Prospective Identification of Malaria Parasite Genes under Balancing Selection

Kevin K. A. Tetteh¹, Lindsay B. Stewart¹, Lynette Isabella Ochola¹,²,³, Alfred Amambua-Ngwa², Alan W. Thomas⁴, Kevin Marsh³, Gareth D. Weedall¹,⁵, David J. Conway¹,²

¹ Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom, ² MRC Laboratories, Fajara, Banjul, The Gambia, ³ KEMRI Centre for Geographic Medicine Research, Coast, Kilifi, Kenya, ⁴ Biomedical Primate Research Centre, Rijswijk, The Netherlands, ⁵ School of Biological Sciences, University of Liverpool, Liverpool, United Kingdom

Abstract

Background: Endemic human pathogens are subject to strong immune selection, and interrogation of pathogen genome variation for signatures of balancing selection can identify important target antigens. Several major antigen genes in the malaria parasite Plasmodium falciparum have shown such signatures in polymorphism-versus-divergence indices (comparing with the chimpanzee parasite P. reichenowi), and in allele frequency based indices.

Methodology/Principal Findings: To compare methods for prospective identification of genes under balancing selection, 26 additional genes known or predicted to encode surface-exposed proteins of the invasive blood stage merozoite were first sequenced from a panel of 14 independent P. falciparum cultured lines and P. reichenowi. Six genes at the positive extremes of one or both of the Hudson-Kreitman-Aguade (HKA) and McDonald-Kreitman (MK) indices were identified. Allele frequency based analysis was then performed on a Gambian P. falciparum population sample for these six genes and three others as controls. Tajima’s D (TjD) index was most highly positive for the msp3/6-like PF10_0348 (TjD = 1.96) as well as the positive control ama1 antigen gene (TjD = 1.22). Across the genes there was a strong correlation between population TjD values and the relative HKA indices (whether derived from the population or the panel of cultured laboratory isolates), but no correlation with the MK indices.

Conclusions/Significance: Although few individual parasite genes show significant evidence of balancing selection, analysis of population genomic and comparative sequence data with the HKA and TjD indices should discriminate those that do, and thereby identify likely targets of immunity.

Introduction

Dynamic interactions between hosts and pathogens result in positive selection on molecules responsible for pathogen invasion, host resistance, and pathogen evasion of host resistance [1–3]. Many surface protein genes reveal signatures of positive selection, with several clear examples in malaria parasites [4–11]. These include signatures of directional selection that increases fixation rates and divergence among populations and species [7–9] and balancing selection that maintains diversity within local populations [4–6]. Although heterozygote advantage might operate during the brief gamete fertilization and diploid stages inside the mosquito host, balancing selection on proteins in the haploid asexual blood stage is probably due to negative frequency-dependent immune selection [12–18]. Predictions that blood-stage proteins under balancing selection are important targets of acquired immunity have been supported by antibody inhibition assays in culture [19–22], and by studies of naturally acquired antibodies and incidence of clinical malaria in endemic populations [17,23–25].

The ~23 Mb P. falciparum genome that encodes ~5300 proteins presents a challenge for identifying targets of immunity, but scans of currently available genome sequence data from different isolates can already identify loci with unusually high levels of polymorphism [5–7]. With available data, such scans do not discriminate loci under transient directional selection (such as drug resistance genes) [5], from those under balancing selection [6]. In parallel with the increasing availability of data on genome sequence diversity, there have been many developments of tests for evidence of positive directional selection [26,27], but less focus on identifying genes under balancing selection [28]. The data requirements of different tests vary, so choices among these should determine the strategic sampling of parasite isolates for whole genome sequencing. Allele frequency based tests require sequences of many isolates from at least one defined population for Tajima’s D (TjD) index [29,30], or multiple populations for Wright’s fixation (FST) indices [31], while polymorphism-versus-divergence tests such as the Hudson-Kreitman-Aguade (HKA) [32] or McDonald-Kreitman (MK) [33] indices can be performed.
on fewer isolate sequences but require comparator sequence from a closely-related species.

To evaluate indices for large scale identification of genes under balancing selection, analysis was performed on a prospective sample of genes encoding surface-accessible proteins in a single parasite stage. Gene transcription and proteomic data on the merozoite of *P. falciparum* [34] and ongoing identification of proteins specifically located on the surface or in the apical organelles [35–37], allows components of this important erythrocyte invasive stage to be investigated. Studies comparing different members of small gene families expressed at this stage, including five *efa* [15,16], three *Rh* [39] and five *RhopH1/Clag* [39] genes had previously shown how variable and locus-specific the signatures of selection are. The present study investigates a prospective panel of twenty six additional merozoite protein-coding genes, by sequencing from diverse laboratory cultured *P. falciparum* isolates and *P. reichenowi* to enable polymorphism-versus-divergence tests. A subset of the genes, together with positive and negative controls, was then sequenced from an endemic population sample in The Gambia to give an allele frequency based analysis with independent data. The HKa and TjD indices with the respective types of data sets are promising for large-scale analyses to detect the important minority of all parasite genes that are under balancing selection.

**Results**

**Polymorphism and divergence analyses**

A screen for signatures of non-neutrality was first applied to a set of 26 genes known or predicted to encode surface-exposed proteins of the merozoite stage of the parasite. Alleles of each of the genes were sequenced from 14 cultured lines of *P. falciparum*, representing species-wide polymorphism, along with the *P. reichenowi* orthologue of each gene (Accession numbers are listed in Supplementary Table S1). Figure 1 shows the positions of insertions, deletions, and nucleotide polymorphisms and fixed differences between the species, as well as repetitive sequences (omitted from alignment-based analyses). Full alignments of the sequences are shown in Supplementary Figures S1, S2, S3, and the repetitive sequences in 15 of the genes are shown in Supplementary Figure S4. For one gene (MRS92) there was a stop codon in the *P. reichenowi* orthologue, and for another (PFI0_0348) there was a stop codon in the allele of *P. falciparum* clone RO33; for analysis, these stop codons were removed and the remainder of each sequence was included in frame. For PFI0_0348, three of the *P. falciparum* isolates unexpectedly contained two distinct gene sequences, one of which was identical across the three isolates but different from all others (alignment shown in Supplementary Figure S5); this extra sequence was termed ‘copy B’ and omitted from analyses.

Table 1 shows the overall nucleotide diversity (π) and inter-specific divergence (K) indices for each gene, and the HKAr polymorphism-versus-divergence index (π/K ratio). The π values per gene ranged from 0 (no polymorphism in rap3 or the msp7-like gene PFI0_0348) up to 0.0568 (the highly polymorphic msp3/6-like gene PFI0_0348). If the repeat sequences had not been identified and appropriately removed prior to analysis, the apparent diversity (π) values for several genes would have been elevated (Supplementary Figure S6). Also shown are the numbers of synonymous and nonsynonymous polymorphisms among the aligned *P. falciparum* alleles, and fixed differences from the *P. reichenowi* orthologues, together with the results of the MK test on these proportions.

Both HKAr and MK indices express polymorphism relative to divergence, although HKAr does this directly while MK skew expresses the imbalance in the ratios of nonsynonymous and synonymous polymorphisms versus fixed differences. Over all the genes there was no correlation between the HKAr and MK indices (Spearman’s ρ = −0.26, P = 0.23) (Figure 2). The three genes showing the highest HKAr indices were the msp3/6-like gene PFI0_0348, msp7, and sera5. Three other genes showed the most positively skewed MK indices (with higher ratio of nonsynonymous to synonymous polymorphisms than fixed differences), two of which were significant (P92/16-cys and rap148, P < 0.05), and one nearly significant (P93/16-cys; P = 0.06) (Table 1).

**Allele frequency based tests in an endemic *P. falciparum* population sample**

The six genes with extreme positive values of either the HKAr or MK indices noted above were chosen for sequencing from a Gambian *P. falciparum* population, to allow application of allele frequency-based analyses. A gene that was previously shown to be under balancing selection in other populations (ama1, PF11_0344) was incorporated as a positive control, and a merozoite stage-specific gene encoding an internally expressed protein (etamp10.2, PFI0_0323) was included as a negative control as well as one of the genes that had neither an extreme MK skew nor HKAr index in the analysis above (rama).

First, for the positive control gene ama1, a sample of 114 allele sequences was derived from the Gambian population, and random subsets of the data were sampled to examine the relationship between sample size and Tajima’s D (TjD) value. This showed that there is an increase in the point estimate of TjD with increasing sample size, but that above a sample size of ~50 alleles the rate of increase becomes much slower (Figure 3). Therefore, to obtain a minimum sample size of 50 allele sequences per gene, amplification and direct sequencing was performed from a random panel of 89 of the Gambian isolates together with positive and negative control samples in a 96-well array. After the expected dropout of some isolates with mixed allele sequences that produced superimposed electropherogram traces (particularly a problem for reading sequences of genes with variable repeat lengths), the numbers of allele sequences obtained ranged from 56 (for sera5) to 88 (for P93/16-cys). (Accession numbers are listed in Supplementary Table S2).

Table 2 shows that the TjD value was positive for the ama1 control, and negative for the etamp10.2 and rama controls, as expected (see Discussion). The value for one of the test genes was highly positive (the msp3/6-like gene PFI0_0348). Values for other genes were either modestly positive (for sera5, P93 and msp7), or negative. Across the 9 genes tested in the population, there was a strong positive correlation between the TjD index and the HKAr index (Spearman’s ρ = 0.92, P<0.001; Figure 4A), but not with the MK index (ρ = −0.07, P = 0.87; Figure 4B). Indeed, apart from ama1, the two genes that had the most positive MK skew (P92/16-cys, and rap148) had the most negative TjD indices, illustrating that positively skewed MK indices are not commonly due to balancing selection. Fu and Li’s F index correlated strongly with TjD (ρ = 0.97, P<0.001), and had similar correlations as TjD with the other indices (significant for HKAr, ρ = 0.88, P = 0.002; not significant for MK, ρ = 0.00, P = 1.0). For the 7 genes included in both data sets (Table 1 and Table 2), TjD indices from the Gambian population correlated with the HKAr indices calculated with the lab isolates (ρ = 0.79, P = 0.036), but did not correlate with the MK indices (ρ = −0.57, P = 0.18).

**Discussion**

This analysis of a large panel of merozoite stage-specific genes for signatures of balancing selection enables recommendations for
scaling up to whole genome analyses. The advantage of using more than one test method is illustrated, particularly where it is possible to perform polymorphism-versus-divergence analysis as well as allele frequency-based analysis, ideally with independent datasets.

Although the MK test was originally developed to detect positive directional selection that has led to differences between species (thus displaying a negative skew in the ratios) [33], it has previously shown a very positive skew for some antigen genes that are under strong balancing selection (e.g. *ana1*, *eba175*, *trap*) [13,15,18]. However, it did not perform well here for prospectively identifying additional genes under such selection. Genes under weak negative selection (suppressing fixation of nonsynonymous changes between species lineages but allowing nonsynonymous

**Figure 1. Scheme of the 26 genes studied showing positions of individual nucleotide polymorphisms among 14 *P. falciparum* isolates as vertical lines above each gene.** Fixed differences between *P. falciparum* and *P. reichenowi* are shown as vertical lines below each gene. Positions of insertions, deletions, and repetitive sequence are also indicated on each gene, as well as verified and predicted protein domains. Stop codons within genes are shown for MRSP3 (in *P. reichenowi* only) and for PF10_0348 (in *P. falciparum* R033 isolate). Horizontal bars above some genes indicate portions not analysed.

doi:10.1371/journal.pone.0005568.g001
have some limitations when used alone [27]. The TjD test is TjD (requiring a population sample) which are informative but disparate laboratory cultured lines or a population sample) and encouraging for the application of both HKAr (to data from allele frequency-based TjD index in independent data. This is purpose. The HKAr indices here did not correlate with the MK divergence data in a manner that is more efficient for the present and synonymous classes can exceed the limitations of the data. in most genes [7], so splitting these further into nonsynonymous and synonymous classes would be associated with low rather than high TjD values [41,42], as seen here for the genes $PF14_0102$, $PF13_0348$, and $hph148$. Another limitation of the MK test is that its power is very low for analysis of most $P. falciparum$ genes, as the number of polymorphic sites is low in most genes [7], so splitting these further into nonsynonymous and synonymous classes can exceed the limitations of the data.

The HKAr index utilizes the overall polymorphism-versus-divergence data in a manner that is more efficient for the present purpose. The HKAr indices here did not correlate with the MK indices for the same set of data, but did correlate strongly with the allele frequency-based TjD index in independent data. This is encouraging for the application of both HKAr (to data from disparate laboratory cultured lines or a population sample) and TjD (requiring a population sample) which are informative but have some limitations when used alone [27]. The TjD test is influenced not only by selection but also by the population history that can alter neutral allele frequency expectations [43], as in the case of previous population expansion that causes the neutral index for $P. falciparum$ to be negative rather than zero [15,44]. This makes the test for balancing selection conservative, and it is possible that the modest positive TjD values here for $sera5$, $PF38/6-cys$, and $msp7$ also reflect balancing selection. The sampling distribution of neutral values could be re-estimated by modeling past population growth parameters, which would lead to negative neutral values of TjD [44] and thus increase the sensitivity of the test for genes under balancing selection. Although we have shown elsewhere how such modifications can potentially benefit the application of the TjD test to $P. falciparum$ genes [15], here we retain a conservative approach by testing TjD values under a constant population size model, as our primary aim is to compare among different types of standard tests with minimal modifications. The between-species divergence (used as denominator for

<table>
<thead>
<tr>
<th>Gene</th>
<th>Locus</th>
<th>Nt</th>
<th>$x$ (10^-3)</th>
<th>$K$ (10^-3)</th>
<th>HKAr ($π/K$)</th>
<th>Syn</th>
<th>Non-syn</th>
<th>MK test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$PF12/6-cys$</td>
<td>PFF0615c</td>
<td>990</td>
<td>1.0</td>
<td>35.9</td>
<td>0.028</td>
<td>12</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>$PF92/6-cys$</td>
<td>PFI3_0338</td>
<td>2319</td>
<td>1.2</td>
<td>41.6</td>
<td>0.029</td>
<td>39</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>$PF38/6-cys$</td>
<td>PFED095c</td>
<td>900</td>
<td>3.4</td>
<td>21.3</td>
<td>0.160</td>
<td>6</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>$PF113$</td>
<td>PFI4_0201</td>
<td>2619</td>
<td>0.5</td>
<td>30.6</td>
<td>0.016</td>
<td>31</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>$MSP10$</td>
<td>PFF0995c</td>
<td>1329</td>
<td>1.7</td>
<td>45.6</td>
<td>0.037</td>
<td>15</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>$MSP6$</td>
<td>PFI0_0346</td>
<td>867</td>
<td>17.5</td>
<td>59.3</td>
<td>0.295</td>
<td>5</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>$MSP3/6-like$</td>
<td>PFI0_0347</td>
<td>1230</td>
<td>1.1</td>
<td>37.2</td>
<td>0.030</td>
<td>14</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>$MSP3/6-like$</td>
<td>PFI0_0348</td>
<td>1935</td>
<td>56.8</td>
<td>82.8</td>
<td>0.686</td>
<td>16</td>
<td>72</td>
<td>58</td>
</tr>
<tr>
<td>$MSP7$</td>
<td>PFI3_0197</td>
<td>1044</td>
<td>5.6</td>
<td>59.8</td>
<td>0.094</td>
<td>13</td>
<td>1</td>
<td>44</td>
</tr>
<tr>
<td>$MRSP1$</td>
<td>PFI3_0196</td>
<td>1143</td>
<td>0.4</td>
<td>43.8</td>
<td>0.009</td>
<td>8</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td>$MRSP2$</td>
<td>MAL13P1.174</td>
<td>819</td>
<td>0.5</td>
<td>36.9</td>
<td>0.014</td>
<td>10</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>$MRSP3$</td>
<td>PFI3_0193</td>
<td>897</td>
<td>0.2</td>
<td>53.9</td>
<td>0.028</td>
<td>6</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>$MRSP4$</td>
<td>MAL13P1.173</td>
<td>897</td>
<td>1.2</td>
<td>26.9</td>
<td>0.045</td>
<td>11</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>$MRSP5$</td>
<td>PFI3_0191</td>
<td>870</td>
<td>0.3</td>
<td>19.7</td>
<td>0.015</td>
<td>7</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>$MRSP-like$</td>
<td>PFI3_0192</td>
<td>1446</td>
<td>1.1</td>
<td>29.6</td>
<td>0.037</td>
<td>8</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>$MRSP-like$</td>
<td>PFI3_0194</td>
<td>567</td>
<td>0.0</td>
<td>5.3</td>
<td>0.000</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$MRSP9/A8R4$</td>
<td>PFL1385c</td>
<td>1761</td>
<td>1.4</td>
<td>48.2</td>
<td>0.029</td>
<td>29</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>$SERA5$</td>
<td>PF00340c</td>
<td>609</td>
<td>16.0</td>
<td>98.2</td>
<td>0.163</td>
<td>7</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>$RAMA$</td>
<td>MAL7P1.208</td>
<td>1326</td>
<td>1.0</td>
<td>26.8</td>
<td>0.036</td>
<td>10</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>$Rhoph148$</td>
<td>PFI3_0348</td>
<td>3132</td>
<td>0.7</td>
<td>45.1</td>
<td>0.016</td>
<td>64</td>
<td>1</td>
<td>76</td>
</tr>
<tr>
<td>$Prohibitin$</td>
<td>PFI0_0144</td>
<td>894</td>
<td>0.8</td>
<td>9.0</td>
<td>0.089</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$RAPI$</td>
<td>PFI4_0102</td>
<td>2271</td>
<td>2.0</td>
<td>29.1</td>
<td>0.069</td>
<td>18</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>$RAP2$</td>
<td>PFED0980c</td>
<td>1053</td>
<td>1.5</td>
<td>29.5</td>
<td>0.051</td>
<td>8</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>$RAP3$</td>
<td>PFED0075c</td>
<td>1119</td>
<td>0.0</td>
<td>26.8</td>
<td>0.000</td>
<td>17</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>$PD4$</td>
<td>PFDD0555w</td>
<td>894</td>
<td>0.5</td>
<td>28.5</td>
<td>0.018</td>
<td>13</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1. HKAr indices summarizing polymorphism ($π$, among 14 $P. falciparum$ laboratory isolates) and divergence ($K$, from $P. reichenowi$) of 26 merozoite stage genes together with MK tests on synonymous (syn) and nonsynonymous (non-syn) polymorphic and fixed differences.

$N$, number of aligned nucleotides.
$\circ$, repeats removed from gene sequences.
$\bullet$, internal stop codon in $P. reichenowi$ mrsp3 gene, and in an allele of $P. falciparum$ PF10_0348 (codons removed from analysis).
$\ast$, complex codons in rap1 and rap2 not analysed.
$\dagger$, new sequence aligned when $P. reichenowi$ added (PF10_0347 N = 1200, PF10_0348 N = 1866, PF10_0352 N = 1104, MRSP1 N = 1140, MRSP2 N = 816, MRSP3 N = 705, MRSP9 N = 1758, SERA5 N = 561, PD4 N = 882); Sequences submitted to Genbank (Accession numbers are listed in Supplementary Table S1).

doi:10.1371/journal.pone.0005568.t001

Selection on Parasite Genes
the HKAr index) is less sensitive to the effects of demographic history than the number of polymorphic nucleotides in a population sample (used towards both the numerator and denominator of the TjD index). Therefore, concordant high values of HKAr and TjD indices for these genes are very supportive of balancing selection, encouraging a two-dimensional approach to identifying genes under such selection [45].

Of the panel of malaria parasite merozoite stage genes prospectively investigated here, the one with the strongest signature of balancing selection, as indicated by the HKAr as well as the TjD index, was PF10_0348 (a member of the msp3/6-like family). This gene had two unusual features, however. Firstly, one of the laboratory isolates and a minority of field isolates had an internal stop codon, so these may be null with regard to protein expression. Secondly, a minority of isolates contained a second, more divergent, PF10_0348-like sequence that does not match with other known loci, and although this was not included in the analysis it suggests there are paralogous genes in some parasites. Further studies of PF10_0348 are needed, including characterization of its transcription and protein expression. Generally, if a protein is known to be encoded by multiple gene copies that undergo gene conversion or ectopic recombination, non-classical approaches to analyzing

Figure 2. Scatterplot of McDonald-Kreitman (MK) skew (log2 transformation of the neutrality index with 0 representing no skew) and Hudson-Kreitman-Aguade ratio (HKAr) of polymorphism (in P. falciparum) and divergence (between P. falciparum and P. reichenowi) for each of the genes studied in Table 1. Two of the 26 genes had no polymorphism and are not plotted. There was no significant correlation between the two indices (Spearman’s ρ = −0.26). Six genes at the outer fringe of the distributions, and one in the middle, are labelled and were selected for population based analysis.

doi:10.1371/journal.pone.0005568.g002

Figure 3. Sample size optimisation for Tajima’s D test assessed using a large sample of ama1 gene sequences from a Gambian P. falciparum population. 100 random samples of each size (n) were taken from a set of 114 ama1 sequences covering the region encoding the ectodomain of the protein (nucleotides 442–1743). The median values and 95% confidence limits (3rd and 97th centile values of 100 estimates) of π, θi, and Tajima’s D indices were calculated for each sample size.

doi:10.1371/journal.pone.0005568.g003
sequence polymorphism may need to be developed [46,47], and it is possible that deeper sequencing of *P. falciparum* genomes will reveal this to be the case for more merozoite stage genes than has yet been appreciated [38,39,48].

It should be noted that any test, or combination of tests, may be conservative if a protein is under positive directional selection as well as balancing selection. Previous data indicate this to be the case for the erythrocyte binding antigen EBA175, a merozoite ligand adapted for binding to species-specific receptor structures on host glycoporphin A (leading to an excess of sequence substitutions between species) [49] and also under selection from acquired immune responses (leading to an excess of polymorphisms) [15,16]. Different types of selection that lead towards fixed differences or maintenance of allelic diversity should be discriminated where possible by information on protein function or antigenicity, as their co-occurrence limits the power of general tests.

Means of scaling up to a comprehensive investigation of balancing selection in a pathogen can be determined by available funding, accessibility of appropriate genome samples, as well as the epidemiological and population genetic structure of the pathogen. Capillary sequencing is rate-limiting for large population analyses of many genes, so it can be efficient to focus such analysis on genes with more than a minimal amount of polymorphism, and for *P. falciparum* initial screening can now be performed using emerging genome sequence diversity data, accessible through PlasmoDB (www.plasmodb.org) [5–7]. However, solid phase ‘next generation’ sequencing methods should soon allow whole genome sequencing to be performed on population samples [30]. As such methods can now be successfully applied to *P. falciparum* genome sequencing [51], it is possible that >50 isolates from an endemic population could be realistically sequenced for a complete genome screen to detect signatures of balancing and directional selection. Subsequent comparison of allele frequency distributions in multiple population genomic samples could further test these signatures, and such data could also be used to refine less intensive approaches to identifying selection in pathogens with large genomes.

**Figure 4.** Correlations between the allele frequency-based Tajima’s D index and two different polymorphism-versus-divergence indices for 9 merozoite stage genes (N = 55 alleles of each gene sampled from a Gambian *P. falciparum* population, and interspecies comparison with *P. reichenowi*). A. Scatterplot showing strong positive correlation (Spearman’s $\rho = 0.92$, $P < 0.001$) between the HKA ratio (HKAr) index and the allele frequency-based Tajima’s D index. B. Scatterplot indicating no correlation (Spearman’s $\rho = -0.07$, N.S.) between the McDonald Kreitman (MK) skew and Tajima’s D index.

doi:10.1371/journal.pone.0005568.g004

**Materials and Methods**

**Ethics Statement**

Written informed consent was given by the parents of each child, and verbal assent by each child, for collection of a <5 ml venous blood sample for analyses including investigation of malaria parasite DNA. The study and protocol was approved by the Scientific Co-ordinating Committee and the Ethics Committee of the MRC Gambia Unit and the Gambian Government.

**Gene sequencing from *P. falciparum* cultured isolates and *P. reichenowi***

Twenty six merozoite stage-expressed genes were chosen for analysis as they encode surface or apical organelle proteins in *P. falciparum* merozoites, or are predicted to do so by close homology with proteins that are so localized. These encode five surface proteins that are GPI-anchored (*PF12*, *PF38*, *PF92*, *PF113*, and *msp10*), 14 known or predicted to be surface associated but not membrane anchored (4 members of the msp3/6-like family; *msp6*, *PF10_0347*, *PF10_0348*, *PF10_0352*; 6 members of the msp7-like family; *msp7*, *msp1*, *msp2*, *msp3*, *msp4*, *msp5*, and 4 others; the msp7-linked *PF13_0192* and *PF13_0194*, and the unlinked *msp3* and *msp9*, and 7 expressed in the apical rhoptry organelle (*rama*, *shap148*, *rap1*, *rap2*, *rap3*, prohibita, *Pf13*). Each gene was amplified from genomic DNA of 14 genotypically distinct cultured *P. falciparum* lines from diverse original sources (3D7, cloned from an airport malaria case in The Netherlands; D6, RO33 and Palo Alto, from Africa; K1, T9/96, T9/102, Dd2, FCC2 and D10, from Southeast Asia; FCR3 and Wellcome, nominally from Africa but suspected to have been previously cross-contaminated by parasites of unknown source during culture; HB3 from Honduras; 7G8 from Brazil and the only existing known isolate of *P. reichenowi* (CDC-1 strain isolated over 50 years ago from a chimpanzee from the Belgian Congo), using primers and amplification conditions listed in Supplementary Tables S3 and S4. PCR products were purified with the QiAquick PCR purification kit (QIAGEN, UK), and sequenced using the outer amplification primers and several internal sequencing primers, using BIG DYE.
Table 2. Analysis of polymorphism in 9 merozoite stage expressed *P. falciparum* genes in a Gambian population sample.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Locus</th>
<th>No. of isolates</th>
<th>Nt</th>
<th>π (×10^{-2})</th>
<th>K (×10^{-2})</th>
<th>HKAr (π/K)</th>
<th>McDonald-Kreitman</th>
<th>Tajima’s D</th>
<th>Fu &amp; Li F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SYN</td>
<td>NONSYN</td>
<td>MK Fixed</td>
</tr>
<tr>
<td>MSP3/6-like</td>
<td>PF10_0348</td>
<td>66</td>
<td>1896</td>
<td>38.7</td>
<td>72.3</td>
<td>0.535</td>
<td>17</td>
<td>68</td>
<td>57</td>
</tr>
<tr>
<td>msp7</td>
<td>PF13_0197</td>
<td>85</td>
<td>1050</td>
<td>6.4</td>
<td>62.0</td>
<td>0.103</td>
<td>13</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>rhop148</td>
<td>PF13_0348</td>
<td>81</td>
<td>1200</td>
<td>0.3</td>
<td>47.7</td>
<td>0.006</td>
<td>30</td>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>sera5</td>
<td>PF80340c</td>
<td>56</td>
<td>624</td>
<td>11.2</td>
<td>93.6</td>
<td>0.119</td>
<td>11</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>PF38/6-cys</td>
<td>PFEO0395c</td>
<td>88</td>
<td>903</td>
<td>3.6</td>
<td>21.3</td>
<td>0.169</td>
<td>6</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>PF92/6-cys</td>
<td>PF13_0338</td>
<td>87</td>
<td>2319</td>
<td>1.3</td>
<td>41.9</td>
<td>0.031</td>
<td>37</td>
<td>6</td>
<td>56</td>
</tr>
<tr>
<td>Rama</td>
<td>MAL7P1208</td>
<td>68</td>
<td>1323</td>
<td>0.7</td>
<td>26.4</td>
<td>0.027</td>
<td>11</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>Etramp10.2</td>
<td>PF10_0323</td>
<td>78</td>
<td>990</td>
<td>1.4</td>
<td>50.0</td>
<td>0.028</td>
<td>14</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>ama1 (+ve control)</td>
<td>PF11_0344</td>
<td>114</td>
<td>1302</td>
<td>16.3</td>
<td>40.0</td>
<td>0.420</td>
<td>11</td>
<td>3</td>
<td>21</td>
</tr>
</tbody>
</table>

Nt, number of aligned nucleotide positions analysed.
* − stop codon in 7 alleles of PF10_0348 (codon removed from analysis).
\(\circ\) − repeats removed from gene sequences for analysis.
\(\circ\) − only one region of msp7 gene (of most dense polymorphism) was studied in the population.
\(\circ\) − 5 complex codons in ama1 not analysed.
\(\circ\) − less sequence aligned when *P. reichenowi* added (PF10_0348 N = 1824, msp7 N = 1041, sera5 N = 600).
\(\circ\) − p<0.05.
\(\circ\) − p<0.001.
\(\circ\) − p<0.001.
Values significantly different to 0 are shown in italics; for ama1 Tajima’s D is significant for Domain II (p<0.05) and Domain III (p<0.001), and Fu and Li’s F is significant for Domain III (p<0.02); Sequences submitted to Genbank (Accession numbers for population datasets of allele sequences are listed in Supplementary Table S2).
 terminator v3.1 chemistry (Applied Biosystems, UK) and an ABI 3130xl capillary sequencer (Applied Biosystems, UK). Sequences were assembled, edited and aligned using SeqMan and MegAlign (DNASTAR, Madison, WI). The data covered the complete or near complete coding sequence of each of the 26 genes, except sens5 for which analysis was focused on the most polymorphic exon 2.

Gene sequencing from a Gambian P. falciparum population
Parasites were studied from children aged <13 years living in the coastal urban/peri-urban area within 40 km south of Banjul who presented with P. falciparum malaria to the Medical Research Council (MRC) outpatients clinic at Fajara, or to the Royal Victoria Teaching Hospital in Banjul during a single malaria season (between September 2005 and January 2006). Heparinised venous blood samples were collected and centrifuged to remove plasma and leukocytes for other studies, and erythrocytes were then washed three times in RPMI medium. The DNA was extracted from packed erythrocytes using the QIAamp DNA Blood Mini Kit (QIAGEN, UK). The ama1 gene was chosen as a positive control gene as results from other populations indicate that it is consistently under balancing selection [12–14]. A sample of 114 ama1 alleles from the Gambian population (one allele from each of 114 independent clinical isolates) was sequenced to enable estimation of smaller sample sizes that would be efficient for detecting positive Tajima’s D values (see below). Six of the genes analysed in Table 1 (and Figure 2) that had positive HKAr or MK signatures, one that did not (rama), and another negative control (PF10_0323) encoding an internal protein, were then amplified from genomic DNA from a panel of 89 isolates, and products were directly sequenced to obtain a final sample size of >55 allele sequences of each gene (more than half of the isolates yielded a clear single allele sequence and those showing mixed sequences were discarded). All nucleotide alleles that had not been previously seen were confirmed by independent re-amplification and re-sequencing of the gene from each relevant sample.

Tests of neutrality
Tests were performed using DnaSP4 [52]. Tajima’s D (TJD) test detects departures from neutrality in allele frequency distributions by considering the number of polymorphic sites and the pairwise nucleotide diversity [29], while Fu and Li’s F test is based on the number of polymorphic sites with singleton alleles [53]. The McDonald-Kreitman (MK) test [33], compares numbers of nonsynonymous (\( \delta_s \)) and synonymous (\( \delta_s \)) nucleotide changes polymorphic (P) within species, and apparently fixed (F) between closely-related species, with a Fisher’s exact test on the 2×2 contingency table. A measure of MK skew was calculated: \[ \log_2 \left( \frac{[P_{NS}/F_{NS}]/[P_{FS}/F_{FS}]}{[P_{NS}/F_{NS}]/[P_{FS}/F_{FS}]} \right) \] where 0 represents no skew, positive values represent an excess of nonsynonymous polymorphism (or deficiency of nonsynonymous differences between species) and negative values represent selection against nonsynonymous polymorphism (or elevated fixation of nonsynonymous differences). The Hudson-Kreitman-Aguade ratio (HKAr) index for a gene is the pairwise nucleotide polymorphism (\( \pi \)) divided by the divergence (\( \delta \)) from a closely-related species, and is used here to identify genes with exceptionally high ratios [45].

Supporting Information
Figure S1
Found at: doi:10.1371/journal.pone.0005568.s001 (0.17 MB PDF)

Figure S2
Found at: doi:10.1371/journal.pone.0005568.s002 (0.36 MB PDF)

Figure S3
Found at: doi:10.1371/journal.pone.0005568.s003 (0.27 MB PDF)

Figure S4
Found at: doi:10.1371/journal.pone.0005568.s004 (0.05 MB PDF)

Figure S5
Found at: doi:10.1371/journal.pone.0005568.s005 (1.42 MB PDF)

Figure S6
Found at: doi:10.1371/journal.pone.0005568.s006 (0.03 MB DOC)

Table S1
Found at: doi:10.1371/journal.pone.0005568.s007 (0.05 MB DOC)

Table S2
Found at: doi:10.1371/journal.pone.0005568.s008 (0.03 MB DOC)

Table S3
Found at: doi:10.1371/journal.pone.0005568.s009 (0.06 MB DOC)

Table S4
Found at: doi:10.1371/journal.pone.0005568.s010 (0.06 MB DOC)

Acknowledgments
We are grateful to Michael Walther, Natalia Gomez-Escobar, and Joseph Okebe for help with collection of blood samples for analysis of Gambian parasite DNA, to Clemens Kocken for assistance in preparation of the P. reichenowi parasite material, and to Brett Lowe for helpful discussions. The Wellcome Trust Sanger Institute Pathogen Genomes group gave valuable support by making available the P. reichenowi partial genome shotgun sequence.

Author Contributions
Conceived and designed the experiments: KKAT KM GDW DJC. Performed the experiments: KKAT LBS AAN DJC. Analyzed the data: KKAT LBS LIO GDW DJC. Contributed reagents/materials/analysis tools: KKAT LBS AWT DJC. Wrote the paper: KKAT KM GDW DJC. Contributed reagents/materials/analysis tools: KKAT LBS AWT DJC. Wrote the paper: KKAT KM GDW DJC.

References


