Supplementary Information

Methods

Microscopy. Smears from participants were evaluated prior to enrollment and scored on a scale of 0 to 4+ by two microscopists on site. The score is based upon the following scale: 0 is not infected, 1+ is 1-9 rings per 100 microscope fields, 2+ is 10-100 rings per 100 fields, 3+ is 1-10 rings per field, and 4+ is >10 rings per field. Participants were required to have a score of 2+ or greater for enrollment.

Amplification and Sequencing of *csp*. DNA from filter paper blood spots was extracted using the Invitrogen Pro 96 Genomic DNA kit (Invitrogen, Carlsbad, CA). The region of *csp* containing the TH2 and TH3 epitopes was amplified using previously described primers [1], which we modified for 454 sequencing by inclusion of a linker, tag and a Multiplex Identifier (MID) sequence. The samples were amplified on an Eppendorf Master cycler (Eppendorf, Haupaugge, NJ) under the conditions previously described using Roche FastStart High fidelity Taq (Roche, Madison, WI) [1]. PCR amplicons were purified using the Purelink PCR purification kit (Invitrogen). Final quality (OD > 1.8) was checked and concentration determined using a Nanodrop 1000 spectrophotometer (Thermo Scientific, Waltham, MA). Amplicons were pooled and sequenced on a 454 Life Sciences sequencer using the Titanium chemistry at the UNC High Throughput Sequencing Facility.

Definition of TH2 and TH3 epitopes. Within our amplicon, TH2 was defined as nt121-156 and TH3 was defined as nt223-258. These correspond to nt946-981 and nt1048-1083 in strain 3d7 (PFC0210c), as well as nt1068-1103 and nt1170-1205 in strain 7g8 (K02194.1).

Haplotype determination from ultra-deep sequencing. Sequence, flow cell intensities, and base quality scores were extracted from the sff files using the program sffinfo (454 Life Sciences). An in-house Perl program pyro_tools was used to process the raw sffinfo text output sorting and return high quality sequences for haplotype prediction. In this program, we first identified and removed the tag, MID and forward primer, requiring all to exactly match without error. Based on the MID and plate location, reads were sorted into the distinct amplicons (the 2 PCR amplifications per participant). As 454 sequencing error rate increases over the length of the read, we sought to minimize the inclusion of poor quality sequences by concurrently trimming low quality sequence from the 3'-end of reads. Low quality sequence was determined by two measures: the default trim position as defined by the 454 base calling software and by direct examination of the underlying flow intensities. For the latter, the trim position was the third instance of a noisy fluorescent signal intensity, which was/were defined as flows with intensities between 0.4 and 0.7 or between 1.2 and 1.6 [2]. We also trimmed the reads to remove the reverse primer sequence identified with blast2seq (National Center for Biotechnology Information, NCBI). Finally, we required that trimmed reads represent at least 200 bases of the amplified region in order to ensure that each read provided adequate haplotypic information to facilitate accurate ShoRAH prediction. Combined this filtering produced high-quality read sets representing individual PCR amplicons (2 per participant). These were further analyzed by ShoRAH (**Sho**rt **R**ead **A**ssembly into **H**aplotypes) to predict the most likely haplotypes within the patient [3]. ShoRAH is a Bayesian model treating reads as discrete samples from a sequencing process which is error prone. A local analysis was performed using this software to correct for sequencing errors by clustering all reads that overlap the same region of the genome of length approximately equal to the read length [3]. The consensus sequence of each cluster represents the true haplotype from which the erroneous reads are predicted to emanate [3]. The number of reads associated with the cluster estimates the prevalence of the haplotype in the population [3]. We removed improbable haplotypes (ShoRAH posterior probability <0.9) and further refined the number of reads representing each high-probability haplotype by assigning each read to its most similar haplotype based on global optimal pairwise alignment (Needleman-Wunsch algorithm as implemented in the program *needleall* in EMBOSS suite [4].

ShoRAH is limited in that it models a uniform error rate across the sequence. This leads to spurious haplotypes due to differences in 3' sequence—particularly in terms of indels. To correct for this, we heuristically clustered the predicted ShoRAH haplotypes for each participant (the combined haplotypes from the two independent amplicon). Clustering was based on differences determined by pairwise global alignments between haplotypes. As haplotypes were clustered, further pairwise alignments were then based on the consensus sequence as determined from a *clustalw* multiple alignment [5] weighted for the number of reads represented by each ShoRAH haplotype within a given cluster. The clustering was done in a stepwise manner allowing for increasing degree of differences reasoning that the vast majority of errors would separate sequences from the true haplotype by only a few differences. Clustering proceed from the smallest (based on the number of reads) to largest. In the case of a cluster that was equally distant from two or more clusters, the assignment was to the largest cluster. We allowed up to a single substitution and five small insertion/deletions of up to two bases, which would be biologically implausible in this indispensible/required gene, as they would result in frame shifts. As each sample was amplified in duplicate and sequenced independently, we required that the final haplotype cluster be composed of initial ShoRAH haplotypes from both amplicons. The halplotype clusters were also required to represent ≥1% of the total reads for a participant.

To examine haplotypes at the population level, heuristic clustering and consensus determination was performed as above across the combined haplotypes from all individuals excepting that substitutions were not allowed (only small indels). The vast majority differences were due to one or two small indels. The resulting weighted consensi provided the final haplotypes for analysis and was assigned a unique population identifier (pUID).

Data Analysis. The final haplotypes were stored, managed and analyzed in Microsoft Access 2007 and Microsoft Excel 2007 (Microsoft, Seattle, WA). DNA alignments and figures were generated using MegAlign and GeneVison software (DNAStar, Madison, WI). Additional figures were generated using Graphpad Prism v5 (GraphPad Software Inc., La Jolla, CA). Ecological indexes of diversity and rarefaction curves were determined using EstimateS v8.2 [6]. The rarefaction curves were made using the Mao Tao estimator as described in EstimateS [6-7]. Calculations of molecular diversity and evolution were done using Arlequin v3.5.1.2 and DnaSP v5.0 [8-9]. DnaSP was used to calculate Fu and Li D*, Fu and Li F*, and Tajima D. Arlequin was used to determine mean pairwise differences, theta estimators of molecular diversity, allele frequencies, expected heterozygosity, inter-haplotypic distance matrices and nucleotides under selection. Nucleotides under selection were detected using coalescent simulations to get p-values of locus specific F-statistics conditioned on observed levels of heterozygosity [9-10]. Since a single population structure was used (total parasite population), a non-hierarchical finite island model was used with 20,000 simulations [9]. The Median-Joining Network was created using DNA Alignment v1.2.1.1 and Network v4.6.0.0 [11].

For this study, we were primarily interested in the diversity of CS. Therefore, multiplicity of infection (MOI) has been defined as the number of different CS variant contained within an individual infection. This may be an under representation of the true MOI for two reasons. First, CS is not as highly diverse as other surface antigens traditionally used for studying diversity, such as merozoite surface protein-2. Second, single locus genotyping has the potential to under represent diversity due to variants haring a similar genotype at the locus studied, which are divergent at additional sites.

References

1. Alloueche A, Silveira H, Conway DJ, et al. High-throughput sequence typing of T-cell epitope polymorphisms in Plasmodium falciparum circumsporozoite protein. Mol Biochem Parasitol **2000**; 106:273-82.

2. Quinlan AR, Stewart DA, Stromberg MP, Marth GT. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods **2008**; 5:179-81.

3. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics **2011**; 12:119.

4. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **2000**; 16:276-7.

5. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **1994**; 22:4673-80.

6. Colwell RK. EstimateS: Statistical estimation of species richness and shared species from samples. Available at: http://viceroy.eeb.uconn.edu/estimates. 2011.

7. Colwell RK, Mao CX, Chang J. Interpolating, extrapolating and comparing incidence-based species accumulation curves. Ecology **2005**; 85:2717-2727.

8. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **2009**; 25:1451-2.

9. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour **2010**; 10:564-7.

10. Excoffier L, Hofer T, Foll M. Detecting loci under selection in a hierarchically structured population. Heredity **2009**; 103:285-98.

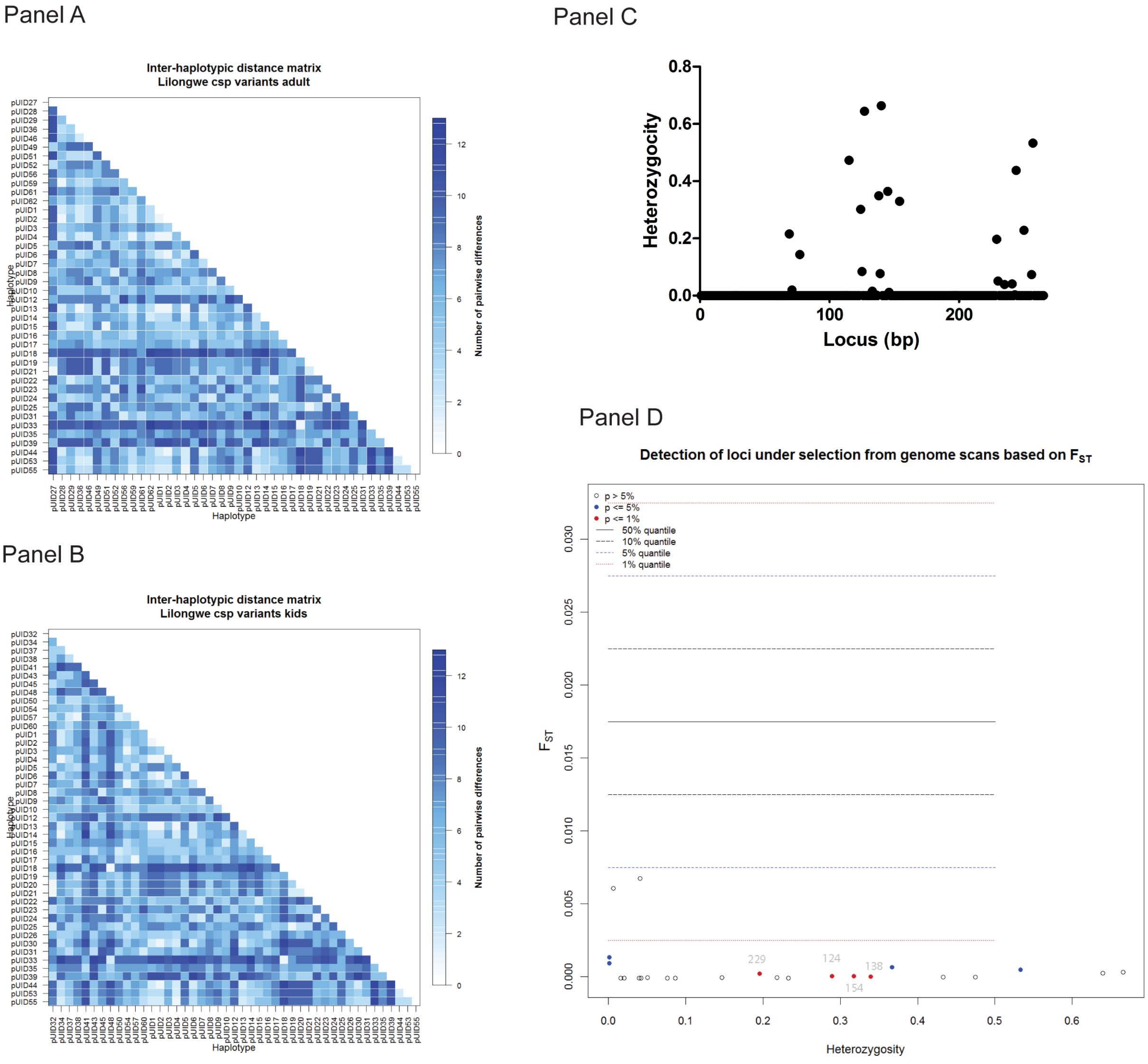
11. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol **1999**; 16:37-48.

Figure S1. Genetic Variation and Selection of *csp* Haplotypes in Adults and Children. Panel A. shows the relative genetic distance between haplotypes found in adult participants in the study. The figure represents the number of pairwise differences between all variants found in the population. Differences between variants appear to be diffuse among the population, with no specific variants being more distant than others. Panel B. show a similar figure for variants found in children. Panel C shows the expected heterozygosity (H_e) for the 24 polymorphic loci identified in the parasite population. Panel D shows loci under selection from genome scans based on F_{st}. Using Arlequin and based on the methods described by Excoffier, there were only 8 sites showing evidence of selection [9-10]. Four sites (loci 124, 138, 154 and 229, red dots) all had p-values of <1%. The other 4 sites (136, 145, 243 and 257, blue dots) all had p-value<5%. These loci correspond to nucleotides 949, 961, 963, 970, 979, 1054, 1068 and 1082 in the 3d7 strain of falciparum (PFC0210c). Of these, 5 loci fall in the TH2 epitope (nucleotides 114-155) and three fall in the TH3 epitope (nucleotides 227-258).

Table S1. Allele Frequency of Polymorphic Sites

Nucleotide	# of Alleles	Allele Frequency			
69	2	T : 0.8771	A : 0.1229		
71	2	C : 0.9902	G : 0.0098		
77	2	A : 0.9220	G : 0.0780		
115	2	C : 0.3830	A : 0.6170		
124	2	G : 0.8147	A : 0.1853		
125	2	A : 0.9560	C : 0.0440		
127	3	A : 0.3886	C : 0.3993	G : 0.2121	
133	2	T : 0.9921	A : 0.0079		
136	2	A : 0.9990	C : 0.0010		
138	2	G : 0.7749	C : 0.2251		
139	2	A : 0.9599	G : 0.0401		
140	4	T : 0.1135	C : 0.3822	A : 0.4110	G : 0.0933
145	2	C : 0.7602	A : 0.2398		
146	2	A : 0.9942	G : 0.0058		
154	2	C : 0.7917	A : 0.2083		
229	2	A : 0.8891	G : 0.1109		
230	2	A : 0.9740	G : 0.0260		
235	2	C : 0.9801	T : 0.0199		
241	2	G : 0.9791	A : 0.0209		
243	2	C : 0.9986	A : 0.0014		
244	2	C : 0.6761	G : 0.3239		
250	2	G : 0.8687	A : 0.1313		
256	2	G : 0.9619	A : 0.0381		

257	3	C : 0.5190	A : 0.4429	T : 0.0381	


Panel A
Inter-haplotypic distance matrix
Lilongwe csp variants adult
Haplotype
pUID27
pUID28
pUID29
pUID36
pUID46
pUID49
pUID51
pUID52
pUID56
pUID59
pUID61
pUID62
pUID1
pUID2
pUID3
pUID4
pUID5
pUID6
pUID7
pUID8
pUID9
pUID10
pUID12
pUID13
pUID14
pUID15
pUID16
pUID17
pUID18
pUID19
pUID21
pUID22
pUID23
pUID24
pUID25
pUID31
pUID33
pUID35
pUID39
pUID44
pUID53
pUID55
Number of pairwise differences
0
2
4
6
8
10
12
Panel B
Inter-haplotypic distance matrix
Lilongwe csp variants kids
Haplotype
pUID32
pUID34
pUID37
pUID38
pUID41
pUID43
pUID45
pUID48
pUID50
pUID54
pUID57
pUID60
pUID1
pUID2
pUID3
pUID4
pUID5
pUID6
pUID7
pUID8
pUID9
pUID10
pUID12
pUID13
pUID14
pUID15
pUID16
pUID17
pUID18
pUID19
pUID20
pUID21
pUID22
pUID23
pUID24
pUID25
pUID26
pUID30
pUID31
pUID33
pUID35
pUID39
pUID44
pUID53
pUID55
Number of pairwise differences
0
2
4
6
8
10
12
Panel C
Heterozygocity
0.0
0.2
0.4
0.6
0.8
Locus (bp)
0
100
200
Panel D
Detection of loci under selection from genome scans based on F_{ST}
○ p > 5%
● p <= 5%
● p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
F_{ST}
0.000
0.005
0.010
0.015
0.020
0.025
0.030
229
124
138
154
Heterozygosity
0.0
0.1
0.2
0.3
0.4
0.5
0.6