Geographical Distribution of Selected and Putatively Neutral SNPs in Southeast Asian Malaria Parasites

Tim J. C. Anderson,* Shalini Nair,* Dan Sudimack,* Jeff T. Williams,* Mayfong Mayxay;††
Paul N. Newton,‡‡§ Jean-Paul Guthmann,¶ Frank M. Smithuis,¶¶ Tran Tinh Hien,§
Ingrid V.F. van den Broek,** Nicholas J. White,§§ and François Nosten§§§

*Southwest Foundation for Biomedical Research, San Antonio, Texas; ‡Faculty of Medical Sciences, National University of Laos, Vientiane, Lao PDR; ¶Wellcome Trust-Mahosot-Oxford Tropical Medicine Research Collaboration, Mahosot Hospital, Vientiane, Lao PDR; §§Centre for Tropical Medicine and Vaccinology, Churchill Hospital, Oxford, United Kingdom; †Épicentre (Médecins Sans Frontières-France), Paris, France; ‡‡Artsen Zonder Grenzen, Médecins Sans Frontières-Holland, Yangon, Myanmar; #Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam; **Médecins Sans Frontières (UK), London, United Kingdom; ††Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand; and §§§Shoklo Malaria Research Unit, Mae Sot, Tak, Thailand

Loci targeted by directional selection are expected to show elevated geographical population structure relative to neutral loci, and a flurry of recent papers have used this rationale to search for genome regions involved in adaptation. Studies of functional mutations that are known to be under selection are particularly useful for assessing the utility of this approach. Antimalarial drug treatment regimes vary considerably between countries in Southeast Asia selecting for local adaptation at parasite loci underlying resistance. We compared the population structure revealed by 10 nonsynonymous mutations (nonsynonymous single-nucleotide polymorphisms [nsSNPs]) in four loci that are known to be involved in antimalarial drug resistance, with patterns revealed by 10 synonymous mutations (synonymous single-nucleotide polymorphisms [ssSNPs]) in housekeeping genes or genes of unknown function in 755 Plasmodium falciparum infections collected from 13 populations in six Southeast Asian countries. Allele frequencies at known nsSNPs underlying resistance varied markedly between locations ($F_{ST}$ = 0.18–0.66), with the highest frequencies on the Thailand-Burma border and the lowest frequencies in neighboring Lao PDR. In contrast, we found weak but significant geographic structure ($F_{ST}$ = 0–0.14) for 8 of 10 ssSNPs. Importantly, all 10 nsSNPs showed significantly higher $F_{ST}$ ($P < 8 \times 10^{-5}$) than simulated neutral expectations based on observed $F_{ST}$ values in the putatively neutral ssSNPs. This result was unaffected by the methods used to estimate allele frequencies or the number of populations used in the simulations. Given that dense single-nucleotide polymorphism (SNP) maps and rapid SNP assay methods are now available for $P. falciparum$, comparing genetic differentiation across the genome may provide a valuable aid to identifying parasite loci underlying local adaptation to drug treatment regimes or other selective forces. However, the high proportion of polymorphic sites that appear to be under balancing selection (or linked to selected sites) in the $P. falciparum$ genome violates the central assumption that selected sites are rare, which complicates identification of outlier loci, and suggests that caution is needed when using this approach.

Introduction

Lewontin and Krakauer (1973) originally suggested that patterns of genetic population structure could be used to identify loci that are under selection. They reasoned that loci involved in local adaptation should show greater levels of differentiation between populations than neutrally evolving loci in which allele frequencies are determined by genetic drift alone. For many years, this appealingly simple idea was disregarded largely as a result of problems with the theoretical approximations of the distribution of $F_{ST}$ values of Lewontin and Krakauer (Nei and Maruyama 1975; Robertson 1975a, 1975b). A number of groups have recently revived this idea, improved the statistical methodology, and used this approach to search for loci involved in adaptation in organisms including humans, fish, intertidal snails, and mice (Beaumont and Nichols 1996; Wilding, Buttin, and Grahame 2001; Akery et al. 2002; Storz and Nachman 2003; Campbell and Bernatchez 2004; Storz and Dubach 2004; Storz, Paysier, and Nachman 2004; Vasemagi, Nilsson, and Primmer 2005). Typically, population structure is assessed at a number of randomly chosen loci, and those showing “extreme” levels of differentiation are identified as putatively selected genes. Both empirical and model-based approaches have been used to identify markers showing extreme patterns of genetic structure suggestive of selection. Work by Akery et al. (2002) exemplifies the empirical approach (see also McDonald 1994). They examined empirical distributions of $F_{ST}$ at 26,530 single-nucleotide polymorphisms (SNPs) in humans and found 156 genes showing higher than expected values and 18 genes showing lower than expected values. The model-based approach has been formulated by Bowcock et al. (1991) and Beaumont and Nichols (1996). These authors used coalescent simulations of metapopulations to generate distributions of $F_{ST}$ values expected under neutrality and used these distributions to detect outlier loci and assess probabilities for deviations from neutrality.

One problem with $F_{ST}$-based approaches is that we cannot tell how many of the outlier loci are actually evolving neutrally (type I errors) and how many selected loci are not detected (type II errors). Typically, in such studies, large numbers of loci may be compared which means that neutral loci may be misclassified as selected by chance alone. This problem can be minimized by raising the cutoff for statistical significance but comes at the cost of increasing the type II error rate and so is not entirely satisfactory. Beaumont and Balding (2004) examined the efficiency of both Bayesian and coalescent-based approaches for detecting selection in a simulated metapopulation under a wide variety of demographic conditions. They found that both

Key words: Plasmodium falciparum, genetic structure, single-nucleotide polymorphism, local adaptation, drug resistance, pfcr, pfmdr, dhfr, dhps.

E-mail: tanderso@darwin.sfbfr.org.

doi:10.1093/molbev/msi235

Advance Access publication August 10, 2005

© The Author 2005. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oupjournals.org
approaches detected selected loci efficiently using stringent statistical criteria with minimal type I or type II errors, but only when selection was strong (selection coefficients \( s > 0.1 \)). False negatives were common with weak selection \( s < 0.1 \). As a general guideline, they suggest that selected loci may be detected when the selection coefficient is five times greater than the migration rate.

Studies of functional polymorphisms that are known to be under contrasting selection in different populations provide an alternative empirical approach for validating the utility of such between-population comparisons for detecting selected regions of the genome (Taylor, Shen, and Kreitman 1995; Kohn, Pelz, and Wayne 2003). Malaria parasites infecting humans provide an especially useful system in this respect because four different genes are known to be involved in resistance to antimalarial treatments. Selection has been intense as the parasites are host specific and the drugs are very widely used. Furthermore, all the drugs responsible for selection at these loci have been used in Southeast Asian countries over the past 50 years, although drug treatment history differs between countries (fig. 1). For example, mefloquine-artemisinin combination therapy is currently used in Thailand, while in neighboring Lao PDR (Laos) to the east, chloroquine (CQ) and sulfadoxine-pyrimethamine (SP) are still the first- and second-line treatments in most of the country. We have previously described considerable heterogeneity in allele frequencies at the dihydrofolate reductase \((dhfr)\) locus on Chr 4 that underlies resistance to pyrimethamine (Plowe et al. 1997); (2) the CQ resistance transporter \((pfcrt)\) (Chr 7), a major gene determining CQ resistance (Fidock et al. 2000; Wootton et al. 2002); and (3) the multidrug resistance gene \((pfmdr)\) (Chr 5), an ATP binding cassette (ABC) transporter that influences resistance to multiple drug classes including CQ, quinine (QN), mefloquine (MFQ), and artemisinin derivatives (Reed et al. 2000; Price et al. 2004). To evaluate the relative roles of selection and genetic drift in determining the geographical distributions of SNPs observed at resistance loci, we also examine differentiation in 10 synonymous single-nucleotide polymorphisms (sSNPs) in housekeeping genes or genes of unknown function. We used these data in three ways: (1) to evaluate the efficacy of \(F_{ST}\)-based methods for detecting selection, (2) to determine patterns of neutral genetic structure in Southeast Asian malaria parasites, and (3) to map allele frequencies at four key genes involved in antimalarial resistance in Southeast Asia.

Materials and Methods

Geographical Sampling

We collected finger prick blood samples (~50 μl of blood absorbed on Whatmann 3-mm filter paper and dried) from malaria-infected patients from 13 locations in six Southeast Asian countries (Myanmar, Thailand, Cambodia, Laos, Vietnam, and Bangladesh). Ten of these populations have been described previously (Nair et al. 2003). In addition to these, we obtained samples from an additional population in Laos and two locations in Bangladesh. These are all areas of relatively low seasonal transmission. The samples from Laos were collected from Muang Feuang in Vientiane Province in 2000 (Mayxay et al. 2004). The samples from Bangladesh were collected from patients visiting the Médecins Sans Frontières clinics in Dighinala and Panchari (Khagrachari District) in the Chittagong Hill Tracts (van den Broek et al. 2004). These two sites are ~25 km apart. Use of blood samples described in this paper was approved by the institutional review board at the University of Texas Health Science Center at San Antonio and by review boards in the countries where blood samples were collected.

DNA Preparation

We cut 3-mm discs from each blood spot using a sterile hole punch and prepared DNA using the Generation Card Capture Kit (Gentra Systems, Minneapolis, Minn.). We also prepared DNA concurrently from a disc of plain 3-mm Whatman paper. These preparations were then used as negative controls in all amplification experiments. This procedure provides efficient control for contamination during both DNA preparation and amplification. We prepared one negative DNA preparation for every 12 samples prepared.

Genotyping of Mutations in Drug Resistance Genes

We used primer extension to genotype five key mutations in both dihydrofolate reductase \((dhfr)\) and
dihydropteroate synthase (dhps). These mutations are known to play important roles in the resistance against pyrimethamine and sulfadoxine, the two components of the combination drug SP (Fansidar) (Plowe et al. 1997). Details of the primer extension method are reported elsewhere (Nair et al. 2002). Pfert alleles conferring resistance to CQ contain multiple point mutations (in codons 72, 74, 75, 76, 220, 271, 326, 356, and 371) of which the pfert-K76T mutation is critical for resistance. Only codon 76 was genotyped here because the additional mutations were in complete linkage disequilibrium (LD) in a large sample of Thai parasites analyzed previously (Anderson et al. 2005). Both the K76T mutation in pfert and the N86Y mutation in pfmdr were genotyped by Apol digestion fluorescent end-labeled polymerase chain reaction (PCR) products and scoring of product sizes on a capillary sequencer (Anderson et al. 2003). The methods are similar to those used for scoring of sSNP polymorphisms in housekeeping genes (below).

SNPs on Chr 2 and 3

For comparison with the nonsynonymous single-nucleotide polymorphisms (nsSNPs) in known drug resistance genes, we scored point mutations at synonymous point mutations (sSNPs) in housekeeping genes or genes of unknown function on Chr 2 and 3. We used SNPs on just two chromosomes because SNP locations on the other 12 chromosomes of Plasmodium falciparum are not yet available (J. Mu and X.-Z. Su, personal communication). However, because linkage disequilibrium decays rapidly with distance in the P. falciparum genome (Conway et al. 1999; Anderson 2004), we believe that this is unlikely to substantially bias our results. The Chr 3 SNPs were identified by Mu et al. (2002). They sequenced 204 genes on Chr 3 in a panel of five parasites from Africa (two clones), Southeast Asia, South America, and Papua New Guinea, identifying 62 synonymous mutations. We excluded sSNPs in known antigen genes, in genes containing excessive levels of polymorphism suggestive of selection, or those for which restriction digest assays could not be designed. This left a total of 19 sSNPs. The approximate locations of the Chr 2 SNPs were identified by Volkman et al. (2002) by hybridization with Affymetrix chips. We identified 29 housekeeping genes or genes of unknown function that showed differential hybridization with 25mer oligo probes. We then sequenced these in the four laboratory malaria strains (W2, D6, 7G8, and HB3) studied in the microarray experiments and four samples from Thailand to identify the exact position of the SNPs within the oligohybridization probes. We screened the sSNPs on both Chr 2 and 3 for polymorphism in Southeast Asia using a panel of 25 samples (five each from Bangladesh, Thailand, Laos, Vietnam, and Myanmar). This was done by restriction digestion of PCR products using appropriate restriction enzymes and running the digested products on agarose gels. In cases where no restriction sites were available, sites were manufactured by incorporating a mismatch of 3–4 bp from the 3' end of an oligonucleotide. SNPs that were polymorphic in at least one population were included in the study. These SNPs were then genotyped in all 755 samples by restriction digestion of fluorescent-labeled PCR products. Primer sequences, amplification conditions, restriction enzymes, and positions of SNPs are described in table 1. We also provide details of additional SNPs on Chr 2 and 3 that were not polymorphic in the panel of 25 samples and were therefore not included in this study (Supplementary Tables S1 and S2, Supplementary Material online).

Measurement of Allele Frequencies

Accurate estimation of allele frequencies is critical for assessing population structure. Measurement of allele frequencies in microparasite populations is complicated by the presence of multiple clones within many infections (Hill and Babiker 1995; Rannala, Qiu, and Dykhuijen 2000). Counting all alleles identified within an infection results in overestimation of frequencies of rare alleles and underestimation of common alleles. To minimize bias, we used three different methods to estimate allele frequencies: (1) we excluded infections in which >1 allele was observed. Infections were deemed as containing multiple clones if peaks at the minor allele were >10% the height of those for the predominant allele on the ABI electropherograms. This approach is unbiased but results in considerable loss of data. (2) We picked the predominant allele present in each infection. Using this approach, a single genotype is obtained from each infection scored. Predominant alleles were identified by comparing the relative height of peaks on electropherograms. (3) We used a maximum likelihood procedure (Hill and Babiker 1995) to estimate both the allele frequency and the mean number of clones per infection. For each locus, we counted the numbers of infections containing allele 1, allele 2, or alleles 1 and 2. We assumed a positive Poisson distribution of parasite clones per host and used maximum likelihood to obtain the best-fit values for frequencies of alleles 1 and 2 and mean infection rate. A positive Poisson distribution was used because we sampled only P. falciparum–positive patients, so we had no hosts carrying zero parasite alleles.

Geographical Differentiation

We assessed genetic differentiation at each of the selected and putatively neutral loci using allele frequencies estimated using the three procedures described. We calculated FST in the total population sample using FSTAT (Goudet 2001), and confidence errors for each locus were derived by jackknifing over populations to determine if values differed significantly from zero. To identify which sampling locations contribute most to the differentiation observed, we also calculated FST between all pairwise combinations of locations.

We used two approaches to compare population differentiation in selected and putatively neutral SNPs. First, we simply compared FST values in the two categories of loci using nonparametric tests (Barbujani 1985). This approach indicates whether the two groups of loci differ but does not demonstrate that values for individual SNPs differ from neutral expectations. To do this, we used the simulation approach outlined in Beaumont and Nichols (1996). We plotted weighted FST values against heterozygosity for each locus examined. We then used the mean weighted FST values from our sample of putatively neutral sSNPs,
Table 1
Locations and Assay Methods for sSNPs on Chr 2 and 3

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>Gene</th>
<th>Change</th>
<th>Description</th>
<th>Uncut Product (bp)</th>
<th>Restriction Enzymes</th>
<th>Label</th>
<th>Notes</th>
<th>Oligos</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>619397</td>
<td>PFB0685c</td>
<td>G-A</td>
<td>Acyl-CoA synthetase</td>
<td>220</td>
<td>Apo1</td>
<td>6FAM</td>
<td></td>
<td>TCATACATACATAATATACTG TATATGATTCTTATCTCCATTA TCAATTTTTTTATGGTAAAT</td>
</tr>
<tr>
<td>2</td>
<td>649509</td>
<td>PFB0715w</td>
<td>C-T</td>
<td>DNA-directed RNA polymerase II second largest subunit</td>
<td>150</td>
<td>BamHI</td>
<td>HEX</td>
<td>*Made RE site</td>
<td>TGAATATCGCTAGTAGACACTGGCACTTTCTGTATA CAAAGGCTACTTTGGTATT</td>
</tr>
<tr>
<td>3</td>
<td>197185</td>
<td>PFC0180c</td>
<td>G-T</td>
<td>Hypothetical protein</td>
<td>138</td>
<td>XmnI</td>
<td>NED</td>
<td></td>
<td>AAGAAAAATGAAACTACAGCA TTTGTGTACTAAGGAGGTATAAACCAAAAGAT</td>
</tr>
<tr>
<td>3</td>
<td>221432</td>
<td>PFC0215c</td>
<td>C-G</td>
<td>Hypothetical protein</td>
<td>176</td>
<td>MlII</td>
<td>6FAM</td>
<td></td>
<td>ATTAAAGAGGAAATGACCCAAATATGCTCTTCCATAGTACATGGGAAATGGGAC</td>
</tr>
<tr>
<td>3</td>
<td>309640</td>
<td>PFO0295</td>
<td>T-G</td>
<td>40S ribosomal protein S12</td>
<td>104</td>
<td>BsrY1</td>
<td>NED</td>
<td>*Made RE site</td>
<td>AATTATTTTGATAGGAAGCACACTGTTAAAATCCATGAGGACCTTTCAGCAGTAATGGGAAATGGGAC</td>
</tr>
<tr>
<td>3</td>
<td>360148</td>
<td>PFC0350c</td>
<td>G-C</td>
<td>Putative T-complex protein eta subunit</td>
<td>207</td>
<td>AcII</td>
<td>NED</td>
<td></td>
<td>TCGAAGAACAAACATGATAATAATTACTTACTTACTCTGTTAAAATCCATGAGGACCTTTCAGCAGTAATGGGAAATGGGAC</td>
</tr>
<tr>
<td>3</td>
<td>454628</td>
<td>PFC0440c</td>
<td>T-A</td>
<td>Putative helicase</td>
<td>152</td>
<td>DraI</td>
<td>HEX</td>
<td></td>
<td>CCCAAATACATTATTCTAT CAAATTTGATCGTACGTAATGGGAAATGGGACCTTTCAGCAGTAATGGGAAATGGGAC</td>
</tr>
<tr>
<td>3</td>
<td>526792</td>
<td>PFC0530w</td>
<td>G-A</td>
<td>Hypothetical protein</td>
<td>175</td>
<td>DraI</td>
<td>HEX</td>
<td></td>
<td>AACAATACATTATTCTAT CAAATTTGATCGTACGTAATGGGAAATGGGACCTTTCAGCAGTAATGGGAAATGGGAC</td>
</tr>
<tr>
<td>3</td>
<td>679987</td>
<td>PFC0745c</td>
<td>G-A</td>
<td>Proteasome component C8</td>
<td>143</td>
<td>NsiI</td>
<td>HEX</td>
<td></td>
<td>AACCATATCAGATTTGATTGGTCAGTACGTAATGGGAAATGGGACCTTTCAGCAGTAATGGGAAATGGGAC</td>
</tr>
<tr>
<td>3</td>
<td>888282</td>
<td>PFC0940c</td>
<td>G-A</td>
<td>Hypothetical protein</td>
<td>122</td>
<td>HinFl</td>
<td>6FAM</td>
<td></td>
<td>GAGCATATTTTTGTAGATT GGACATATTTTTTTGATT GGTTTTGTTTGTATGCTTTG</td>
</tr>
</tbody>
</table>

NOTE.—All primers are listed 5’→3’. For each locus, the first and second primers are used in the primary reaction and the second and third primers are used in the secondary reaction. Primers listed third are end labeled with fluorescent tags. The primary reaction (25 μl) contained 2.5 μl of template, 2.5 μl of 10× buffer, 0.75 units of Taq polymerase (Takara, Otsu, Shiga, Japan), 3 mM Mg²⁺, and 0.1 pM of each primer. The secondary reaction (10 μl) contained 1 μl of template, 1 μl of 10× buffer, 0.35 units of Taq polymerase (Takara), 2.5 mM Mg²⁺, and 0.4 pM of each primer. Cycling conditions for both primary and secondary rounds were as follows: 2 min, 94°C; (30 s, 94°C; 30 s, 50°C; 30 s, 60°C) × 5 cycles, followed by (30 s, 94°C; 30 s, 45°C; 30 s, 60°C) × 25 cycles and an extension of 60°C for 2 min. Second round products were digested overnight using 0.8–1 unit of appropriate restriction enzyme. In some cases (marked by *), restriction enzyme sites were created by making a mismatch in 3–4 bp from the 3’ end of the unlabeled oligo in the second reaction. The digestion products were pooled together in the following combinations for capillary electrophoresis: (PFB0685c, PFB0715w, PFC0530w, PFC0180c, PFC015c, PFC0745c) and (PFC0440c, PFC0530w, PFC0940c, PFC0295). We also designed assays for 11 additional sSNPs on Chr 3 and 6 additional sSNPs on Chr 2. These were not polymorphic in initial inspection of 25 Southeast Asian isolates (Supplementary Table S1, Supplementary Material online).
together with sample size and population number information to seed coalescent simulations of genetic differentiation within metapopulations. Initially, we used an island model with 100 islands and repeated the procedure 50,000 times to generate 99% confidence intervals for neutral differentiation and estimated \( P \) values for departure of selected loci from these expectations. This analysis was conducted using the FDIST2 program (http://www.rubic.rdg.ac.uk/~mab/software.html). This approach is fairly robust to variation in mutation rate between loci, to sample size, and whether populations are in equilibrium or nonequilibrium (Beaumont and Nichols 1996). However, reducing the number of populations used in the simulations can influence the distribution of \( F_{ST} \), depressing it at high heterozygosities (Flint et al. 1999). We do not have a good idea of the appropriate number of islands in the Southeast Asian \textit{Plasmodium} population. We therefore reran simulations using 15 populations to examine the robustness of our conclusions to variation in this parameter.

Isolation by Distance

We investigated the fit of the data from each sSNP to isolation-by-distance models using Mantel tests. For these analyses, we measured distances between locations from maps and used the natural log of distance. We used the \( F_{ST}/(1 - F_{ST}) \) as our measure of genetic similarity (Rousset 1997). The significance of Mantel correlations was assessed by random permutation of the matrices 10,000 times and recalculation of test statistics. Tests were significant if the observed test statistics were within the top 5% of the distribution of test statistics measured in the permuted data sets. These analyses were performed using permutation tests implemented using POPTOOLS (http://www.cse.csiro.au/poptools/).

Results

Allele-Frequency Estimation

We analyzed a total of 755 malaria infections for 20 point mutations (10 nsSNPs and 10 sSNPs, with sample sizes ranging from 21 [Meuang Feuang, Laos] to 107 [Puj gang, Bangladesh]). The three different procedures used to estimate allele frequencies gave extremely similar values—these estimates are listed for each population in the Supplementary Table S3 (Supplementary Material online). The maximum difference between estimators was found in Ratchaburi (Thailand) where the predominant allele measure was 13% higher than both the maximum likelihood and single-allele estimates for the \textit{dhfr}-51. We measured correlations between the 260 allele-frequency estimates (20 loci \( \times \) 13 populations) derived using the three methods. For each comparison, the slopes were indistinguishable from 1, indicating minimal bias, and the correlation coefficients were >0.995 (fig. 2). We used the predominant allele method for most analyses reported here because it utilizes data from all people sampled and makes few assumptions. However, for the analysis of \textit{dhfr} and \textit{dhps}, where alleles are defined by multiple SNPs, we used the single-allele method because this allows unambiguous construction of haplotypes.

nsSNPs in Resistance Genes

Figure 3 summarizes allele frequencies observed at \textit{dhfr}, \textit{dhps}, \textit{pfcrt}, and \textit{pfmdr}. (1) \textit{pfcrt}: The K76T at \textit{pfcrt} critical for CQ resistance is found at frequencies of 0.58–1.00 in the populations examined. Eight of 13 populations examined had >95% frequency of this allele, and all four populations sampled from the Thailand-Burma border were fixed for this allele. Only three populations in Burma, Laos, and Vietnam showed high frequencies (>30%) of the wild-type allele. (2) \textit{pfmdr}: The N86Y mutation in \textit{pfmdr} is at low frequency in Thailand, Laos, Cambodia, and Vietnam (<30%) but at high frequency in one population from Western Burma (73%) and Bangladesh (61%--79%). (3) \textit{dhfr}: Data from 10 of the 13 populations have been presented previously (Nair et al. 2003). We found between
zero and four mutations in $dhfr$ alleles. The highest level of resistance alleles was found in Thai populations, where most parasites contain two to four mutations and the I164L mutation conferring high-level resistance to pyrimethamine is common (67%–100%).

$dhps$: We found between zero and four mutations in $dhps$ alleles. The highest level of resistance was found in Thailand, where most parasites carry three mutations and no wild-type parasites were found, while the lowest levels of resistance was found in neighboring Laos, where 66%–85% of parasites carried wild-type (sensitive) alleles.

sSNPs in Putatively Neutral Loci

Of the 62 sSNPs identified on Chr 3 by Mu et al. (2002), we screened 19 for variation in Southeast Asian samples. Of these, eight loci were variable (alternative alleles seen in at least one parasite) in the 25 Southeast Asian parasites screened. These were included in the full population survey. Of the 29 regions sequenced on Chr 2, in which 25mer Affymetrix probes suggested presence of SNPs (Volkman et al. 2002), 25 gave good sequence. Eighteen of 25 regions contained variable sites, including 14 sSNPs and 15 nsSNPs (Supplementary Table 3, Supplementary Material online). Of these, only two of eight sSNPs examined were polymorphic in the 25 Southeast samples screened. Six additional sSNPs on Chr 2 were not genotyped because restriction digest assays could not be designed or they were only variable in the 3D7 sequence. We genotyped all 10 sSNPs (eight on Chr 3 and two on Chr 2) in 755 infections. One of these loci (PFB0685c) had rare
minor alleles with maximal allele frequencies of 6% in the 13 populations. Eight of the 10 sSNPs showed weak but significant differentiation between populations (fig. 4) with average $F_{ST}$ of 0.029 (range 0–0.14). Much of the differentiation is explained by populations from Bangladesh and Laos. Parasites from these countries are significantly differentiated from those collected from other countries in the region (fig. 5).

Comparison of Selected and Putatively Neutral SNPs

$F_{ST}$ values for all nsSNPs in resistance genes were higher than those for the sSNPs examined (fig. 4). The differences between these two groups of SNPs are highly significant (Mann-Whitney test: $Z = -3.78$, $P < 0.0002$). The four mutations within both $dhfr$ and $dhps$ are not independent and may elevate the significance of the differences observed. To avoid this, we reanalyzed these data using the mean $F_{ST}$ over all four polymorphic sites within both $dhfr$ and $dhps$, but the result is still highly significant ($Z = -2.83$, $P < 0.0047$). To evaluate whether individual SNPs show patterns of genetic differentiation suggestive of selection, we used coalescent simulations to generate the distribution of expected values for neutral loci using FDIST2 (Beaumont and Nichols 1996). To seed these simulations, we used the mean weighted $F_{ST}$ values for the 11 sSNPs typed and 15 or 100 populations. In both cases, all 10 nsSNPs within resistance genes show strong deviations from neutral expectations with $P$ values ranging from $8 \times 10^{-5}$ to $<10^{-6}$ (table 2). The $P$ values for these deviations were derived by comparing observed results with the distributions derived from 50,000 coalescent simulations. To investigate the robustness of this result, we also ran simulations seeded with an $F_{ST}$ value 2 standard deviations (SDs) greater than that observed to seed the coalescent simulations. This suggests that any inaccuracies in our measure of $F_{ST}$ due to sampling of putatively neutral SNPs are not independent and may elevate the significance of the differences observed. To avoid this, we reanalyzed these data using the mean $F_{ST}$ over all four polymorphic sites within both $dhfr$ and $dhps$, but the result is still highly significant ($Z = -2.83$, $P < 0.0047$).

**Discussion**

**Efficacy of $F_{ST}$ for Detecting Selection**

We found dramatic differences in geographical structure between nsSNPs in drug resistance genes and sSNPs in putatively neutral loci. While this result was expected, the strength and consistency of the differences observed were not. All 10 nsSNPs within the four resistance genes show higher $F_{ST}$ than the 10 sSNPs, and simulations suggest that the $F_{ST}$ values observed deviate significantly from neutral expectations with $P$ values for all comparisons below $8 \times 10^{-5}$. Eight of 10 nsSNPs remain as outliers even when we use an $F_{ST}$ value 2 SDs greater than that observed to seed the coalescent simulations. This suggests that any inaccuracies in our measure of $F_{ST}$ due to sampling of putatively neutral sSNPs is unlikely to explain the patterns observed. These empirical results provide strong support for recent simulation studies that demonstrated efficient detection of loci under directional selection, with few false-negative

![Diagram](image_url)
or false-positive results when selection coefficients were
>0.1 (Beaumont and Balding 2004). We also observed that
the combined sSNPs fitted a model of isolation by distance
better than the combined nsSNPs. However, many individ-
ual loci from both groups showed poor fit to isolation-by-
distance models, and differences between the nsSNPs and
sSNPs were not significant. Hence, analyses of patterns of
differentiation (isolation by distance) were less useful for
discriminating selected loci than analyses of the extent
of differentiation (measured using FST) in this system.

Two features of the drug resistance may contribute to
the distinctive differences in the extent of population struc-
ture revealed by nsSNPs in resistance loci and putatively
neutral sSNPs. First, selection is strong. Selection co-
efficients driving antimalarial drug resistance have been es-
timated at between 0.03 and 0.3, based on allele-frequency
change over time (Nair et al. 2003; Roper et al. 2003;
Anderson 2004). Selection coefficients are expected to
be particularly high in many Southeast Asian countries,
where most infections are symptomatic and a high pro-
portion of infections is exposed to drug treatment
(Luxemburger et al. 1996, 1997). However, while the selec-
tion coefficients driving differentiation in allele frequencies
at drug resistance loci are strong, they are by no means
atypical of selection coefficients measured in many natural
populations (Endler 1986; Hoekstra et al. 2001). Second,
drug resistance has evolved very recently in malaria para-
sites. The clarity of the differences in population structure
observed in the two groups of SNPs may in part reflect the
fact that drug resistance mutations are not in migration-drift
equilibrium, and allele frequencies have not yet been ho-
mogenized by migration. These data provide strong empir-
ical support for the idea that genome regions containing
selected loci could be located by comparing FST values
(Lewontin and Krakauer 1973).

In this study, we examined 13 parasite populations.
Two of the drug resistance polymorphisms studied
(pfdmr-56 and pfcrt-76) show homogeneous distribution
in the majority of populations examined and their high
FST values result from distinctive allele frequencies in just
a few of the populations. In the case of pfcrt, for example,
resistance allele frequencies are >95% in eight of the
populations examined, and wild-type alleles are found at
frequencies >30% in only three parasite populations. It
would be prohibitively expensive to genotype parasites
from all 13 populations in a genome-wide FST screen.

or false-positive results when selection coefficients were
>0.1 (Beaumont and Balding 2004). We also observed that
the combined sSNPs fitted a model of isolation by distance
better than the combined nsSNPs. However, many individ-
ual loci from both groups showed poor fit to isolation-by-
distance models, and differences between the nsSNPs and
sSNPs were not significant. Hence, analyses of patterns of
differentiation (isolation by distance) were less useful for
discriminating selected loci than analyses of the extent
of differentiation (measured using FST) in this system.

Two features of the drug resistance may contribute to
the distinctive differences in the extent of population struc-
ture revealed by nsSNPs in resistance loci and putatively
neutral sSNPs. First, selection is strong. Selection co-
efficients driving antimalarial drug resistance have been es-
timated at between 0.03 and 0.3, based on allele-frequency
change over time (Nair et al. 2003; Roper et al. 2003;
Anderson 2004). Selection coefficients are expected to
be particularly high in many Southeast Asian countries,
where most infections are symptomatic and a high pro-
portion of infections is exposed to drug treatment
(Luxemburger et al. 1996, 1997). However, while the selec-
tion coefficients driving differentiation in allele frequencies
at drug resistance loci are strong, they are by no means
atypical of selection coefficients measured in many natural
populations (Endler 1986; Hoekstra et al. 2001). Second,
drug resistance has evolved very recently in malaria para-
sites. The clarity of the differences in population structure
observed in the two groups of SNPs may in part reflect the
fact that drug resistance mutations are not in migration-drift
equilibrium, and allele frequencies have not yet been ho-
mogenized by migration. These data provide strong empir-
ical support for the idea that genome regions containing
selected loci could be located by comparing FST values
(Lewontin and Krakauer 1973).

In this study, we examined 13 parasite populations.
Two of the drug resistance polymorphisms studied
(pfdmr-56 and pfcrt-76) show homogeneous distribution
in the majority of populations examined and their high
FST values result from distinctive allele frequencies in just
a few of the populations. In the case of pfcrt, for example,
resistance allele frequencies are >95% in eight of the
populations examined, and wild-type alleles are found at
frequencies >30% in only three parasite populations. It
would be prohibitively expensive to genotype parasites
from all 13 populations in a genome-wide FST screen.
Table 2
Significance of Deviations from Neutral Expectations in Drug Resistance Loci

<table>
<thead>
<tr>
<th>Locus</th>
<th>H2</th>
<th>FST</th>
<th>15pops</th>
<th>100pops</th>
<th>Mean FST + 2 SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>pf405</td>
<td>0.098</td>
<td>0.038</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
</tr>
<tr>
<td>dhfr540</td>
<td>0.001</td>
<td>0.003</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
</tr>
<tr>
<td>dhfr59</td>
<td>0.001</td>
<td>0.004</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>10^-6</td>
</tr>
<tr>
<td>dhfr108</td>
<td>0.001</td>
<td>0.004</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>2 x 10^-6</td>
</tr>
<tr>
<td>dhfr118</td>
<td>0.001</td>
<td>0.004</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>2 x 10^-6</td>
</tr>
<tr>
<td>dhfr164</td>
<td>0.001</td>
<td>0.004</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>2 x 10^-6</td>
</tr>
<tr>
<td>dhfr51</td>
<td>0.001</td>
<td>0.004</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>2 x 10^-6</td>
</tr>
<tr>
<td>dhfr540</td>
<td>0.001</td>
<td>0.004</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>2 x 10^-6</td>
</tr>
<tr>
<td>dhfr59</td>
<td>0.001</td>
<td>0.004</td>
<td>&lt;10^-5</td>
<td>&lt;10^-5</td>
<td>2 x 10^-6</td>
</tr>
</tbody>
</table>

Note.—The table shows the weighted mean heterozygosity and FST and P values for deviation from neutral expectations for each of the nsSNPs within drug resistance loci. The coalescent simulations used to derive these P values were based on average FST values observed for the 10 sSNPs, sample sizes of 50, and either 15 or 100 subpopulations within a metapopulation. We also simulated expectations using FST values 2 SD greater than that observed in our sample of sSNPs to investigate the robustness of these results. Significant (P < 0.05) results are shown in bold text.

For a genome-wide comparison of FST, examination of smaller numbers of populations for which we suspect differences in selection regimes may be more practical. Comparisons between neighboring countries of Laos and Thailand may be particularly informative because these neighboring countries show the most marked differences in allele frequencies at loci encoding resistance. In Southern Laos, allele frequencies in drug resistance genes are consistently lower than in Thailand for all four loci, suggesting strong local adaptation in these neighboring countries. We also have independent evidence that drug selection is weaker in Laos than in Thailand because hitchhiking events associated with resistance to both pyrimethamine and CQ affect larger genome regions in Thailand than in Laos (Nash et al. 2005).

Some caveats are important to mention. FST-based approaches for detecting selection rely strongly on the assumption that selected sites (or SNPs in LD with selected sites) are rare relative to neutrally evolving sites and therefore show up as outliers. This may be broadly true in many free-living organisms. However, it is questionable whether this is the case for P. falciparum. Recent studies suggest that a high proportion of polymorphisms are under balancing selection in this organism as a result of immune selection. An oligo microarray study on Chr 2 of Plasmodium showed that 69% of all polymorphisms detected were in membrane-associated genes in telomeric and subtelomeric regions and that polymorphism is rather rare elsewhere in the genome (Volkman et al. 2002). The highly polymorphic genes included Var genes involved in antigenic variation (Bull et al. 1998; Newbold 1999), vaccine targets such as pf230 (Bustamante et al. 2000; Williamson 2003), and merozoite antigens (M. K. Hughes and A. L. Hughes 1995) for which there is independent evidence of positive selection from sequence comparisons. Furthermore, many other polymorphisms were found in unknown genes that may also be surface expressed and exposed to immune selection (Volkman et al. 2002). Random sampling of point mutations from the P. falciparum genome using a technique such as amplified fragment length polymorphism (AFLP), as used in recent studies of free-living organisms (Wilding, Butlin, and Grahame 2001; Campbell and Bernatchez 2004), would tend to sample subtelomeric SNPs that are predominantly from antigen genes under balancing selection. This would result in underestimates of FST for randomly sampled SNPs presumed to represent neutrality and a consequent overestimation of the numbers of outlier SNPs that are identified as being under selection. We expect that this problem will apply to other parasites and pathogens in which many proteins are under balancing or frequency-dependent selection due to immune pressure.

The opposite bias—overestimation of FST for randomly sampled SNPs presumed to represent neutrality—could also generate bias in studies of free-living organisms if a substantial proportion of loci are locally adapted. In this case, this would lead to underestimation of the numbers of loci inferred to be under positive selection. For example, a large survey of human SNP variation showed higher FST for nsSNPs than sSNPs, suggesting that many amino acid changes may be locally adapted (Hinds et al. 2005).
Neutral Population Structure

We observed weak but significant population structure for 8 of 10 loci examined. We have previously examined population structure on a global scale using microsatellite markers (Anderson et al. 2000). This study revealed a spectrum of population structures with strong population structure in South America (\(F_{ST} = 0.4\)) and minimal differentiation between African populations. The putatively neutral sSNPs examined in this study reveal low \(F_{ST}\) (0–0.14) but significant genetic structure for 8 of 10 loci and a good fit to isolation by distance \((P < 0.0001)\). The level of population structure is intermediate between that observed in South America and Africa, consistent with our previous data (Anderson et al. 2000). The weak but significant population structure observed in the sSNPs caution against combining parasite isolates collected from different regions of Southeast Asia for use in association analyses or candidate gene studies because population structure can generate spurious associations (Pritchard et al. 2000; Pritchard and Donnelly 2001). Contrasting patterns of isolation by distance have recently been documented in malaria parasite populations from Brazil (no evidence for isolation by distance) (Machado et al. 2004) and Indonesia (strong isolation by distance) (Anthony et al. 2005). The Southeast Asian populations examined here show weaker isolation by distance than seen in the Indonesian study.

The low \(F_{ST}\) results suggest extensive gene flow between Southeast Asia parasite populations. Gene flow of parasites may result from movement of both human and mosquito hosts. Direct measures of human traffic reveal that ~20 million people per year travel across international borders between China, Burma, Laos, and Vietnam (Hu et al. 1998). Actual numbers are probably much higher,
if we include people moving illegally between countries. We have no direct measures of mosquito migration, but indirect estimates can be assessed using population genetic approaches. For example, mtDNA and microsatellite analyses of Anopheles dirus species complex suggest little structure within each of the vector species A and D within Thailand (Walton et al. 2001). However, while these results suggest high levels of movement between populations, they could also be explained by rapid population expansion (Walton et al. 2000).

Distribution of Drug Resistance Alleles and Patterns of Drug Selection

Heterogeneity in allele frequencies among locations may result from both selection for resistance alleles in the presence of drug selection and/or selection against resistance alleles in the absence of drug selection. The highest frequencies of resistance alleles in \( dhfr \), \( dhps \), and \( pfcr \) are seen in populations on the Thailand-Burma border. This is consistent with the fact that malaria is restricted to a few border regions of Thailand, where treatment regimes are carefully optimized and treatment coverage is good. In comparison, in neighboring countries such as Laos and Burma, health infrastructure is weaker, malaria is more widespread, and treatment regimes are frequently suboptimal utilizing CQ and SP which are relatively ineffective. Neither CQ nor SP has been used for treating \( P. falciparum \) in Thailand for 25 years, although CQ is still used widely for treatment of \( Plasmodium \) \( vivax \) infections and trimethoprim-sulfamethoxazole (an antifol-sulfa combination with activity similar to SP) is still widely used as an antibiotic throughout the region. The high frequency of mutations in \( pfcr \), \( dhfr \), and \( dhps \) is surprising because resistance alleles at two of these loci (\( pfcr \) and \( dhfr \)) carry fitness costs. In the case of \( pfcr \), evidence for fitness costs comes from three sources. First, removal of CQ pressure results in decline in resistance allele frequencies in Malawi (Kublin et al. 2003), Hainan Island (Wang et al. 2005), and Vietnam (T. T. Hien, unpublished data). Second, parasites carrying back mutations in residue 76 have been found in nature (Fidock et al. 2000). Finally, laboratory selection of mutations in residue 76 results in parasites with reduced growth rates (R. Cooper, personal communication). Similarly, for \( dhfr \), enzyme manufactured from mutant alleles shows weaker binding to substrate than wild-type enzyme (Sirawaraporn et al. 1997), and yeast transfected with resistant alleles show lower growth rates than parasites transfected with wild-type alleles (Cortese and Plowe 1998).

Selection pressures acting on \( pfmdr \) are thought to be particularly complex, with both point mutations (Reed et al. 2000; Djimde et al. 2001) and copy number changes playing important roles (Price et al. 2004). Furthermore, while CQ selects for mutant alleles and low copy number, MFQ, QN, and artemisinin derivatives act in the opposite direction, selecting for wild-type alleles and copy number amplification. We genotyped the N86Y mutation in \( pfmdr \). This mutation plays an important role in CQ drug resistance in African countries (Hayton and Su 2004), although its role in Southeast Asia is less well supported. We see low levels of the N86Y in all locations, except for Bangladesh and one Burmese population. The differentiation that we see at N86Y may not reflect direct selection on this SNP. Instead, differentiation may result from LD between this SNP and copy number amplification. Recent data suggest that \( pfmdr \) copy number plays the predominant role in determining resistance to MFQ in Southeast Asia (Price et al. 2004). Consistent with this, real-time assays show that 34% of samples from the Thailand-Burma border carry \( >1 \) copy of \( pfmdr \), while in neighboring Laos all 80 infections examined carried a single copy (Anderson et al., unpublished data).

The geographical distribution of resistance alleles has important practical implications for resistance management. Mapping of resistance alleles provides a powerful approach to monitoring spread of resistance and allows rational choices to be made about treatment policy. The efficacy of this approach to resistance mapping is dependent on the spatial scale over which allele frequencies vary. In this study, we examined 13 populations including four populations from Thailand, three from Laos, and two from both Burma and Bangladesh. We found homogeneous allele frequencies within countries for alleles at loci (\( dhfr \) and \( dhps \)) underlying antifolate resistance, with most variation observed in between-country comparisons. These data suggest that low-density sampling of parasite populations may provide sufficient information for developing rational antimalarial policy. On the other hand, \( pfcr \) polymorphism is more heterogeneous within countries and is less encouraging. The K76T has a patchy distribution, with locations situated \(<200\) miles apart showing \( \geq 30\% \) difference in allele frequency. In this case, more intensive sampling may be required to evaluate countrywide resistance levels.

Supplementary Material

 Supplementary Tables S1–S3 are available at Molecular Biology and Evolution online (http://www.mbe. oxfordjournals.org/).

Acknowledgments

Jianbing Mu provided details of Chr 3 SNP locations, while Elizabeth Winzeler provided unpublished details from her group hybridization study of Chr 2 SNPs. The following people assisted in sample collection: Lucy Phaipun, Alan Brockman, and the staff of the Shoklo Malaria Research Unit; Samlane Phompida, Rattanaxay Phetsouvanh, Maniphone Khanthavone, and the staff of the Centre for Malaria, Parasitology, and Entomology, Vientiane, Laos; Wim Van Damme, Louis Goubert, Nuong Sophy, and Louis Goubert (Médecins Sans Frontières, Cambodia); staff at Médecins Sans Frontières-Holland (Myanmar); and in Vietnam, the Wellcome Unit, Hospital for Tropical Diseases, Ho Chi Min City. This investigation was conducted in a facility constructed with support from Research Facilities Improvement Program Grant Number C06 RR-013556 from the National Center for Research Resources at National Institutes of Health (NIH) and supported by NIH grant AI48071 (T.J.C.A.). The SMRU is part of the Wellcome-Trust Mahidol University-Oxford Tropical Medicine Research Program supported by the Wellcome Trust of Great Britain. F.N., P.N., M.M., and N.W. are supported by the Wellcome Trust.
Literature Cited


Edward Holmes, Associate Editor

Accepted July 25, 2005