested for use in oligonucleotide array design, as well as for functional analysis of the components. The table is not exhaustive.
Found at: doi:10.1371/journal.pcbi.1000933.s011 (0.09 MB PDF)

Text S1 PfEMP1 domain classification by alignment and distance tree analysis.
Found at: doi:10.1371/journal.pcbi.1000933.s012 (0.15 MB PDF)

Text S2 Defining PfEMP1 homology blocks.
Found at: doi:10.1371/journal.pcbi.1000933.s013 (0.81 MB PDF)

Text S3 PfEMP1 DBL domain relations to CIDR and paralog DBL domains.
Found at: doi:10.1371/journal.pcbi.1000933.s014 (0.31 MB PDF)

Acknowledgments

We are very grateful to those who generated and made available the sequence data of the seven genomes. HB3, DD2, IGH and RAJ116 parasite isolates were sequenced at the NIAID Microbial Sequencing Center at the Broad Institute. The sequence data from the IT clone, the clinical isolate (PFCLIN) and clone 3D7 were produced by the Pathogen Genomics group at the Wellcome Trust Sanger Institute.

Author Contributions

Conceived and designed the experiments: TSR AGP TL. Performed the experiments: TSR DAH TL. Analyzed the data: TSR DAH TGT AGP TL. Contributed reagents/materials/analysis tools: TSR AGP. Wrote the paper: TSR TL.

References

1. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, et al. (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.
2. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, et al. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell 82: 89–100.
3. Staalsoe T, Shulman CE, Bulmer JN, Kawuondo K, Marsh K, et al. (2004) Variant surface antigen-specific IgG and protection against clinical consequences of pregnancy-associated Plasmodium falciparum malaria. Lancet 363: 283–289.
4. Marsh K, Howard RJ (1986) Antigens induced on erythrocytes by P. falciparum: expression of diverse and conserved determinants. Science 231: 150–153.
5. Fried M, Nosten F, Brockman A, Brabin BJ, Duffy PE (1998) Maternal antibodies block malaria. Nature 395: 851–852.
6. Baruch DI, Gamain B, Barnwell JW, Sullivan JS, Stowers A, et al. (2002) Immunization of Aotus monkeys with a functional domain of the Plasmodium falciparum variant antigen induces protection against a lethal parasite line. Proc Natl Acad Sci U S A 99: 3860–3865.
7. Salanti A, Dahlback M, Turner L, Nielsen MA, Barfod L, et al. (2004) Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. J Exp Med 200: 1197–1203.
8. Lusingu JP, Jensen AT, Vestergaard LS, Minja DT, Dalgaard MB, et al. (2006) Levels of plasma immunoglobulin G with specificity against the cysteine-rich interdomain regions of a semiconserved Plasmodium falciparum erythrocyte membrane protein 1, VAR4, predict protection against malarial anemia and febrile episodes. Infect Immun 74: 2867–2875.
9. Magistrado PA, Lusingu J, Vestergaard LS, Lemnge M, Lavstsen T, et al. (2007) Immunoglobulin G antibody reactivity to a group A Plasmodium falciparum erythrocyte membrane protein 1 and protection from P. falciparum malaria. Infect Immun 75: 2415–2420.
10. Lavstsen T, Salanti A, Jensen AT, Arnot DE, Theander TG (2003) Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions. Malar J 2: 27.
11. Kraemer SM, Smith JD (2003) Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family. Mol Microbiol 50: 1527–1538.
12. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498–511.
13. Taylor HM, Kyes SA, Newbold CI (2000) Var gene diversity in Plasmodium falciparum is generated by frequent recombination events. Mol Biochem Parasitol 110: 391–397.
14. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, et al. (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. Nature 407: 1018–1022.
15. Duffy MF, Byrne TJ, Carret C, Ivens A, Brown GV (2009) Ectopic recombination of a malaria var gene during mitosis associated with an altered var switch rate. J Mol Biol 389: 453–469.
16. Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, et al. (2007) Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates. BMC Genomics 8: 45.
17. Salanti A, Staalsoe T, Lavstsen T, Jensen AT, Sowa MP, et al. (2003) Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering Plasmodium falciparum involved in pregnancy-associated malaria. Mol Microbiol 49: 179–191.
18. Viebig NK, Gamain B, Scheidig C, Lepolard C, Przyborski J, et al. (2005) A single member of the Plasmodium falciparum var multigene family determines cytoadhesion to the placental receptor chondroitin sulphate A. EMBO Rep 6: 775–781.
19. Ralph SA, Bischoff E, Mattei D, Sismeiro O, Dillies MA, et al. (2005) Transcriptome analysis of antigenic variation in Plasmodium falciparum–var silencing is not dependent on antisense RNA. Genome Biol 6: R93.
20. Duffy MF, Maier AG, Byrne TJ, Marty AJ, Elliott SR, et al. (2006) VAR2CSA is the principal ligand for chondroitin sulfate A in two allogeneic isolates of Plasmodium falciparum. Mol Biochem Parasitol 148: 117–124.
21. Fried M, Duffy PE (1996) Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. Science 272: 1502–1504.
22. Marsh K, Forster D, Waruiru C, Mwangi I, Winstanley M, et al. (1995) Indicators of life-threatening malaria in African children. N Engl J Med 332: 1399–1404.
23. Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C (1999) Immunity to non-cerebral severe malaria is acquired after one or two infections. Nat Med 5: 340–343.
24. Bull PC, Lowe BS, Kortok M, Molyneux CS, Newbold CI, et al. (1998) Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. Nat Med 4: 358–360.
25. Nielsen MA, Staalsoe T, Kurtzhals JA, Goka BQ, Dodoo D, et al. (2002) Plasmodium falciparum variant surface antigen expression varies between

isolates causing severe and nonsevere malaria and is modified by acquired immunity. J Immunol 168: 3444–3450.
26. Bian Z, Wang G (2000) Antigenic variation and cytoadherence of PfEMP1 of Plasmodium falciparum-infected erythrocyte from malaria patients. Chin Med J (Engl) 113: 981–984.
27. Kaestli M, Cockburn IA, Cortes A, Baea K, Rowe JA, et al. (2006) Virulence of malaria is associated with differential expression of Plasmodium falciparum var gene subgroups in a case-control study. J Infect Dis 193: 1567–1574.
28. Jensen AT, Magistrado P, Sharp S, Joergensen L, Lavstsen T, et al. (2004) Plasmodium falciparum associated with severe childhood malaria preferentially expresses PfEMP1 encoded by group A var genes. J Exp Med 199: 1179–1190.
29. Rottmann M, Lavstsen T, Mugasa JP, Kaestli M, Jensen AT, et al. (2006) Differential expression of var gene groups is associated with morbidity caused by Plasmodium falciparum infection in Tanzanian children. Infect Immun 74: 3904–3911.
30. Kirchgatter K, Portillo Hdel A (2002) Association of severe noncerebral Plasmodium falciparum malaria in Brazil with expressed PfEMP1 DBL1 alpha sequences lacking cysteine residues. Mol Med 8: 16–23.
31. Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, et al. (2006) Differential var gene transcription in Plasmodium falciparum isolates from patients with cerebral malaria compared to hyperparasitaemia. Mol Biochem Parasitol 150: 211–218.
32. Falk N, Kaestli M, Qi W, Ott M, Baea K, et al. (2009) Analysis of Plasmodium falciparum var genes expressed in children from Papua New Guinea. J Infect Dis 200: 347–356.
33. Warimwe GM, Keane TM, Fegan G, Musyoki JN, Newton CR, et al. (2009) Plasmodium falciparum var gene expression is modified by host immunity. Proc Natl Acad Sci U S A 106: 21801–21806.
34. Cham GK, Turner L, Lusingu J, Vestergaard L, Mmbando BP, et al. (2009) Sequential, ordered acquisition of antibodies to Plasmodium falciparum erythrocyte membrane protein 1 domains. J Immunol 183: 3356–3363.
35. Rowe JA, Claessens A, Corrigan RA, Arman M (2009) Adhesion of Plasmodium falciparum-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications. Expert Rev Mol Med 11: e16.
36. Smith JD, Subramanian G, Gamain B, Baruch DI, Miller LH (2000) Classification of adhesive domains in the Plasmodium falciparum erythrocyte membrane protein 1 family. Mol Biochem Parasitol 110: 293–310.
37. Higgins MK (2008) The structure of a chondroitin sulfate-binding domain important in placental malaria. J Biol Chem 283: 21842–21846.
38. Oleinikov AV, Amos E, Frye IT, Rossnagle E, Mutabingwa TK, et al. (2009) High throughput functional assays of the variant antigen PfEMP1 reveal a single domain in the 3D7 Plasmodium falciparum genome that binds ICAM1 with high affinity and is targeted by naturally acquired neutralizing antibodies. PLoS Pathog 5: e1000386.
39. Howell DP, Levin EA, Springer AL, Kraemer SM, Phippard DJ, et al. (2008) Mapping a common interaction site used by Plasmodium falciparum Duffy binding-like domains to bind diverse host receptors. Mol Microbiol 67: 78–87.
40. Smith JD, Craig AG, Kriek N, Hudson-Taylor D, Kyes S, et al. (2000) Identification of a Plasmodium falciparum intercellular adhesion molecule-1 binding domain: a parasite adhesion trait implicated in cerebral malaria. Proc Natl Acad Sci U S A 97: 1766–1771.
41. Chen Q, Heddini A, Barragan A, Fernandez V, Pearce SF, et al. (2000) The semiconserved head structure of Plasmodium falciparum erythrocyte membrane protein 1 mediates binding to multiple independent host receptors. J Exp Med 192: 1–10.
42. Rowe JA, Moulds JM, Newbold CI, Miller LH (1997) P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. Nature 388: 292–295.
43. Russell C, Mercereau-Puijalon O, Le Scanf C, Steward M, Arnot DE (2005) Further definition of PfEMP-1 DBL-1alpha domains mediating rosetting adhesion of Plasmodium falciparum. Mol Biochem Parasitol 144: 109–113.
44. Vigan-Womas I, Guillotte M, Le Scanf C, Igonet S, Petres S, et al. (2008) An in vivo and in vitro model of Plasmodium falciparum rosetting and autoagglutination mediated by varO, a group A var gene encoding a frequent serotype. Infect Immun 76: 5565–5580.
45. Baruch DI, Ma XC, Singh HB, Bi X, Pasloske BL, et al. (1997) Identification of a region of PfEMP1 that mediates adherence of Plasmodium falciparum infected erythrocytes to CD36: conserved function with variant sequence. Blood 90: 3766–3775.
46. Gamain B, Smith JD, Miller LH, Baruch DI (2001) Modifications in the CD36 binding domain of the Plasmodium falciparum variant antigen are responsible for the inability of chondroitin sulfate A adherent parasites to bind CD36. Blood 97: 3268–3274.
47. Robinson BA, Welch TL, Smith JD (2003) Widespread functional specialization of Plasmodium falciparum erythrocyte membrane protein 1 family members to bind CD36 analysed across a parasite genome. Mol Microbiol 47: 1265–1278.
48. Klein MM, Gittis AG, Su HP, Makobongo MO, Moore JM, et al. (2008) The cysteine-rich interdomain region from the highly variable plasmodium falciparum erythrocyte membrane protein-1 exhibits a conserved structure. PLoS Pathog 4: e1000147.
49. Singh SK, Hora R, Belrhali H, Chitnis CE, Sharma A (2006) Structural basis for Duffy recognition by the malaria parasite Duffy-binding-like domain. Nature 439: 741–744.
50. Andersen P, Nielsen MA, Resende M, Rask TS, Dahlback M, et al. (2008) Structural insight into epitopes in the pregnancy-associated malaria protein VAR2CSA. PLoS Pathog 4: e42.
51. Van Dongen S (2000) A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May 2000.
52. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
53. Van Dongen S (2009) MCL.
54. Labeit S, Kolmerer B (1995) Titins: giant proteins in charge of muscle ultrastructure and elasticity. Science 270: 293–296.
55. Duan Y, DeKeyser JG, Damodaran S, Greaser ML (2006) Studies on titin PEVK peptides and their interaction. Arch Biochem Biophys 454: 16–25.
56. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281–288.
57. Trimnell AR, Kraemer SM, Mukherjee S, Phippard DJ, Janes JH, et al. (2006) Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. Mol Biochem Parasitol 148: 169–180.
58. Mu J, Awadalla P, Duan J, McGee KM, Joy DA, et al. (2005) Recombination hotspots and population structure in Plasmodium falciparum. PLoS Biol 3: e335.
59. Dahlback M, Rask TS, Andersen PH, Nielsen MA, Ndam NT, et al. (2006) Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in P. falciparum placental sequestration. PLoS Pathog 2: e124.
60. Durbin R (1998) Biological sequence analysis : probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press. xi, 357 p.
61. Tolia NH, Enemark EJ, Sim BK, Joshua-Tor L (2005) Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite Plasmodium falciparum. Cell 122: 183–193.
62. Howell DP, Samudrala R, Smith JD (2006) Disguising itself–insights into Plasmodium falciparum binding and immune evasion from the DBL crystal structure. Mol Biochem Parasitol 148: 1–9.
63. Mo M, Lee HC, Kotaka M, Niang M, Gao X, et al. (2008) The C-terminal segment of the cysteine-rich interdomain of Plasmodium falciparum erythrocyte membrane protein 1 determines CD36 binding and elicits antibodies that inhibit adhesion of parasite-infected erythrocytes. Infect Immun 76: 1837–1847.
64. Bull PC, Berriman M, Kyes S, Quail MA, Hall N, et al. (2005) Plasmodium falciparum variant surface antigen expression patterns during malaria. PLoS Pathog 1: e26.
65. Awadalla P (2003) The evolutionary genomics of pathogen recombination. Nat Rev Genet 4: 50–60.
66. Minin VN, Dorman KS, Fang F, Suchard MA (2007) Phylogenetic mapping of recombination hotspots in human immunodeficiency virus via spatially smoothed change-point processes. Genetics 175: 1773–1785.
67. Takagi J (2007) Structural basis for ligand recognition by integrins. Curr Opin Cell Biol 19: 557–564.
68. Nichols SA, Dirks W, Pearse JS, King N (2006) Early evolution of animal cell signaling and adhesion genes. Proc Natl Acad Sci U S A 103: 12451–12456.
69. Stewart PL, Nemerow GR (2007) Cell integrins: commonly used receptors for diverse viral pathogens. Trends Microbiol 15: 500–507.
70. Ouaissi MA (1988) Role of the RGD sequence in parasite adhesion to host cells. Parasitol Today 4: 169–173.
71. Nobbs AH, Shearer BH, Drobni M, Jepson MA, Jenkinson HF (2007) Adherence and internalization of Streptococcus gordonii by epithelial cells involves beta1 integrin recognition by SspA and SspB (antigen I/II family) polypeptides. Cell Microbiol 9: 65–83.
72. Hostetter MK (2000) RGD-mediated adhesion in fungal pathogens of humans, plants and insects. Curr Opin Microbiol 3: 344–348.
73. Lu X, Lu D, Scully MF, Kakkar VV (2006) Integrins in drug targeting-RGD templates in toxins. Curr Pharm Des 12: 2749–2769.
74. Siano JP, Grady KK, Millet P, Wick TM (1998) Short report: Plasmodium falciparum: cytoadherence to alpha(v)beta3 on human microvascular endothelial cells. Am J Trop Med Hyg 59: 77–79.
75. Ruoslahti E (1996) RGD and other recognition sequences for integrins. Annu Rev Cell Dev Biol 12: 697–715.
76. Calvete JJ, Marcinkiewicz C, Sanz L (2007) KTS and RTS-disintegrins: anti-angiogenic viper venom peptides specifically targeting the alpha 1 beta 1 integrin. Curr Pharm Des 13: 2853–2859.
77. Calvete JJ, Moreno-Murciano MP, Theakston RD, Kisiel DG, Marcinkiewicz C (2003) Snake venom disintegrins: novel dimeric disintegrins and structural diversification by disulphide bond engineering. Biochem J 372: 725–734.
78. Bray PG, Barrett MP, Ward SA, de Koning HP (2003) Pentamidine uptake and resistance in pathogenic protozoa: past, present and future. Trends Parasitol 19: 232–239.
79. Yipp BG, Robbins SM, Resek ME, Baruch DI, Looareesuwan S, et al. (2003) Src-family kinase signaling modulates the adhesion of Plasmodium falciparum on human microvascular endothelium under flow. Blood 101: 2850–2857.
80. Ho M, Hoang HL, Lee KM, Liu N, MacRae T, et al. (2005) Ectophosphorylation of CD36 regulates cytoadherence of Plasmodium falciparum to microvascular endothelium under flow conditions. Infect Immun 73: 8179–8187.

81. Marti M, Good RT, Rug M, Knuepfer E, Cowman AF (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 306: 1930–1933.
82. Resh MD (1999) Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins. Biochim Biophys Acta 1451: 1–16.
83. Gunaratne RS, Sajid M, Ling IT, Tripathi R, Pachebat JA, et al. (2000) Characterization of N-myristoyltransferase from Plasmodium falciparum. Biochem J 348 Pt 2: 459–463.
84. Resh MD (2006) Trafficking and signaling by fatty-acylated and prenylated proteins. Nat Chem Biol 2: 584–590.
85. Peitzsch RM, McLaughlin S (1993) Binding of acylated peptides and fatty acids to phospholipid vesicles: pertinence to myristoylated proteins. Biochemistry 32: 10436–10443.
86. Sigal CT, Zhou W, Buser CA, McLaughlin S, Resh MD (1994) Amino-terminal basic residues of Src mediate membrane binding through electrostatic interaction with acidic phospholipids. Proc Natl Acad Sci U S A 91: 12253–12257.
87. Nadolski MJ, Linder ME (2007) Protein lipidation. Febs J 274: 5202–5210.
88. Batistic O, Sorek N, Schultke S, Yalovsky S, Kudla J (2008) Dual fatty acyl modification determines the localization and plasma membrane targeting of CBL/CIPK Ca2+ signaling complexes in Arabidopsis. Plant Cell 20: 1346–1362.
89. Farazi TA, Waksman G, Gordon JI (2001) The biology and enzymology of protein N-myristoylation. J Biol Chem 276: 39501–39504.
90. Rahlfs S, Koncarevic S, Iozef R, Mailu BM, Savvides SN, et al. (2009) Myristoylated adenylate kinase-2 of Plasmodium falciparum forms a heterodimer with myristoyltransferase. Mol Biochem Parasitol 163: 77–84.
91. Russo I, Oksman A, Goldberg DE (2009) Fatty acid acylation regulates trafficking of the unusual Plasmodium falciparum calpain to the nucleolus. Mol Microbiol 72: 229–245.
92. Moskes C, Burghaus PA, Wernli B, Sauder U, Durrenberger M, et al. (2004) Export of Plasmodium falciparum calcium-dependent protein kinase 1 to the parasitophorous vacuole is dependent on three N-terminal membrane anchor motifs. Mol Microbiol 54: 676–691.
93. Rees-Channer RR, Martin SR, Green JL, Bowyer PW, Grainger M, et al. (2006) Dual acylation of the 45 kDa gliding-associated protein (GAP45) in Plasmodium falciparum merozoites. Mol Biochem Parasitol 149: 113–116.
94. Struck NS, de Souza Dias S, Langer C, Marti M, Pearce JA, et al. (2005) Re-defining the Golgi complex in Plasmodium falciparum using the novel Golgi marker PfGRASP. J Cell Sci 118: 5603–5613.
95. Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. J Mol Biol 317: 541–557.
96. Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2002) N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences. J Mol Biol 317: 523–540.
97. Struck NS, Herrmann S, Langer C, Krueger A, Foth BJ, et al. (2008) Plasmodium falciparum possesses two GRASP proteins that are differentially targeted to the Golgi complex via a higher- and lower-eukaryote-like mechanism. J Cell Sci 121: 2123–2129.
98. Boddey JA, Moritz RL, Simpson RJ, Cowman AF (2009) Role of the Plasmodium export element in trafficking parasite proteins to the infected erythrocyte. Traffic 10: 285–299.
99. Chang HH, Falick AM, Carlton PM, Sedat JW, DeRisi JL, et al. (2008) N-terminal processing of proteins exported by malaria parasites. Mol Biochem Parasitol 160: 107–115.
100. Utsumi T, Ohta H, Kayano Y, Sakurai N, Ozoe Y (2005) The N-terminus of B96Bom, a Bombyx mori G-protein-coupled receptor, is N-myristoylated and translocated across the membrane. Febs J 272: 472–481.
101. Denny PW, Gokool S, Russell DG, Field MC, Smith DF (2000) Acylation-dependent protein export in Leishmania. J Biol Chem 275: 11017–11025.
102. Ames JB, Tanaka T, Stryer L, Ikura M (1996) Portrait of a myristoyl switch protein. Curr Opin Struct Biol 6: 432–438.
103. Wu Y, Nelson MM, Quaile A, Xia D, Wastling JM, et al. (2009) Identification of phosphorylated proteins in erythrocytes infected by the human malaria parasite Plasmodium falciparum. Malar J 8: 105.
104. Hora R, Bridges DJ, Craig A, Sharma A (2009) Erythrocytic casein kinase II regulates cytoadherence of Plasmodium falciparum-infected red blood cells. J Biol Chem 284: 6260–6269.
105. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
106. Muller J, Creevey CJ, Thompson JD, Arendt D, Bork P (2010) AQUA: automated quality improvement for multiple sequence alignments. Bioinformatics 26: 263–265.
107. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
108. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059–3066.
109. Thompson JD, Thierry JC, Poch O (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 19: 1155–1161.
110. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O (2001) Towards a reliable objective function for multiple sequence alignments. J Mol Biol 314: 937–951.
111. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
112. Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57: 758–771.
113. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
114. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.
115. Felsenstein J (1989) Mathematics vs. Evolution: Mathematical Evolutionary Theory. Science 246: 941–942.
116. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
117. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
118. Margush TaM, FR (1981) Consensus n-trees. Bulletin of Mathematical Biology 43: 239–244.
119. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.
120. Smith TF, Waterman MS, Burks C (1985) The statistical distribution of nucleic acid similarities. Nucleic Acids Res 13: 645–656.
121. Sheneman L, Evans J, Foster JA (2006) Clearcut: a fast implementation of relaxed neighbor joining. Bioinformatics 22: 2823–2824.
122. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.
123. Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294: 1351–1362.

Appendix B

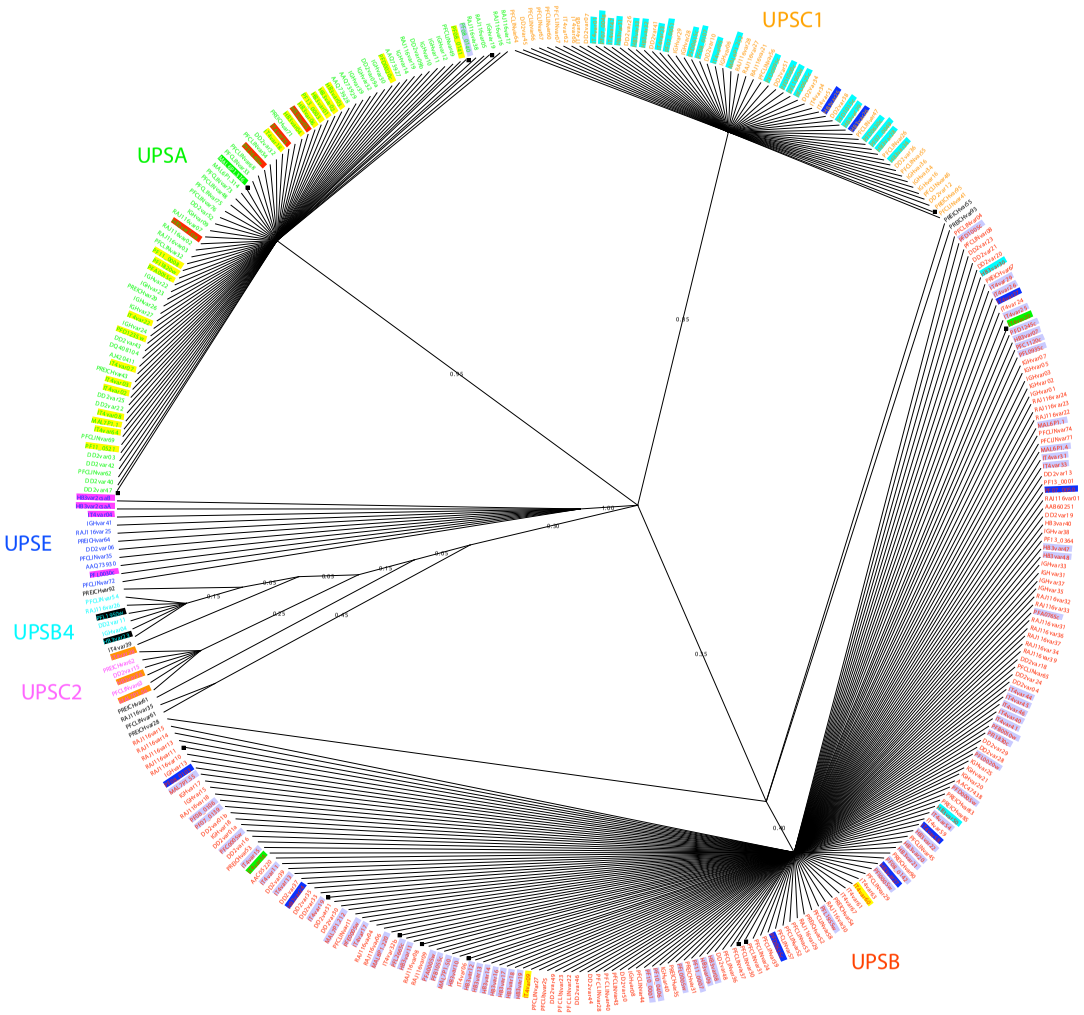
Supplementary material from paper I

This section contains supplementary figure S2N and S2O from paper I (appendix A). References to other figures and publications in the figure legends below refer to paper I. All the supplementary material for paper I is freely available online at the publishers webpage (doi: 10.1371/journal.pcbi.1000933) [108].

Supplementary figure S2N. Assignment of sequences to UPS groups by Markov clustering. The UPS groups were named as indicated by the text color. The background colors show the group membership assigned by Kraemer et al. 2007 [16]. Sequences found upstream of domain cassette 8 (Figure 3) are marked with black squares. The branch labels show the fraction of Markov clusters with this group present.

Supplementary figure S2O. Assignment of sequences to UPS groups by neighbor joining. The UPS groups were named as indicated by the text color. The background colors show the group membership assigned by Kraemer et al. 2007 [16]. Sequences found upstream of domain cassette 8 (Figure 3) are marked with black squares. The branch labels show the bootstrap values as fractions of 1000 bootstraps. Monophyletic subgroups with a bootstrap support above 0.7 and containing sequences from at least four different strains of P. falciparum are highlighted with thick red branches. Some subgroups were further expanded (without bootstrap support) to form larger monophyletic groups: UPSA2 and UPSB3 are expanded to include additional sequences annotated to UPSA2 and UPSB3 respec-

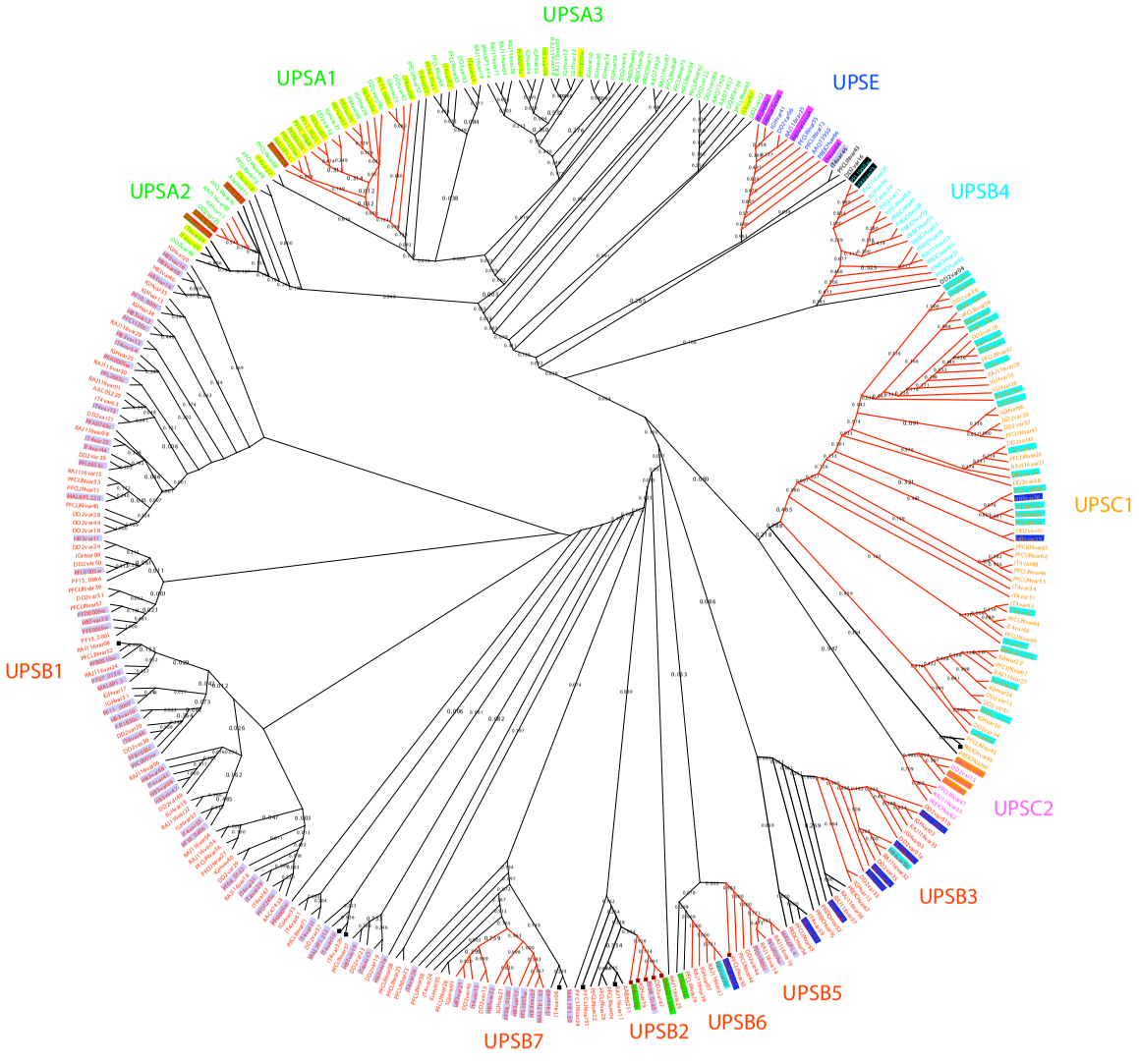
tively by Kraemer et al. 2007 [16], UPSB2 is expanded to include two genes with same domain architecture, and UPSC1 is expanded to include three sequences that fall between UPSC1 and UPSC2 but within the larger monophyletic group comprising all UPSC sequences. The sequences are shown with thick black branches. The additional sequences included by this expansion are denoted with an asterisk in the annotation in Figure S4 and S5. UPSA3 and UPSB1 are groups that contain all the sequences not assigned to any other subgroup in UPSA and UPSB respectively. ND: Not Determined.



Color legend

Text color: UPSA UPSB UPSB4 UPSC1 UPSC2 UPSE ND

Background color: UPSA1 UPSA2 UPSB1 UPSB2 UPSB3 UPSB4 UPSC1 UPSC2 UPSE



Color legend

Text color: UPSA UPSB UPSB4 UPSC1 UPSC2 UPSE ND

Background color: UPSA1 UPSA2 UPSB1 UPSB2 UPSB3 UPSB4 UPSC1 UPSC2 UPSE

Appendix C

Classification and diversity of *var* genes

var gene	Ups group	*var* gene	Ups group
DD2var09b	UpsA1	DQ408104	UpsA3
DD2var32	UpsA1	HB3var06	UpsA3
DD2var42	UpsA1	IGHvar10	UpsA3
HB3var01	UpsA1	IGHvar12	UpsA3
HB3var02	UpsA1	IGHvar14	UpsA3
HB3var03	UpsA1	IGHvar22	UpsA3
HB3var04	UpsA1	IGHvar24	UpsA3
HB3var05	UpsA1	IGHvar30	UpsA3
IGHvar09	UpsA1	IGHvar32	UpsA3
IGHvar23	UpsA1	IGHvar39	UpsA3
IGHvar26	UpsA1	IT4var02	UpsA3
IGHvar27	UpsA1	IT4var03	UpsA3
IT4var08	UpsA1	IT4var07	UpsA3
MAL7P1.1	UpsA1	IT4var18	UpsA3
PF11_0008	UpsA1	IT4var22	UpsA3
PF11_0521	UpsA1	IT4var64	UpsA3
PF13_0003	UpsA1	MAL6P1.314	UpsA3
PFD0020c	UpsA1	PF08_0141	UpsA3
PFD1235w	UpsA1	PFA0015c	UpsA3
DD2var25	UpsA2	PFCLINvar32	UpsA3
IGHvar11	UpsA2	PFCLINvar33	UpsA3
PFE1640w	UpsA2	PFCLINvar34	UpsA3
RAJ116var02	UpsA2	PFCLINvar48	UpsA3
DD2var40	UpsA2*	PFCLINvar49	UpsA3
HB3var1csa	UpsA2*	PFCLINvar62	UpsA3
IT4var35	UpsA2*	PFCLINvar68	UpsA3
PFCLINvar76	UpsA2*	PFCLINvar69	UpsA3
AAQ73927	UpsA3	PFCLINvar73	UpsA3
AAQ73928	UpsA3	PFCLINvar75	UpsA3
AAQ73929	UpsA3	PFI1820w	UpsA3
AJ420411	UpsA3	PREICHvar29	UpsA3
DD2var03	UpsA3	PREICHvar43	UpsA3
DD2var09a	UpsA3	PREICHvar71	UpsA3
DD2var22	UpsA3	RAJ116var03	UpsA3
DD2var43	UpsA3	RAJ116var05	UpsA3
DD2var52	UpsA3	RAJ116var07	UpsA3

Table C.1. The assignment of sequences upstream of *var* genes to Ups groups. ND: Not determined.

var gene	Ups group
RAJ116var16	UpsA3
RAJ116var17	UpsA3
RAJ116var19	UpsA3
RAJ116var38	UpsA3
AAC05220	UpsB1
AAC47438	UpsB1
DD2var04	UpsB1
DD2var18	UpsB1
DD2var19	UpsB1
DD2var20	UpsB1
DD2var21	UpsB1
DD2var23	UpsB1
DD2var24	UpsB1
DD2var28	UpsB1
DD2var29	UpsB1
DD2var30	UpsB1
DD2var31	UpsB1
DD2var37	UpsB1
DD2var39	UpsB1
DD2var48	UpsB1
DD2var49	UpsB1
DD2var50	UpsB1
HB3var08	UpsB1
HB3var09	UpsB1
HB3var10	UpsB1
HB3var11	UpsB1
HB3var12	UpsB1
HB3var13	UpsB1
HB3var14	UpsB1
HB3var16	UpsB1
HB3var18	UpsB1
HB3var19	UpsB1
HB3var20	UpsB1
HB3var40	UpsB1
HB3var47	UpsB1
HB3var48	UpsB1

var gene	Ups group
IGHvar05	UpsB1
IGHvar08	UpsB1
IGHvar13	UpsB1
IGHvar17	UpsB1
IGHvar18	UpsB1
IGHvar20	UpsB1
IGHvar25	UpsB1
IGHvar31	UpsB1
IGHvar33	UpsB1
IGHvar35	UpsB1
IGHvar37	UpsB1
IGHvar38	UpsB1
IGHvar40	UpsB1
IT4var06	UpsB1
IT4var11	UpsB1
IT4var13	UpsB1
IT4var16	UpsB1
IT4var17	UpsB1
IT4var19	UpsB1
IT4var24	UpsB1
IT4var25	UpsB1
IT4var26	UpsB1
IT4var29	UpsB1
IT4var31	UpsB1
IT4var32b	UpsB1
IT4var33	UpsB1
IT4var40	UpsB1
IT4var41	UpsB1
IT4var44	UpsB1
IT4var45	UpsB1
IT4var46	UpsB1
IT4var54	UpsB1
IT4var61	UpsB1
IT4var63	UpsB1
IT4var67	UpsB1
MAL6P1.1	UpsB1

Table C.2. The assignment of sequences upstream of *var* genes to Ups groups. ND: Not determined.

var gene	Ups group	*var* gene	Ups group
MAL7P1.212	UpsB1	PFI0005w	UpsB1
MAL7P1.50	UpsB1	PFI1830c	UpsB1
MAL8P1.220	UpsB1	PFL0005w	UpsB1
PF07_0139	UpsB1	PFL0935c	UpsB1
PF08_0142	UpsB1	PFL2665c	UpsB1
PF10_0001	UpsB1	RAJ116var01	UpsB1
PF10_0406	UpsB1	RAJ116var04	UpsB1
PF11_0007	UpsB1	RAJ116var06	UpsB1
PF13_0001	UpsB1	RAJ116var08	UpsB1
PF13_0364	UpsB1	RAJ116var09	UpsB1
PFA0005w	UpsB1	RAJ116var15	UpsB1
PFA0765c	UpsB1	RAJ116var18	UpsB1
PFB0010w	UpsB1	RAJ116var22	UpsB1
PFB1055c	UpsB1	RAJ116var23	UpsB1
PFC0005w	UpsB1	RAJ116var24	UpsB1
PFC1120c	UpsB1	RAJ116var29	UpsB1
PFCLINvar08	UpsB1	RAJ116var30	UpsB1
PFCLINvar11	UpsB1	RAJ116var34	UpsB1
PFCLINvar23	UpsB1	DD2var47	UpsB2
PFCLINvar24	UpsB1	IGHvar19	UpsB2
PFCLINvar25	UpsB1	IT4var20	UpsB2
PFCLINvar27	UpsB1	MAL6P1.316	UpsB2
PFCLINvar28	UpsB1	PF08_0140	UpsB2
PFCLINvar36	UpsB1	AAB60251	UpsB2*
PFCLINvar37	UpsB1	PFCLINvar04	UpsB2*
PFCLINvar39	UpsB1	PFCLINvar22	UpsB2*
PFCLINvar40	UpsB1	PFCLINvar29	UpsB2*
PFCLINvar52	UpsB1	PFCLINvar31	UpsB2*
PFCLINvar53	UpsB1	RAJ116var11	UpsB2*
PFCLINvar57	UpsB1	DD2var01a	UpsB3
PFCLINvar58	UpsB1	DD2var01b	UpsB3
PFCLINvar71	UpsB1	DD2var33	UpsB3
PFCLINvar74	UpsB1	DD2var35	UpsB3
PFD0005w	UpsB1	HB3var24	UpsB3
PFD1245c	UpsB1	HB3var50	UpsB3
PFE0005w	UpsB1	IGHvar02	UpsB3

Table C.3. The assignment of sequences upstream of *var* genes to Ups groups. ND: Not determined.

var gene	Ups group
IGHvar03	UpsB3
IGHvar15	UpsB3
IT4var27	UpsB3
PF07_0050	UpsB3
PFD0635c	UpsB3
RAJ116var32	UpsB3
RAJ116var33	UpsB3
HB3var27	UpsB3*
IT4var58	UpsB3*
IT4var59	UpsB3*
PFCLINvar65	UpsB3*
PREICHvar35	UpsB3*
PREICHvar52	UpsB3*
PREICHvar54	UpsB3*
PREICHvar67	UpsB3*
RAJ116var36	UpsB3*
RAJ116var37	UpsB3*
DD2var11	UpsB4
HB3var23	UpsB4
IGHvar04	UpsB4
PFCLINvar54	UpsB4
PFL1950w	UpsB4
PREICHvar31	UpsB4
PREICHvar53	UpsB4
PREICHvar83	UpsB4
PREICHvar85	UpsB4
PREICHvar92	UpsB4
RAJ116var13	UpsB4
RAJ116var26	UpsB4
DD2var44	UpsB5
MAL6P1.4	UpsB5
PFCLINvar30	UpsB5
PFCLINvar44	UpsB5
PFD1005c	UpsB5
PFL0020w	UpsB5
RAJ116var10	UpsB5
RAJ116var14	UpsB5
HB3var30	UpsB6
IGHvar07	UpsB6
PF08_0103	UpsB6
RAJ116var31	UpsB6
RAJ116var39	UpsB6
DD2var13	UpsB7
DD2var46	UpsB7
HB3var07	UpsB7
HB3var17	UpsB7
HB3var21	UpsB7
HB3var22	UpsB7
IGHvar01	UpsB7
IGHvar21	UpsB7
IT4var15	UpsB7
MAL7P1.55	UpsB7
PF08_0106	UpsB7
PFL1955w	UpsB7
DD2var07	UpsC1
DD2var10	UpsC1
DD2var12	UpsC1
DD2var26	UpsC1
DD2var34	UpsC1
DD2var36	UpsC1
DD2var38	UpsC1
DD2var41	UpsC1
DD2var45	UpsC1
DD2var51	UpsC1
HB3var25	UpsC1
HB3var26	UpsC1
HB3var28	UpsC1
HB3var29	UpsC1
HB3var31	UpsC1
HB3var32	UpsC1
HB3var33	UpsC1
HB3var34	UpsC1

Table C.4. The assignment of sequences upstream of *var* genes to Ups groups. ND: Not determined.

var gene	Ups group	*var* gene	Ups group
IGHvar06	UpsC1	PFD1015c	UpsC1
IGHvar16	UpsC1	PFL1960w	UpsC1
IGHvar28	UpsC1	RAJ116var21	UpsC1
IGHvar29	UpsC1	RAJ116var27	UpsC1
IGHvar34	UpsC1	RAJ116var28	UpsC1
IGHvar36	UpsC1	PREICHvar95	UpsC1*
IT4var01	UpsC1	DD2var15	UpsC2
IT4var05	UpsC1	HB3var36	UpsC2
IT4var23	UpsC1	IT4var28	UpsC2
IT4var34	UpsC1	MAL7P1.56	UpsC2
IT4var47	UpsC1	PFCLINvar61	UpsC2
IT4var51	UpsC1	PFCLINvar63	UpsC2
IT4var62	UpsC1	PREICHvar62	UpsC2
IT4var66	UpsC1	RAJ116var35	UpsC2
IT4var68	UpsC1	AAQ73930	UpsE
MAL6P1.252	UpsC1	DD2var06	UpsE
PF07_0048	UpsC1	HB3var2csaA	UpsE
PF07_0049	UpsC1	HB3var2csaB	UpsE
PF07_0051	UpsC1	IGHvar41	UpsE
PF08_0107	UpsC1	IT4var04	UpsE
PFCLINvar07	UpsC1	PFCLINvar35	UpsE
PFCLINvar26	UpsC1	PFCLINvar72	UpsE
PFCLINvar41	UpsC1	PFL0030c	UpsE
PFCLINvar46	UpsC1	PREICHvar64	UpsE
PFCLINvar47	UpsC1	RAJ116var25	UpsE
PFCLINvar55	UpsC1	DD2var16	ND
PFCLINvar56	UpsC1	IT4var09	ND
PFCLINvar60	UpsC1	IT4var39	ND
PFCLINvar64	UpsC1	IT4var60	ND
PFCLINvar66	UpsC1	PFCLINvar43	ND
PFCLINvar67	UpsC1	PFCLINvar45	ND
PFD0615c	UpsC1	PREICHvar28	ND
PFD0625c	UpsC1	PREICHvar55	ND
PFD0630c	UpsC1	PREICHvar61	ND
PFD0995c	UpsC1	PREICHvar90	ND
PFD1000c	UpsC1	PREICHvar93	ND

Table C.5. The assignment of sequences upstream of *var* genes to Ups groups. ND: Not determined.

Appendix D

Classification of MHC-binding peptides

c	n	k	$\mathcal{L}(\hat{\theta}\|data)$	AIC	AIC_c	w	w_c
1	4 050	171	−10 846.9	22 035.7	22 050.7	0.000000	0.000000
2	4 050	342	−10 441.7	21 567.5	21 630.5	0.000000	0.999999
3	4 050	513	−10 241.5	21 509.1	21 658.1	0.000000	0.000001
4	4 050	684	−10 046.3	21 460.7	21 738.7	0.999957	0.000000
5	4 050	855	−9 885.4	21 480.8	21 938.8	0.000043	0.000000
6	4 050	1 026	−9 753.4	21 558.7	22 255.7	0.000000	0.000000
7	4 050	1 197	−9 632.4	21 658.9	22 663.9	0.000000	0.000000
8	4 050	1 368	−9 575.5	21 887.0	23 284.0	0.000000	0.000000
9	4 050	1 539	−9 492.6	22 063.1	23 951.1	0.000000	0.000000
10	4 050	1 710	−9 385.9	22 191.9	24 692.9	0.000000	0.000000

Table D.1. Comparison of models and model probabilities for models with different number of clusters. The table shows the result of clustering the three alleles HLA-A*0101, HLA-A*0301 and HLA-B*4402 when the amino acid frequencies used to calculate the likelihood had been adjusted with both pseudocounts and sequence weighting. c: number of clusters in the model, n: sample size, k: estimable parameters, $\mathcal{L}(\hat{\theta}|data)$: likelihood, AIC: Akaike's information criterion, AIC_c: AIC corrected for small sample size, w: Akaike weight or model probability based on AIC, w_c Akaike weight or model probability based on AIC_c.

Data set	Alleles	AIC_c^1	$CAIC$	AIC_c^2	AIC_c^3	AIC_c^4	AIC^5	AIC_c^5
unbal	3	1	4	4	4	4	4	2
bal	1	1	1	1	1	1	1	1
bal	2	1	2	2	2	2	2	2
bal	3	1	3	3	3	3	3	2
bal	4	1	4	4	4	4	4	3
bal	5	1	4	4	4	4	5	4
nn 1	2	2	9	9	9	9	6	5
nn 2	3	3	9	9	9	9	7	5

Table D.2. The table shows the number of clusters suggested when using alternative ways to calculate Akaike's information criterion. unbal: the unbalanced data set with data from three alleles, bal: the balanced data sets using from one to five alleles, nn: the two data sets based on the neural networks with two or three alleles, AIC_c^1: small sample AIC using the number of sequences as the sample size rather than the number of amino acids (equation 5.4) [98], $CAIC$: Fujikoshi and Satoh's small sample AIC for multiple linear regression (equation 5.14) [104], AIC_c^2: Burnham and Anderson's hypothesized generalized multidimensional small sample AIC (equation 5.15) [98] using $v = 1$, AIC_c^3: same as AIC_c^2 but using $v = (p(p+1)/2 - 1)/2$, AIC_c^4: same as AIC_c^2 but using $v = p(p+1)/2$, AIC^5: Akaike's information criterion when the amino acid frequencies used to calculate the likelihood has been adjusted with both pseudocounts and sequence weighting (equation 5.3) [98], AIC_c^5: small sample AIC when the amino acid frequencies used to calculate the likelihood has been adjusted with both pseudocounts and sequence weighting (equation 5.4) [98].

Data set	Alleles	AIC_c^1	$CAIC$	AIC_c^2	AIC_c^3	AIC_c^4	AIC^5	AIC_c^5
sim 1	1	1	1	1	1	1	1	1
sim 2	1	1	1	1	1	1	1	1
sim 3	2	1	2	2	2	2	2	1
sim 4	2	1	2	2	2	2	2	1
sim 5	2	1	2	2	2	2	2	1
sim 6	2	1	2	2	2	2	2	1
sim 7	3	1	3	3	3	3	3	2
sim 8	3	1	2	2	2	2	3	2
sim 9	3	1	3	3	3	3	3	2
sim 10	3	1	3	3	3	3	3	2
sim 11	3	1	3	3	3	3	3	2
sim 12	3	1	3	3	3	3	3	2
sim 13	3	1	3	3	3	3	3	2
sim 14	3	1	3	3	3	3	3	2
sim 15	1	1	1	1	1	1	1	1
sim 16	1	1	1	1	1	1	1	1
sim 17	1	1	1	1	1	1	1	1
sim 18	1	1	1	1	1	1	1	1
sim 19	2	1	2	2	2	2	2	1
sim 20	2	1	2	2	2	2	2	1
sim 21	2	1	1	1	1	1	2	1
sim 22	2	1	1	1	1	1	2	1
sim 23	2	1	2	2	2	2	2	1
sim 24	2	1	2	2	2	2	2	1
sim 25	2	1	2	2	2	2	2	1
sim 26	2	1	2	2	2	2	2	1
sim 27	3	1	3	3	3	3	3	2
sim 28	3	1	3	3	3	3	3	2
sim 29	3	1	3	3	3	3	3	2
sim 30	3	1	2	2	2	2	3	2
sim 31	1	1	3	3	3	3	3	2
sim 32	1	1	3	3	3	3	3	2
sim 33	3	2	8	8	8	8	8	5
sim 34	3	2	10	10	10	10	9	6
sim 35	3	2	8	8	8	8	8	6
sim 36	3	2	10	10	10	10	8	6
sim 37	2	1	3	3	3	3	3	3
sim 38	2	1	3	3	3	3	4	3
sim 39	2	1	3	3	3	3	4	2
sim 40	2	1	3	3	3	3	3	3

Table D.3. The table shows the number of clusters suggested when using alternative ways to calculate Akaike's information criterion for the simulated data sets 1–40. Refer to table D.2 for an explanation of the different AIC's.

Data set	Alleles	AIC_c^1	$CAIC$	AIC_c^2	AIC_c^3	AIC_c^4	AIC^5	AIC_c^5
sim 101	2	1	2	2	2	2	2	2
sim 102	2	1	2	2	2	2	2	2
sim 103	2	1	2	2	2	2	2	2
sim 104	2	1	2	2	2	2	2	2
sim 105	2	1	2	2	2	2	2	2
sim 106	2	1	2	2	2	2	2	2
sim 107	2	2	4	4	4	4	4	2
sim 108	2	2	4	4	4	4	4	2
sim 109	2	2	5	5	5	5	5	2
sim 110	2	2	4	4	4	4	4	3
sim 111	2	2	5	5	5	5	4	2
sim 112	2	2	5	5	5	5	4	2
sim 113	2	1	3	3	3	3	3	2
sim 114	2	1	3	3	3	3	3	2
sim 115	2	1	3	3	3	3	3	2
sim 116	2	1	2	2	2	2	2	2
sim 117	2	1	3	4	4	3	3	2
sim 118	2	1	3	3	3	3	3	2

Table D.4. The table shows the number of clusters suggested when using alternative ways to calculate Akaike's information criterion for the simulated data sets 101–118. Refer to table D.2 for an explanation of the different AIC's.

Figure D.1. Sequence logo for simulated matrix 1. The logo was created from 10 000 peptides generated randomly from this matrix.

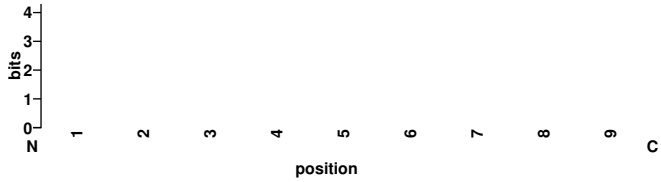


Figure D.2. Sequence logo for simulated matrix 2. The logo was created from 10 000 peptides generated randomly from this matrix.

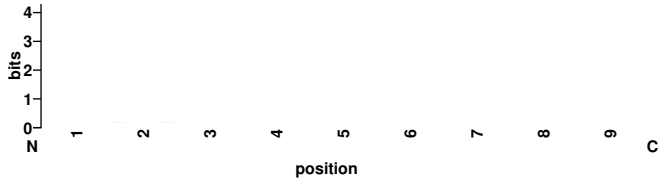


Figure D.3. Sequence logo for simulated matrix 3. The logo was created from 10 000 peptides generated randomly from this matrix.

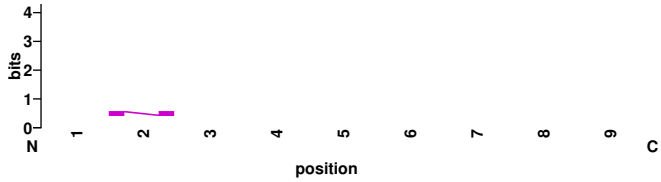


Figure D.4. Sequence logo for simulated matrix 4. The logo was created from 10 000 peptides generated randomly from this matrix.

Figure D.5. Sequence logo for simulated matrix 5. The logo was created from 10 000 peptides generated randomly from this matrix.

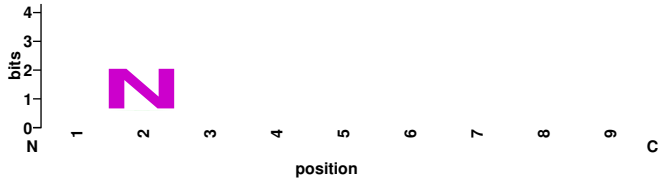


Figure D.6. Sequence logo for simulated matrix 6. The logo was created from 10 000 peptides generated randomly from this matrix.

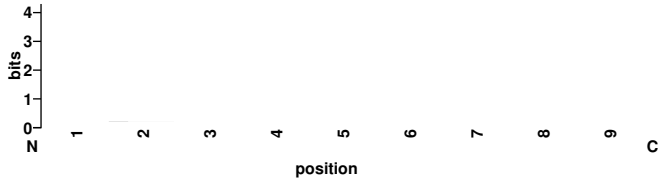


Figure D.7. Sequence logo for simulated matrix 7. The logo was created from 10 000 peptides generated randomly from this matrix.

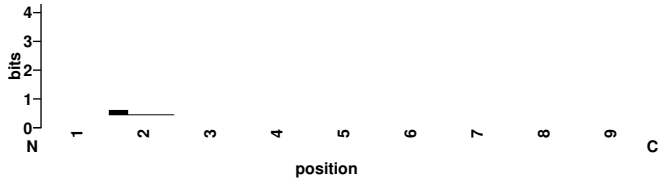


Figure D.8. Sequence logo for simulated matrix 8. The logo was created from 10 000 peptides generated randomly from this matrix.

Figure D.9. Sequence logo for simulated matrix 9. The logo was created from 10 000 peptides generated randomly from this matrix.

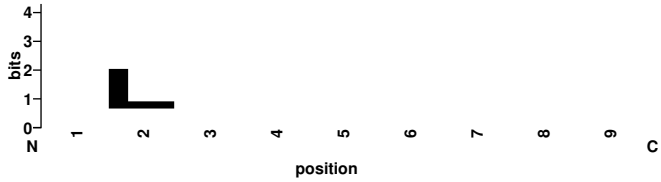


Figure D.10. Sequence logo for simulated matrix 10. The logo was created from 10 000 peptides generated randomly from this matrix.

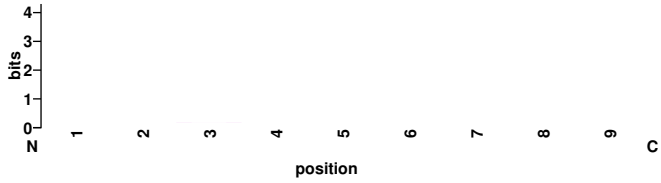


Figure D.11. Sequence logo for simulated matrix 11. The logo was created from 10 000 peptides generated randomly from this matrix.

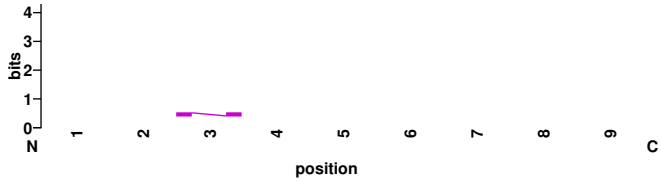


Figure D.12. Sequence logo for simulated matrix 12. The logo was created from 10 000 peptides generated randomly from this matrix.

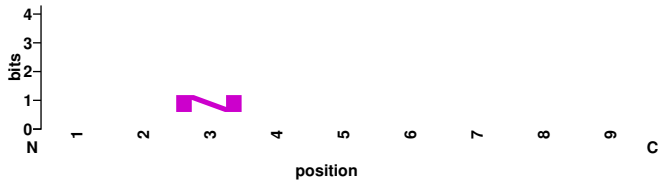


Figure D.13. Sequence logo for simulated matrix 13. The logo was created from 10 000 peptides generated randomly from this matrix.

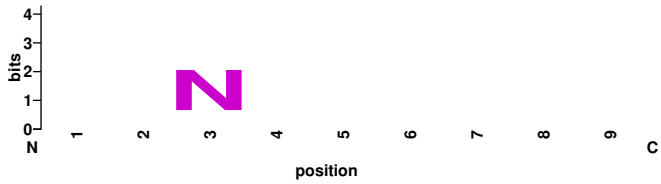


Figure D.14. Sequence logo for simulated matrix 14. The logo was created from 10 000 peptides generated randomly from this matrix.

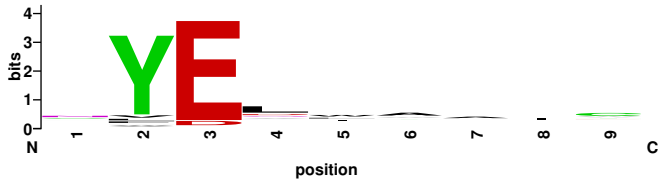


Figure D.15. Sequence logo for simulated matrix 101. The logo was created from 10 000 peptides generated randomly from this matrix.

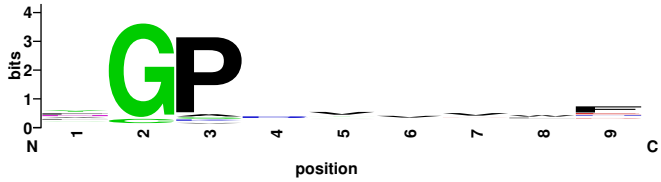


Figure D.16. Sequence logo for simulated matrix 102. The logo was created from 10 000 peptides generated randomly from this matrix.

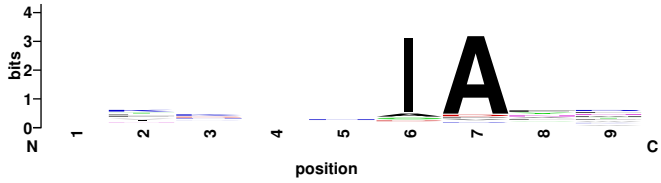


Figure D.17. Sequence logo for simulated matrix 103. The logo was created from 10 000 peptides generated randomly from this matrix.

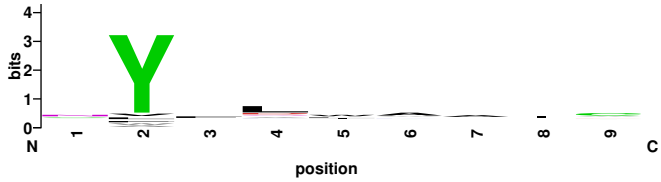


Figure D.18. Sequence logo for simulated matrix 104. The logo was created from 10 000 peptides generated randomly from this matrix.

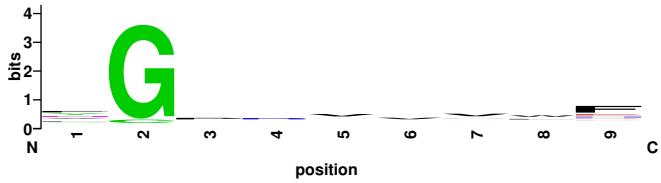


Figure D.19. Sequence logo for simulated matrix 105. The logo was created from 10 000 peptides generated randomly from this matrix.

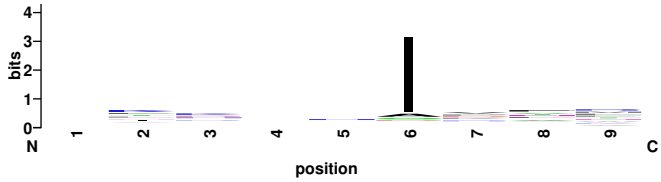


Figure D.20. Sequence logo for simulated matrix 106. The logo was created from 10 000 peptides generated randomly from this matrix.

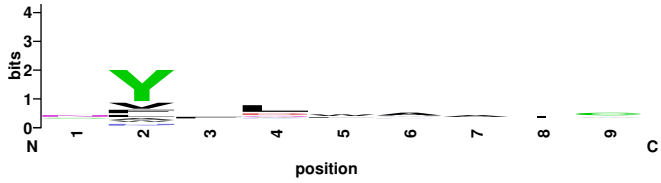


Figure D.21. Sequence logo for simulated matrix 107. The logo was created from 10 000 peptides generated randomly from this matrix.

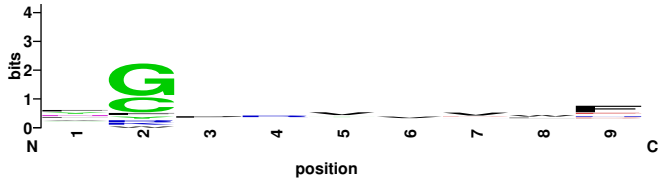


Figure D.22. Sequence logo for simulated matrix 108. The logo was created from 10 000 peptides generated randomly from this matrix.

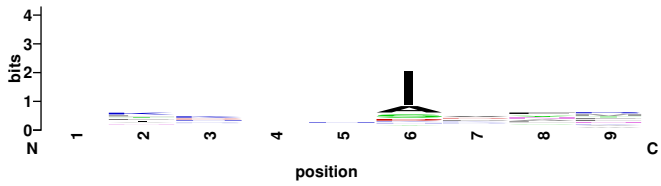


Figure D.23. Sequence logo for simulated matrix 109. The logo was created from 10 000 peptides generated randomly from this matrix.

Bibliography

1. Zheng H, Zhu HmM, Ning BfF, Li XyY (2006) [molecular identification of naturally acquired plasmodium knowlesi infection in a human case]. Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi 24: 273-6.

2. Zhu HmM, Li J, Zheng H (2006) [human natural infection of plasmodium knowlesi]. Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi 24: 70-1.

3. Berry A, Iriart X, Wilhelm N, Valentin A, Cassaing S, et al. (2011) Imported plasmodium knowlesi malaria in a french tourist returning from thailand. Am J Trop Med Hyg 84: 535-8.

4. Hoosen A, Shaw MTM (2011) Plasmodium knowlesi in a traveller returning to new zealand. Travel Med Infect Dis .

5. Barber BE, William T, Jikal M, Jilip J, Dhararaj P, et al. (2011) Plasmodium knowlesi malaria in children. Emerg Infect Dis 17: 814-820.

6. Lee KSS, Divis PCS, Zakaria SK, Matusop A, Julin RA, et al. (2011) Plasmodium knowlesi: Reservoir hosts and tracking the emergence in humans and macaques. PLoS Pathog 7: e1002015.

7. (2010) World Malaria Report 2010. World Health Organization.

8. Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415: 673-9.

9. Jones MK, Good MF (2006) Malaria parasites up close. Nature medicine : 170.

10. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite plasmodium falciparum. Nature 419: 498-511.

11. Lavstsen T, Salanti A, Jensen ATR, Arnot DE, Theander TG (2003) Sub-grouping of plasmodium falciparum 3d7 var genes based on sequence analysis of coding and non-coding regions. Malar J 2: 27.

12. Kraemer SM, Smith JD (2003) Evidence for the importance of genetic structuring to the structural and functional specialization of the plasmodium falciparum var gene family. Molecular Microbiology 50: 1527-1538.

13. Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, et al. (2007) Patterns of gene recombination shape var gene repertoires in plasmodium falciparum: comparisons of geographically diverse isolates. BMC Genomics 8: 45.

14. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, et al. (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.

15. Su XzZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, et al. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. Cell 82: 89 - 100.

16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-402.

17. Katoh K, Misawa K, Kuma Ki, Miyata T (2002) Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res 30: 3059-66.

18. Katoh K, Kuma Ki, Toh H, Miyata T (2005) Mafft version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-8.

19. Katoh K, Toh H (2008) Recent developments in the mafft multiple sequence alignment program. Brief Bioinform 9: 286-98.

20. Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using clustalw and clustalx. Curr Protoc Bioinformatics Chapter 2: Unit 2.3.

21. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal w and clustal x version 2.0. Bioinformatics 23: 2947-8.

22. Van Dongen S (2008) Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl 30: 121-141.

23. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic acids research 30: 1575.

24. Margush T, McMorris FR (1981) Consensus n-trees. Bulletin of Mathematical Biology 43: 239-244.

25. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.

26. Myers EW, Miller W (1988) Optimal alignments in linear space. Comput Appl Biosci 4: 11-7.

27. Liu W, Li Y, Learn GH, Rudicell RS, Robertson JD, et al. (2010) Origin of the human malaria parasite plasmodium falciparum in gorillas. Nature 467: 420-5.

28. Krief S, Escalante AA, Pacheco MA, Mugisha L, André C, et al. (2010) On the diversity of malaria parasites in african apes and the origin of plasmodium falciparum from bonobos. PLoS Pathog 6: e1000765.

29. Pollack Y, Katzen AL, Spira DT, Golenser J (1982) The genome of plasmodium falciparum. i: Dna base composition. Nucleic Acids Res 10: 539-46.

30. Weber JL (1987) Analysis of sequences from the extremely a + t-rich genome of plasmodium falciparum. Gene 52: 103-9.

31. Pollack Y, Kogan N, Golenser J (1991) Plasmodium falciparum: evidence for a dna methylation pattern. Exp Parasitol 72: 339-44.

32. Hattman S (2005) Dna-[adenine] methylation in lower eukaryotes. Biochemistry (Mosc) 70: 550-8.

33. Feng S, Cokus SJ, Zhang X, Chen PYY, Bostick M, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 107: 8689-94.

34. Pelizzola M, Ecker JR (2010). Febs lett. doi:10.1016/j.febslet.2010.10.061.

35. Suzuki MM, Bird A (2008) Dna methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9: 465-76.

36. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human dna methylomes at base resolution show widespread epigenomic differences. Nature 462: 315-22.

37. Ball MP, Li JB, Gao Y, Lee JHH, LeProust EM, et al. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 27: 361-8.

38. Bird AP (1980) Dna methylation and the frequency of cpg in animal dna. Nucleic Acids Res 8: 1499-504.

39. Schorderet DF, Gartler SM (1992) Analysis of cpg suppression in methylated and nonmethylated species. Proc Natl Acad Sci U S A 89: 957-61.

40. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, et al. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. Nat Genet 37: 853-62.

41. Khulan B, Thompson RF, Ye K, Fazzari MJ, Suzuki M, et al. (2006) Comparative isoschizomer profiling of cytosine methylation: the help assay. Genome Res 16: 1046-55.

42. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in arabidopsis. Cell 133: 523-36.

43. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, et al. (2008) Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. Nature 452: 215-9.

44. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic dna methylation. Science 328: 916-9.

45. Xiang H, Zhu J, Chen Q, Dai F, Li X, et al. (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol 28: 516-20.

46. FASTX Toolkit. URL `http://hannonlab.cshl.edu/fastx_toolkit/`.

47. Ewing B, Hillier LD, Wendl MC, Green P (1998) Base-calling of automated sequencer traces usingphred. i. accuracy assessment. Genome research 8: 175.

48. Ewing B, Green P (1998) Base-calling of automated sequencer traces usingphred. ii. error probabilities. Genome research 8: 186.

49. Burrows M, Wheeler DJ (1994) A block-sorting lossless data compression algorithm. Technical report, Palo Alto, CA: Digital Equipment Corporation.

50. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol 10: R25.

51. Krueger F, Andrews SR (2011) Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics .

52. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: pp. 289-300.

53. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, et al. (2009) Plasmodb: a functional genomic database for malaria parasites. Nucleic Acids Res 37: D539-43.

54. Petersen I, Eastman R, Lanzer M (2011) Drug-resistant malaria: Molecular mechanisms and implications for public health. FEBS Lett .

55. White NJ (2008) Qinghaosu (artemisinin): the price of success. Science 320: 330-4.

56. Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, et al. (2009) Artemisinin resistance in plasmodium falciparum malaria. New England Journal of Medicine 361: 455-467.

57. Anderson TJC, Nair S, Nkhoma S, Williams JT, Imwong M, et al. (2010) High heritability of malaria parasite clearance rate indicates a genetic basis for artemisinin resistance in western cambodia. J Infect Dis 201: 1326-30.

58. Llinás M, Deitsch KW, Voss TS (2008) Plasmodium gene regulation: far more to factor in. Trends Parasitol 24: 551-6.

59. Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of apicomplexa and their implication for the evolution of the ap2-integrase dna binding domains. Nucleic Acids Res 33: 3994-4006.

60. Boschet C, Gissot M, Briquet S, Hamid Z, Claudel-Renard C, et al. (2004) Characterization of pfmyb1 transcription factor during erythrocytic development of 3d7 and f12 plasmodium falciparum clones. Mol Biochem Parasitol 138: 159-63.

61. De Silva EK, Gehrke AR, Olszewski K, León I, Chahal JS, et al. (2008) Specific dna-binding by apicomplexan ap2 transcription factors. Proc Natl Acad Sci U S A 105: 8393-8.

62. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, et al. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science 290: 2105-10.

63. Lindner SE, De Silva EK, Keck JL, Llinás M (2009) Structural determinants of dna binding by a p. falciparum apiap2 transcriptional regulator. J Mol Biol .

64. Voss TS, Kaestli M, Vogel D, Bopp S, Beck HPP (2003) Identification of nuclear proteins that interact differentially with plasmodium falciparum var gene promoters. Mol Microbiol 48: 1593-607.

65. Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28: 337-50.

66. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, et al. (2005) A protein interaction network of the malaria parasite plasmodium falciparum. Nature 438: 103-7.

67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic acids research 28: 235.

68. Gamo FJJ, Sanz LM, Vidal J, de Cozar C, Alvarez E, et al. (2010) Thousands of chemical starting points for antimalarial lead identification. Nature 465: 305-10.

69. Knox C, Law V, Jewison T, Liu P, Ly S, et al. (2011) Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39: D1035-41.

70. Goodsell DS, Morris GM (1998) Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 19: 1639-1662.

71. Bréhélin L, Florent I, Gascuel O, Maréchal E (2010) Assessing functional annotation transfers with inter-species conserved coexpression: application to plasmodium falciparum. BMC Genomics 11: 35.

72. de Koning-Ward TF, Gilson PR, Boddey JA, Rug M, Smith BJ, et al. (2009) A newly discovered protein export machine in malaria parasites. Nature 459: 945-9.

73. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The string database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561-8.

74. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, et al. (2010) Mint, the molecular interaction database: 2009 update. Nucleic Acids Res 38: D532-9.

75. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) Pubchem: integrated platform of small molecules and biological activities. Annual Reports in Computational Chemistry 4: 217-241.

76. Read JA, Wilkinson KW, Tranter R, Sessions RB, Brady RL (1999) Chloroquine binds in the cofactor binding site of plasmodium falciparum lactate dehydrogenase. J Biol Chem 274: 10213-8.

77. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S (1999) Syfpeithi: database for mhc ligands and peptide motifs. Immunogenetics 50: 213-9.

78. Brusic V, Rudy G, Harrison LC (1998) Mhcpep, a database of mhc-binding peptides: update 1997. Nucleic Acids Res 26: 368-71.

79. Lundegaard C, Lund O, Buus S, Nielsen M (2010) Major histocompatibility complex class i binding predictions as a tool in epitope discovery. Immunology 130: 309-18.

80. Liao WWP, Arthur JW (2011) Predicting peptide binding to major histocompatibility complex molecules. Autoimmun Rev .

81. Kondo A, Sidney J, Southwood S, del Guercio MF, Appella E, et al. (1997) Two distinct hla-a*0101-specific submotifs illustrate alternative peptide binding modes. Immunogenetics 45: 249-58.

82. Kubo RT, Sette A, Grey HM, Appella E, Sakaguchi K, et al. (1994) Definition of specific peptide motifs for four major hla-a alleles. J Immunol 152: 3913-24.

83. DiBrino M, Parker KC, Shiloach J, Turner RV, Tsuchida T, et al. (1994) Endogenous peptides with distinct amino acid anchor residue motifs bind to hla-a1 and hla-b8. J Immunol 152: 620-31.

84. Geluk A, van Meijgaarden KE, Southwood S, Oseroff C, Drijfhout JW, et al. (1994) Hla-dr3 molecules can bind peptides carrying two alternative specific submotifs. J Immunol 152: 5742-8.

85. Sidney J, Oseroff C, Southwood S, Wall M, Ishioka G, et al. (1992) Drb1*0301 molecules recognize a structural motif distinct from the one recognized by most dr beta 1 alleles. J Immunol 149: 2634-40.

86. Geluk A, Van Meijgaarden KE, Janson AA, Drijfhout JW, Meloen RH, et al. (1992) Functional analysis of dr17(dr3)-restricted mycobacterial t cell epitopes reveals dr17-binding motif and enables the design of allele-specific competitor peptides. J Immunol 149: 2864-71.

87. Chicz RM, Urban RG, Gorga JC, Vignali DA, Lane WS, et al. (1993) Specificity and promiscuity among naturally processed peptides bound to hla-dr alleles. J Exp Med 178: 27-47.

88. Burgdorf S, Kurts C (2008) Endocytosis mechanisms and the cell biology of antigen presentation. Curr Opin Immunol 20: 89-95.

89. van den Hoorn T, Paul P, Jongsma MLM, Neefjes J (2011) Routes to manipulate mhc class ii antigen presentation. Curr Opin Immunol 23: 88-95.

90. Yewdell JW, Bennink JR (1999) Immunodominance in major histocompatibility complex class i-restricted t lymphocyte responses. Annu Rev Immunol 17: 51-88.

91. Rao X, Costa AICAF, van Baarle D, Kesmir C (2009) A comparative study of hla binding affinity and ligand diversity: implications for generating immunodominant cd8+ t cell responses. J Immunol 182: 1526-32.

92. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, et al. (2004) Improved prediction of mhc class i and class ii epitopes using a novel gibbs sampling approach. Bioinformatics 20: 1388-97.

93. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, et al. (1993) Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. Science 262: 208-14.

94. Tatusov RL, Altschul SF, Koonin EV (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A 91: 12091-5.

95. Hobohm U, Scharf M, Schneider R, Sander C (1992) Selection of representative protein data sets. Protein Science 1: 409-417.

96. Henikoff S, Henikoff JG (1994) Position-based sequence weights. J Mol Biol 243: 574-8.

97. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097-100.

98. Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. springer-verlag. New York, New York, USA .

99. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors, Second International Symposium on Information Theory. Budapest: Akademiai Kiado, volume Second edi, pp. 267–281.

100. Hurvich CM, Tsai CLL (1989) Regression and time series model selection in small samples. Biometrika 76: 297-307.

101. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, et al. (2010) The immune epitope database 2.0. Nucleic Acids Res 38: D854-62.

102. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89: 10915-9.

103. Crooks GE, Hon G, Chandonia JMM, Brenner SE (2004) Weblogo: a sequence logo generator. Genome Res 14: 1188-90.

104. Fujikoshi Y, Satoh K (1997) Modified aic and cp in multivariate linear regression. Biometrika 84: 707.

105. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, et al. (2003) Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. Protein Sci 12: 1007-17.

106. R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

107. Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53: 793-808.

108. Rask TS, Hansen DA, Theander TG, Pedersen AG, Lavstsen T (2010) Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes–divide and conquer. PLoS Comput Biol 6: e1000933.