Inferring malaria parasite population structure from serological networks

Caroline O. Buckee1,2,3,*, Peter C. Bull2,4 and Sunetra Gupta1

1Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, UK
2Wellcome Collaborative Research Program, KEMRI, Kilifi 80108, Kenya
3Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
4Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, University of Oxford, CCVTM, Oxford OX3 7LJ, UK

Keywords: pathogen diversity; Plasmodium falciparum; serology; network

1. INTRODUCTION

Individuals living in malaria-endemic areas never develop complete immunity to infection with the highly diverse malaria parasite, Plasmodium falciparum, although they do acquire protection against clinical malaria after a certain period of exposure and appear to mount a protective response against the more severe forms of the disease after relatively few infections (Gupta et al. 1999). The acquisition of clinical, but non-sterile, immunity is still poorly understood; however, it seems to rely on cumulative experience of infection by many antigenically different parasite isolates. An understanding of the antigenic structuring of the parasite population is therefore vital for the design of vaccination and other control programmes, and although large amounts of sequence data from P. falciparum antigen genes are currently being generated, a link between genotype and phenotype remains elusive. Here, we analyse serological data from Kenya using a novel network technique to deconstruct the relationships between patients’ immune responses to different parasite isolates. We show that particular population structures and expression patterns produce distinctive signatures within serological networks of parasite recognition, which can be used to discriminate between competing hypotheses regarding the organization of these genes. Our analysis suggests that different levels of immune selection occur within different groups of the same multigene family leading to mixed population structures.

The malaria parasite Plasmodium falciparum is characterized by high levels of genetic diversity at antigenic loci involved in virulence and immune evasion. Knowledge of the population structure and dynamics of these genes is important for designing control programmes and understanding the acquisition of immunity to malaria; however, high rates of homologous and non-homologous recombination as well as complex patterns of expression within hosts have hindered attempts to elucidate these structures experimentally. Here, we analyse serological data from Kenya using a novel network technique to deconstruct the relationships between patients’ immune responses to different parasite isolates. We show that particular population structures and expression patterns produce distinctive signatures within serological networks of parasite recognition, which can be used to discriminate between competing hypotheses regarding the organization of these genes. Our analysis suggests that different levels of immune selection occur within different groups of the same multigene family leading to mixed population structures.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2008.1122 or via http://journals.royalsociety.org.

* Author and address for correspondence: Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA (caroline.buckee@zoo.ox.ac.uk).

Received 13 August 2008
Accepted 11 September 2008

Published online 30 September 2008


This journal is © 2008 The Royal Society
population is composed of discretely circulating strains of varying prevalence and virulence explains several features of malaria epidemiology, including the rapid acquisition of protection against severe disease. This 'discrete strain' hypothesis has proved contentious, however, and competing hypotheses range from those suggesting that var genes are completely randomly distributed, to those arguing that significant but variable overlap exists between parasite var repertoires (Ghia et al. 1999; Chattopadhyay et al. 2003). Unfortunately, testing these ideas remains extremely challenging for several technical reasons. First of all, the locations of antigenic epitopes within the majority of var genes remain elusive (although see Dahlbäck et al. (2006) and Andersen et al. (2008) for bioinformatic analysis of epitope regions within the conserved var gene associated with placental binding in pregnant women).

Furthermore, var diversity makes the design of universal primers problematic, so most sequencing projects focus on the analysis of small sequence fragments rather than whole genes. It is important to note here that simply observing some level of sequence conservation within certain regions of particular var gene families does not contradict the discrete strain hypothesis. Understanding the results of these projects is also hampered by the complex and variable expression patterns of var genes during infection and in vitro. Serological data, by contrast, provide direct information about relationships between the expression of antigenic epitopes and host responses. Here, we attempt to develop a set of tools to dissect these data with a view to extending these methods to eventually allow us to link the genetic structure of var sequence fragments to patterns of parasite recognition among patients from endemic regions.

Numerous studies comparing the antibody responses of hosts to their own (homologous) and others' (heterologous) parasites have addressed the question of population structure at the serological level by exploring levels of cross-reactivity between isolates (Marsh & Howard 1986; Forsyth et al. 1989; Iqbal et al. 1993; Chattopadhyay et al. 2003). The patterns of parasite recognition generated by these comparisons are generally presented in the form of a chequerboard of agglutination assays, measuring the difference in antibody titre at the time of acute disease and during convalescence (see appendix A). The chequerboard from the largest study of this kind, involving Kenyan children (Bull et al. 1999), is shown in figure 1a. Serological studies of this kind have concluded that patients tend to generate antibodies primarily to the VSAs of their own (homologous) parasite, although some cross-reactive responses have been noted (Bull et al. 1999; Chattopadhyay et al. 2003) and certain parasites are particularly well recognized by heterologous sera (Bull et al. 1999, 2000; Nielsen et al. 2004). These observations of both isolate-specific and cross-reactive immune responses have not been satisfactorily linked to the competing hypotheses about parasite population structure described above, however, partly because the relationships between cross-reactive responses are difficult to examine quantitatively using the chequerboard representation of the data.

In order to analyse these patterns of parasite recognition within the Kenyan data described above (Bull et al. 1999), we use a novel network approach rather than the traditional chequerboard framework. Here, each patient and corresponding parasite isolate is represented as one node within a network, with positive agglutination scores represented as a directed edge between nodes. Edges directed into a node correspond to recognition of the parasite, whereas edges directed out of a node indicate an agglutinating antibody response by the patient (see appendix A for more details). Figure 1b shows networks corresponding to the chequerboards in figure 1a. Note that the network showing increasing antibody titre, which we assume is generated by the current infection and will hereafter call the 'response' network, is simply the difference between the convalescent and acute responses. We use the structure of these networks, in conjunction with simple stochastic models, to illustrate the idea that serological data contain indirect information about the expression patterns and population structure of P. falciparum var gene repertoires. We provide evidence for a hierarchy of infection with different parasites among Kenyan children, and propose a var repertoire structure that represents an intermediate between a random distribution and the discrete strain hypothesis described above.

Figure 1. (a) A chequerboard of agglutination assays from patients in Kenya. Serological responses of patients (on the y-axis) to different parasite isolates (on the x-axis). The parasites are presented in the same order as the patients, such that the diagonal of the matrix corresponds to homologous patient–parasite pairs. (i) The acute responses to parasites (the parasite and serum are isolated upon admission to hospital), and (ii) the responses 3 weeks later. The darkness of the boxes corresponds with the strength of the immune response, in terms of the frequency and size of agglutinates formed on a semi-quantitative scale from 0 to 4 (see Bull et al. (1999) for details). (b) Networks corresponding to the agglutination profiles shown in (a). Note that these are (i) acute and (ii) increasing responses rather than acute and convalescent, since we are interested in increasing antibody titres. Each node corresponds to a patient–parasite pair. Positive recognition of parasites is represented by arrows from the patient towards a parasite. We have not included homologous responses.

2. STRUCTURAL CHARACTERISTICS OF KENYAN SEROLOGICAL NETWORKS

To examine the characteristics of the Kenyan serological networks (figure 1b), various metrics of network structure such as degree distribution (the number of edges directed...
3. EVIDENCE FOR AN INFECTION HIERARCHY

It has previously been suggested, based on the negative correlation between a host’s age and the frequency of recognition of his or her parasite, that there may be a hierarchy of infection with different parasite types (Bull et al. 1999). To determine what effects such an infection hierarchy would have on the expected structure of an acute serological network (reflecting previous exposure rather than current infections), simulations were performed in which hypothetical networks were generated assuming either a random or ordered infection process. Infection histories, which can also be thought of as host antibody repertoires, were generated for each of 100 hosts assuming varying levels of exposure. Parasites were assumed to be discrete antigenic types, for simplicity, and infection histories were either chosen randomly from a circulating pool of types or built up with a pre-defined order. Each infecting parasite was then compared with the infection history (or antibody repertoire) of each host, to create a hypothetical acute serological network. Ten thousand simulations were run for each of the random and ordered models, and compared with randomly rewired networks of the same density.

The ordered infection model yielded different network structures from random expectations, whereas the random infection model did not. The ordered model produced networks with extremely low reciprocity, high transitivity, high variance in $k_{in}$ and $k_{out}$, and a significantly higher...
Table 2. A comparison of structural features of the model results and the observed networks. (For each of 5 metrics, the directions of the difference between model results or real networks and 10,000 randomly rewired networks of the same density are shown. Boxes with arrows are significantly different from randomly rewired networks in the direction shown. See electronic supplementary material for more results.)

<table>
<thead>
<tr>
<th>metric</th>
<th>reciprocity</th>
<th>transitivity</th>
<th>$var\ k_{in}$</th>
<th>$var\ k_{out}$</th>
<th>ffd triads</th>
</tr>
</thead>
<tbody>
<tr>
<td>random model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ordered model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kenyan acute</td>
<td></td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>Kenyan increasing</td>
<td>↑</td>
<td></td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. The relative importance of different structural triads in the serological networks. The SP score is calculated as in Milo et al. (2002, 2004), which measures the relative importance of different triad types. Here, an SP score above 0.5 represents a significant departure from random expectations. The key shows which triads are referred to. Blue bars, acute; yellow bars, response.

density of ‘feed-forward’ loops than expected by chance. Table 2 illustrates the difference between the randomly rewired networks and those simulated under the two different infection processes, as well as the striking structural similarities between the networks generated by the ordered model of infection and the acute network. Intuitively, one would expect these characteristics from a hierarchical infection process. As an example, a patient infected by an isolate ‘higher up’ (for example isolate A) in the hierarchy would recognize one ‘lower down’ (isolate B), but this would not be reciprocated (hence low reciprocity). In addition, the patient infected by isolate A would recognize all isolates below B (e.g. isolate C); the patient infected by isolate B would therefore also be expected to recognize isolate C. Thus, the feed-forward loop $A \rightarrow B \rightarrow C$ and $A \rightarrow C$ is found at high densities and transitivity is high.

The structural characteristics of the acute network are therefore consistent with a hierarchical infection process occurring among these patients, as previously hypothesized (Bull et al. 2000). Thus, the immune system of the host (correlating with age in this case) seems to select for the expression of particular antigenic determinants, and children encounter serological variants in a non-random order as they grow up in areas such as Kenya.

4. THE EFFECTS OF POPULATION STRUCTURE, INFECTION LENGTH AND VAR EXPRESSION PATTERNS ON RESPONSE NETWORK STRUCTURE

We were surprised to find that the response network also exhibited a distinctive hierarchical source–sink structure (see table 2). Unlike the acute networks, edges within response networks are not related to previous infections but rather indicate shared antigenic determinants between currently infecting parasites from different patients. Since the expressed $var$ sequences at the time of parasite sampling will be dominated by only the most recently expressed $vars$, the frequency of heterologous immune responses (i.e. those of one child towards the parasites from another child) will depend on (i) the extent of $var$ repertoire overlap between isolates and (ii) the order in which repertoires are expressed. Simple models were used to generate hypothetical response networks, and compared with the observed network in order to test three hypotheses concerning $var$ repertoire structure and expression patterns. Hosts were infected with parasites defined by 60 $var$ genes, and assigned an infection length corresponding to the proportion of their parasites’ $var$ repertoire that they experienced during infection. Networks were generated by comparing each host’s isolate, which was assumed to be expressing only the last gene in the sequence of $var$ expressed, to every other host’s antibody repertoire generated during their current infection. Figure 3 shows a schematic of the model set-up for one such comparison.

Three different simplified population structures were examined, as well as the effects of random or ordered expression of different $var$ genes. The population structures modelled were as follows: $var$ repertoires could be (i) discrete with non-overlapping combinations of $var$ genes, in accordance with the theoretical models of Gupta et al. (1996, 1998) described above; (ii) randomly drawn from a global pool of circulating $var$ genes, assuming repertoires have been randomized by recombination; or (iii) an intermediate structure combining both these hypotheses, in which each parasite genotype contains a few ‘common’ $var$ genes that are expressed early in infection (Lavstsen et al. 2005) and are discretely structured due to strong immune selection, and many ‘rare’ $var$ genes drawn randomly from a global pool (this hypothesis is based on ideas generated by Bull et al. (1999, 2000) see electronic supplementary material for further explanation). For each population, it was then assumed that $var$ gene expression occurred either in a particular order or randomly (see appendix A for more details).

To find the models that best approximated the Kenyan response network, we measured network density, transitivity and reciprocity, the number of components formed, and the variance in $k_{in}$ and $k_{out}$, since these seemed to define the distinctive structure of the data. By comparing the effects of highly simplified hypothetical population structures on expected network structures, we hoped to
gain insights into the population structures leading to the observed serological data.

5. SUPPORT FOR A PARTITIONED VAR REPETTOIRE STRUCTURE

Figure 4 shows the performance of the three population models; for each structural characteristic, the heights of the bars indicate the parasite diversity at which the model coincides with the observed response network. Absent bars indicate that the model population never reached the value of the parameter observed in the data. Notice that these results are for the best fit parameters for the models in each case (see electronic supplementary material for more comparisons between populations). For the discrete population model, realistic densities and levels of variance in $k_{in}$ and $k_{out}$ were reached when the parasite population was made up of only seven non-overlapping 'strains', although the reciprocity, transitivity and number of components for this population were all extremely high. For populations in which var repertoires were drawn randomly from a global pool, only the reciprocity and variance in $k_{in}$ reached levels comparable to the data, although this occurred at very different var pool sizes (1209 and 532, respectively). The network densities were far below those of the data, and none of the other metrics came close to the values observed. For the best fit intermediate model, however (see Figure 4), all six structural characteristics measured coincided with the data when parasite var repertoires were defined by a discrete set of vars chosen from one of seven non-overlapping strains, and a larger set of vars drawn from a global pool of between 260 and 591. The intermediate model also showed features characteristic of the real network over a much larger parameter range than the other two models.

For all three populations, the models most compatible with the data were those with ordered rather than random var expression. Without ordered var expression, no source–sink structure occurred and transitivity remained low for all three populations—thus some of the features that the response network shared with the hierarchical
infection model described in §4 are in this case due to hierarchical expression patterns. It is important to note that variable infection lengths are also important in generating the structures observed in the real network (see electronic supplementary material).

The differences in the compatibility of the model population structures with the data can be clearly observed by visualizing the networks in figure 5. For the discrete population, as the diversity of the pathogen population increased, the networks fragmented into different components corresponding to clusters of hosts infected by the same strain (figure 5b). Reciprocity remained very high as the number of strains increased, and although the source–sink structure was apparent when var expression was ordered, it was weak due to the different components. The random population model showed the same fragmentation as the global var pool increased, and there was no source–sink structure observed (figure 5a). Densities fell rapidly below that of the observed network—figure 5a shows the densest network for this population, for example (assuming a var pool of only 100 variants). Of the three population structures, only the intermediate model showed distinctive source–sink structure, low reciprocity, low numbers of cliques and high transitivity. This occurred because many parasites shared the common var genes expressed at the beginning of infection (assuming ordered var expression), but not those expressed later on, which were randomly drawn from a large global pool. Thus, the isolates taken from those patients with very short infection lengths were recognized by almost all other patients in a non-reciprocal fashion. As seen in figure 5c, this led to a pronounced source–sink structure that was focused on these isolates. This finding is consistent with the Kenyan data and provides an intermediate hypothesis between the strain theory of malaria transmission and the arguments that var repertoires are randomized. These model population structures are highly artificial, of course, but they do indicate the types of dynamics and population structures that are expected to lead to different types of networks.

6. CONCLUSIONS
We have shown that serological networks can be used to help understand the population structures and infection dynamics of malaria infections. Given the difficulties inherent to directly identifying strains of *P. falciparum*, and our lack of understanding of within-host dynamics of important antigenic determinants such as the var genes, we believe that analysing serological data in this way offers a useful approach to inferring these structures. Our analysis of the Kenyan data suggests a hierarchical structure not only in terms of the infection dynamics of different parasites (the acute network) but also in terms of var gene ordering and expression within genotypes (the response network).

We propose that there are different levels of immune selection occurring on different groups of var genes, which could lead to the intermediate var repertoire structure described above. Here, relatively common var will be structured into discrete, non-overlapping combinations by immune selection, since they are expressed first and by many isolates, and only these restricted combinations will be observed. This hypothesis is compatible with the observation that the group A var family appears to be relatively restricted, and it will be testable once the epitope regions of these var have been elucidated (Jensen et al. 2004). The relative restriction of diversity among the common genes may also be explained by functional constraints with respect to binding particular host receptors—it has been suggested that commonly recognized var genes are optimized for rapid growth among non-immune hosts (Bull et al. 1999), also explaining the apparent association of specific subsets of var with severe disease (Rottmann et al. 2006; Kyriacou et al. 2007). Under our hypothesis, the remaining var genes within each genome are not under the same immune selection pressure and therefore do not show the same discrete structure, and will be relatively randomly distributed between genotypes. We conjecture that these var genes may also be less functionally constrained, accounting for their diversity.

Bull et al. (2008) have generated networks of var gene fragments, in which each gene fragment is a node and edges connecting genes are exact sequence matches (see figure 6). This network of var sequences forms two main clusters characterized by dense within-cluster links and loose between-cluster connections. The smaller lobe appears to be primarily made up of Group A var genes, which form a dense ‘polymorphic block-sharing group’, whereas the larger lobe appears to be composed of both B and C var (based on the classification system of Lavstsen et al. (2003) and Kraemer et al. (2007)). The recombination hierarchy evidenced by this structure is consistent with our models. We suggest that if the subset of var in the group A block-sharing group is expressed often and is
highly immunogenic, immune selection may act to structure it into discrete strains. By contrast, the majority of var genes in each repertoire (belonging to groups B and C) may not be expressed to the same extent and would not be under the same level of immune selection. The effects of this type of partitioned var repertoire on the overall level of var diversity and the development of immunity to disease have recently been explored theoretically (Recker et al. 2008). These models showed that the partitioning of var repertoires not only limited the level of diversity achievable through recombination, but could also explain the rapid acquisition of immunity to severe disease despite the vast diversity of var sequences observed. A better understanding of the link between var groups and disease phenotype is needed to confirm the implications of a recombination hierarchy for patterns of immunity and disease, however.

Our analysis also supports the idea that var gene expression patterns within hosts are ordered, to some extent, rather than random. Although theoretical models (for a review see Frank & Barbour 2006) and empirical studies (Horrocks et al. 2002, 2004) have shown that some level of ordering of expression of different var genes is probably needed to maintain antigenic variation within the host, others have concluded that var genes are expressed randomly (for example Fernandez et al. 2002). The hierarchical structure of the response network, and the fact that low reciprocity between heterologous responses was consistently observed in these data, suggest some intrinsic ordering in expression is likely. The role of the host immune system in the orchestration of these expression patterns is not known, however, and probably plays an important part in determining the order of appearance of different variants (Recker et al. 2004).

The diversity of the var genes and their complex expression patterns within hosts will continue to hinder attempts to understand their population structure. We believe new methods of analysing serological data are needed, particularly while we remain unable to link sequence and phenotype, to provide insights into the underlying population structure of the malaria parasite.

The authors would particularly like to acknowledge Kevin Marsh for his support of this study, as well as Oliver Pybus, Mario Recker and Chris Newbold for helpful comments. C.O.B. is a Sir Henry Wellcome Postdoctoral Research Fellow funded by the Wellcome Trust. P.C.B. is funded by a Wellcome Trust Advanced Training Fellowship in Tropical Medicine (060678) and a Wellcome Trust Project grant (076030).

APPENDIX A

(a) Agglutination assays and network generation

These assays take advantage of the fact that antibodies present in host serum can bind two identical epitopes at once. When antibody binding occurs on different infected red blood cells expressing the same PfEMP1 variant, the cells clump together into ‘agglutinates’, which can be analysed by means of microscopy based on their size and prevalence (Bull et al. 1998). Parasites are isolated from patients upon admission to hospital, representing the acute stage of infection (generally, hosts have not generated antibody responses to their infecting isolate at this stage). The ability of patients’ sera to form agglutinates is then measured for every isolate at this acute stage and also a few weeks later, at the convalescent stage of infection. During this period patients generate antibodies primarily to their own parasite, resulting in the diagonal stripe of increasing, homologous agglutination scores observed in the convalescent chequerboard in figure 1a. We did not use data from patients who were given a blood transfusion when generating the response network, since the source of the blood was likely to be a semi-immune adult with a large antibody repertoire.

The use of directed networks to analyse relationships between individuals has a long history in sociology and has begun to be adopted by other areas of natural science. As a result, many metrics and ways of analysing network structures are available, and can be used to test hypotheses about complex systems. We wanted to use network theory to quantitatively analyse the structure of cross-reactive immune responses between patients, since this has not been explored systematically in the past. In our networks, the nodes represented both the parasite and the patient it was isolated from. We ignored the relative strengths of agglutination when generating networks from these data and instead used a simple binary measure (i.e. positive or negative response, increasing or not increasing response), since the relative differences in scores are likely to reflect inherent differences between patients rather than differences between parasite epitopes.

Analysis was performed using the network and social network analysis packages in R (Butts 2007; Butts et al. 2008; R Core Development Team 2008) and MATLAB v. R2007a (MATLAB 1997), and network figures were produced using Agna (Benta 2005).

(b) Derivation of structural metrics and ZISP scores

Network density is defined as the number of observed edges divided by the total possible number of edges, given the number of nodes. The reciprocity of the network is in
this case the ‘edgewise’ reciprocity, defined simply as the proportion of reciprocated edges. Transitivity here is measured as the proportion of potentially intransitive triads that satisfy the constraint: $a \rightarrow b \rightarrow c \Rightarrow a \rightarrow c$ (this is the ‘weak’ form of transitivity).

$Z$ scores give the prevalence of different triads in the data compared with equivalent, randomly rewired networks (see Milo et al. 2002, 2004). SP scores are normalized $Z$ scores, representing the relative importance of each triad. The $Z$ score for each triad $i$ is derived as follows:

$$Z_i = (N_{\text{real}} - N_{\text{rand}}) / \text{std}(N_{\text{rand}}),$$

where $N_{\text{real}}$ is the observed number of occurrences of triad $i$ in the data, and $N_{\text{rand}}$ is the number of triads in the randomly rewired networks. Here, for each observed network we generated 10 000 randomly rewired networks.

The SP score for each triad $i$ is then derived as

$$SP_i = Z_i \left( \sum Z_i^2 \right)^{1/2},$$

such that the vector of $Z$ scores is normalized to give relative rather than absolute significance of different triads.

(c) Modelling the effect of population structure and expression patterns on hypothetical response networks

The parasite population was generated assuming that each parasite had 60 $\text{var}$ genes that were: (i) drawn randomly from a global $\text{var}$ pool (which varied in size between 70 and 2000 $\text{var}$ genes); (ii) chosen from a pool of one of several strains defined by non-overlapping $\text{var}$ gene repertoires (the number of strains was varied from 2 to 50); or (iii) in part chosen from non-overlapping combinations of $\text{var}$ genes (10 genes per genome) and in part drawn from a global pool. For the latter case, the intermediate population, the number of strains and global $\text{var}$ pool was varied within the same range as the other populations. Hosts were assumed to be exposed to a fraction of their parasite’s $\text{var}$ gene repertoire, drawn from a normal distribution around mean exposure varying between 10 and 60 of the possible 60 $\text{vars}$. The effects of differences in infection length were explored by changing the variance of the mean number of $\text{vars}$ experienced ranging between 2 and 32.

REFERENCES


