Supporting Information

Takala-Harrison et al. 10.1073/pnas.1211205110

SI Text

Concordance Between Genotype Calls from Three Single-Nucleotide Polymorphism-Calling Algorithms. Three algorithms were assessed to call single-nucleotide polymorphisms (SNPs) from normalized nucleotide signal intensities: (*i*) GeneChip Targeted Genotyping Analysis Software (GTGS; www.affymetrix.com/), (*ii*) Illuminus (1), and (*iii*) a heuristic algorithm based on discrete cutoffs of signal intensity strength and contrast. SNP calls from the three algorithms were >90% concordant [median concordance: 91.2% (GTGS vs. heuristic), 90.6% (GTGS vs. Illuminus), and 94.5% (heuristic vs. Illuminus)]. Most discordant calls resulted from SNPs that were not called by one algorithm (undetermined) but were called by another algorithm [median: 8.1% (GTGS vs. heuristic), 8.8% (GTGS vs. illuminus), and 2.9% (heuristic vs. illuminus)]. Very rarely was an allele called by one algorithm and the alternative allele called by another algorithm (median: 0% for all three pairwise algorithm comparisons). SNP calls from the heuristic algorithm were used in the analysis because this algorithm allows for heterozygous calls at a wider range of allele frequencies than algorithms designed to genotype diploid individuals, and may therefore more accurately represent mixtures of haploid organisms observed in *Plasmodium falciparum* field samples.

Thresholds for Undetermined and Heterozygous Calls. Uninformative SNPs that were invariant or had very low minor-allele frequency (<1%) were excluded from the analysis (4,420 SNPs), as were SNPs with an extreme number of undetermined or heterozygous calls (832 additional SNPs). Thresholds for the proportion of missing and heterozygous calls were based on inflection points in plots of the proportion of missing or heterozygous calls per SNP. The results of the data analysis were robust to the thresholds chosen. The thresholds for undetermined calls and heterozygous calls are designated in Fig. S2. SNPs with greater than 15% of calls undetermined were excluded, as were SNPs with heterozygous calls greater than 15%. The final number of SNPs included in the analysis was 2,827 (Fig. S3), including 752 intergenic SNPs. Samples with an extreme number of undetermined SNP calls were also excluded from the analysis (Fig. S2). Samples with greater than 20% undetermined SNP calls were excluded. The final number of samples included in the analysis was 327.

Accounting for Confounding Due to Population Structure. We assessed two approaches to account for population structure in statistical models. First, we included the first four principal components and other covariates as fixed effects in linear regression models [principal components analysis (PCA) approach]. Second, we used linear mixed-regression models with covariates treated as fixed effects and a genetic similarity matrix included as a random effect to account for the lack of independence among genetically similar parasites [efficient mixed-model association (EMMA) approach] (2). The first four principal components accounted for 65% of the genetic variation; however, the PCA approach still resulted in significant residual inflation of *P* values (Fig. S6*A*), whereas the EMMA approach showed little remaining effect of population structure (Fig. S6*B*; Fig. 3). We therefore used the EMMA approach for the analysis.

Distribution and Heritability of Dihydroartemisinin IC_{50}. The distribution of dihydroartemisinin (DHA) IC_{50} values of parasites collected from participants in the artesunate clinical trials is shown in Fig. S4*A*. Comparison of DHA IC_{50} values was complicated by the use of different methods to assess IC_{50} at the different study sites. At the Bandarban, Bangladesh, and Tasanh, Cambodia, study sites, DHA IC_{50} was assessed ex vivo on fresh samples using a histidine-rich protein 2 (HRP2)-based assay. At the Pailin, Cambodia, and Wang Pha, Thailand, sites, DHA IC_{50} was determined from frozen isolates using a tritiated hypoxanthine-based assay. IC_{50}s from the same control strains were not determined at each site, preventing normalization across sites. There was little overlap in IC_{50}s between the Tasanh and Bandarban sites (Fig. S4*A*), with parasites from Tasanh having higher DHA IC_{50}s than parasites from Bandarban. Parasites from the Pailin and Wang Pha showed intermediate DHA IC_{50}s, with parasites from Pailin having greater IC_{50}s than parasites from Wang Pha. The differences in DHA IC_{50}s between the two Cambodian sites were likely due to the use of different methods to assess this phenotype.

Because different methods were used to assess DHA IC_{50} at the two Cambodian study sites, heritability (H^2) for this phenotype was assessed for each site separately (Fig. S4*B*). Without adjustment for confounding variables, DHA IC_{50} did not show significant heritability in identical clones from either Cambodian study site (Tasanh: $H^2 = 0.025$, $F = 1.11$, $P = 0.37$; Pailin: $H^2 = -0.13$, $F = 0.69$, $P = 0.58$; Fig. S4*B*). Adjustment for log-transformed parasitemia at diagnosis increased heritability of DHA IC_{50} (Tasanh: $H^2 = 0.18$, $F = 1.95$, $P = 0.035$; Pailin: $H^2 = 0.15$, $F = 1.46$, $P = 0.32$), but not approaching the considerably higher levels observed for clinical parasite clearance phenotypes. In addition, although DHA IC_{50} values in parasites from the Cambodian sites were greater than those from parasites from Thailand or Bangladesh, within Cambodian clones, the fastest-clearing parasites did not have the lowest IC_{50} values (Fig. S4*B*).

Genotype Associations with DHA IC_{50}. We used EMMA to estimate the association between each SNP and log-transformed DHA IC_{50}. Because DHA IC_{50}s assessed using different methods were not comparable (as evidenced by the two Cambodian parasites having different distributions of IC_{50}s, but similar distributions of clinical phenotypes), we analyzed data from sites using different methods separately (i.e., data from Tasanh, Cambodia, and Bandarban, Bangladesh, were analyzed separately from data from Pailin, Cambodia, and Wang Pha, Thailand). Linear mixed-regression models included age, study site, and log-transformed parasitemia at diagnosis as covariates, with a genetic similarity matrix included as a random effect to account for population structure. A phenotype permutation approach was used to determine the threshold for genome-wide significance (3.2E-05 for Tasanh and Bandarban; 4.9E-05 for Pailin and Wang Pha). Using this approach, no SNPs achieved genome-wide significance for association with DHA IC_{50} for either sample set. The top SNPs from each sample set were different from each other and from the SNPs identified in models of parasite clearance phenotypes [Tasanh and Bandarban: MAL4-266872 ($P = 7.5E\text{-}04$), MAL14-1031954 ($P = 9.8E\text{-}04$), MAL12-592322 (1.1E-03); Pailin and Wang Pha: MAL14-4848 (3.9E-04), MAL6-2368 (1.2E-03), MAL9-2340 (1.4E-03)].

The Random Forests method was used to assess the importance of each SNP in predicting DHA IC_{50}. The importance of individual SNPs and other covariates in predicting DHA IC_{50} was estimated by the percent increase in mean-squared error. The best predictor of DHA IC_{50} in parasites from Tasanh, Cambodia, and Bandarban, Bangladesh, was study site, followed by log-transformed parasitemia at diagnosis and the two SNPs most predictive of parasites from Bangladesh [Figs. S5*A* and S7*E*; percent variance explained by all predictors (%Var) = 73.4)]. When parasites from Tasanh were examined alone, the best predictor of DHA IC_{50} became log-transformed parasitemia at diagnosis (Fig. S5*B*), and the percent of the variance in IC_{50} explained by the predictors dropped to negligible levels (%Var = −7.5), suggesting that variables other than study site did not explain a significant amount of variation in DHA IC_{50} in the analysis of Tasanh and Bandarban parasites. When parasites from Pailin, Cambodia, and Wang Pha, Thailand, were examined, study site was also the best predictor of DHA IC_{50}, followed by SNPs on chromosomes 7, 9, and 6 (Fig. S5*C*; %Var = 42.7). None of these SNPs overlapped with SNPs identified in EMMA analyses of DHA IC_{50} or clinical parasite clearance phenotypes.

Discussion of Genotype Associations with DHA IC_{50}. Previous genome-wide association studies (GWAS) to identify regions of the parasite genome associated with artemisinin resistance have used an in vitro phenotype (IC_{50} to DHA or other artemisinin derivatives) estimated from culture-adapted parasite isolates. Working with in vitro phenotypes from culture-adapted parasites reduces confounding due to host factors; however, it is uncertain whether these standard in vitro measures of drug susceptibility capture the delayed parasite clearance phenomenon that is the hallmark of emerging artemisinin resistance, both because standard IC_{50} assays do not resemble the pharmacokinetics of the drug within the human host (artemisinins are metabolized and eliminated in the human host within a few hours, whereas parasites are continuously exposed to the drug over the entire 48 h duration of an in vitro assay) and because it is possible that important phenotypes can be lost in the process of culture adaptation.

Individual IC_{50} measurements have not correlated well with clinical artemisinin resistance parameters in artesunate efficacy studies (3, 4), and were not significantly different between culture-adapted parasites from Africa (where ACTs are still very effective) and Cambodia (the focus of emerging artemisinin resistance) (5). In this study, DHA IC_{50}s were higher in Cambodia than in Bangladesh, but demonstrated low heritability in identical parasite clones from Cambodia. This lack of heritability may account for the fact that no SNPs were significantly associated with this phenotype in regression analyses; however, lack of power could also be a plausible explanation for this result, because there was very little overlap in DHA IC_{50}s between the Tasanh, Cambodia, and Bandarban, Bangladesh, study sites, and very few parasites from Pailin, Cambodia, and Wang Pha, Thailand, with IC_{50} values (n = 41). Even in the absence of SNPs achieving genome-wide significance, none of the SNPs most strongly associated with DHA IC_{50} overlapped with SNPs significantly associated with parasite clearance phenotypes, suggesting that DHA IC_{50} is not measuring the parasite behavior responsible for the delayed clearance phenotype.

EMMA Models of Parasite Clearance Phenotypes Excluding Parasites from Bangladesh. To further evaluate the adequacy of the EMMA approach in accounting for population structure, we also performed the analysis excluding parasites from Bangladesh, because these parasites were genetically distinct from parasites from the other locations and are not resistant to artemisinins. When parasites from Bangladesh were excluded from models of parasite clearance half-life, the two SNPs most strongly associated with half-life were MAL13-1718319 (P = 1.3E-04) and MAL10-688956 (P = 3.1E-04). These are the same two SNPs that achieved genome-wide significance in models including parasites from Bangladesh, but neither of these SNPs achieved genome-wide significance when parasites from Bangladesh were excluded (significance threshold = 4.2E-05). When parasites from Bangladesh were excluded from models of parasite clearance time, the five SNPs most strongly associated with clearance time were MAL13-1718319 (P = 2.9E-05), MAL13-1719976 (P = 2.3E-04), MAL13-363472 (P = 4.6E-04), MAL8-261424 (P = 6.7E-04), and MAL14-718269 (P = 7.5E-04). Three of these five SNPs are among the four SNPs that achieved genome-wide significance in models including parasites from Bangladesh. Only one SNP, MAL13-1718319, achieved genome-wide significance when Bangladesh parasites were excluded (significance threshold = 3.9E-05).

SNPs Predictive of Parasites from Bandarban, Bangladesh, Based on Random Forests. There is currently no explicit way to account for population structure in the Random Forest analysis, aside from removing populations that appear to be most genetically distinct. In an effort to identify SNPs most likely to show false associations with clearance phenotype due to population structure, we used Random Forests to identify SNPs that were most predictive of parasites from the Bandarban, Bangladesh, study site. Parasites from Bangladesh were genetically distinct from parasites from the other study sites, and no suspected artemisinin resistance was observed at the Bangladesh site. As a result, we would not expect SNPs predictive of Bangladesh to be associated with artemisinin resistance. The Random Forest algorithm was able to classify parasites from Bandarban with zero error, and the SNPs most predictive of parasites from this study site were MAL14-2492091 and MAL9-1042451 (Fig. S7*E*).

SI Materials and Methods

Artemisinin Resistance Confirmation, Characterization, and Containment Clinical Trials. As part of the World Health Organization's Artemisinin Resistance Confirmation, Characterization and Containment (ARC3) collaboration, clinical trials of artesunate efficacy were conducted at two sites in western Cambodia (Pailin and Tasanh) and one site each in northern Thailand (Wang Pha) and Bangladesh (Bandarban). Blood samples were collected, leukocyte-depleted using Plasmodipur filters (EuroProxima), and preserved for parasite genotyping. All trial protocols were approved by the Research Ethics Review Committee of the World Health Organization, as well as local Institutional Review Boards (IRBs) at each study site. The treatment regimens and number of patients receiving each treatment can be found in Table S1. Informed consent was received from all participants at all sites.

Cambodia. The Pailin study was conducted by the Mahidol Oxford Tropical Medicine Research Unit, subsequent to a previously reported trial conducted in 2007 at this site (4). The study protocol was approved by the Ministry of Health in Cambodia, the Ethics Committee of the Faculty of Tropical Medicine of Mahidol University in Thailand, and the Oxford Tropical Medicine Ethical Committee. Patients with uncomplicated *P. falciparum* malaria with parasitemia >10,000/mm^3 were randomized to four study arms (Table S1). A total of 59 patients were enrolled from 2008 to 2009. DNA extracted from leukocyte-depleted blood from 59 clinical infections and four recurrent parasitemias underwent parasite genotyping.

The Tasanh study was conducted by the US Armed Forces Research Institute of Medical Sciences from 2008 to 2009 (3). The study protocol was approved by the Walter Reed Army Institute of Research IRB and the National Ethics Committee for Health Research, Phnom Penh, Cambodia. Patients with parasitemia between 1,000 and 200,000/mm^3 were randomized to three treatment arms (Table S1). A total of 143 patients were enrolled, and DNA extracted from leukocyte-depleted blood from 143 clinical infections and 10 recurrent parasitemias underwent parasite genotyping.

Thailand. The trial conducted in Wang Pha near the Thailand/Myanmar border was conducted by the Shoklo Malaria Research Unit (4). The study protocol was approved by the Ethics Committee of the Faculty of Tropical Medicine of Mahidol University in Thailand and the Oxford Tropical Medicine Ethical Committee. Patients with parasitemia >10,000/mm^3 were randomized to two treatment arms (Table S1). DNA extracted from leukocyte-depleted blood from 30 clinical infections and 11 recurrent parasitemias underwent parasite genotyping.

Bangladesh. The trial in Bandarban, near the Myanmar border, was conducted during 2008–2009 by investigators from the Medical University of Vienna. The study protocol was approved by the Ethics Committee of the Medical University of Vienna and the Ethics Committee of the International Centre for Diarrheal Disease Research, Bangladesh. Patients with parasitemia between 1,000–100,000 parasites per microliter were randomized to three study arms (Table S1). DNA extracted from leukocyte-depleted blood from 101 clinical infections from patients receiving artesunate and two recurrent parasitemias underwent parasite genotyping.

Parasite Genotyping. DNA was extracted from leukocyte-depleted blood samples using the Qiagen QIAamp Blood Midi Kit and extracted DNA underwent whole-genome amplification (WGA) by multiple displacement amplification using the Qiagen REPLI-g Mini Kit. The WGA material was quantified using the Quant-iT PicoGreen dsDNA Kit (Invitrogen). Briefly, a standard curve of Lambda DNA was made consisting of twofold dilutions from 25 ng/μL to 0.4 ng/μL. WGA DNA was diluted 1:10 in water, and 2 μL was mixed with 198 μL of the PicoGreen reagent diluted in TE buffer in duplicate. Fluorescence was read on a SpectraMax plate reader (Molecular Devices).

Quantitative PCR targeting the *apical membrane antigen 1* (*ama1*) gene was performed to estimate the proportion of parasite DNA within each sample. The forward and reverse primer sequences were 5′-CGTTGGATGGATTCTCTTTCGATTTCTTTC and 5′-CGTTGGATGTGCTACTACTGCTTTGTCCC, respectively. The 25-μL reaction consisted of 12.5 μL of 2× SYBR Green QPCR Master Mix (Applied Biosystems), 0.24-μM final primer concentration, and 2 μL of WGA DNA template. A standard curve of genomic DNA diluted from 12 ng/μL down to 0.1 ng/μL by twofold serial dilutions was run with the samples. All samples and standards were tested in duplicate. The quantitative PCR was performed on an Applied Biosystems 7300 platform using the following cycling conditions: 95 °C denaturation for 10 min; five cycles of 94 °C for 45 s, 56 °C for 45 s, and 72 °C for 45 s; then 29 cycles of 94 °C for 45 s, 65 °C for 45 s (data collection), and 72 °C for 45 s. A dissociation curve was run from 95 °C to 55 °C.

WGA samples were genotyped at 8,079 SNPs using a *P. falciparum*-specific molecular inversion probe Affymetrix SNP array. The array was designed by collaborating investigators at the National Institute of Allergy and Infectious Diseases who sequenced ~4,000 gene fragments (1–1.5 kb) from four isolates (Dd2, HB3, 7G8, and D10) and compared these sequences to that of the reference strain 3D7 to identify SNPs. All SNPs within the sequenced regions were printed on the chip (6). A maximum of 4 μg of total genomic WGA DNA or 25 μL of WGA material was enzymatically purified and used for genotyping, resulting in a range of ~1 ng to 400 ng parasite DNA in the assay. The assay was performed following the Affymetrix GeneChip Scanner 3000 Targeted Genotyping System User Guide (www.affymetrix.com/) without modification, except for alterations that were made to the first PCR thermal cycling parameters (5).

Phenotypes. ***Parasite clearance time and rate.*** Following malaria diagnosis and treatment with artesunate, asexual parasitemia was assessed by microscopy at regular intervals in each of the four trials. In the Pailin, Cambodia, and Wang Pha, Thailand, trials, parasitemia was evaluated at 0, 4, 8, and 12 h, and then every 6 h until two consecutive slides were negative for parasites (4). In the Tasanh, Cambodia, trial, parasitemia was evaluated eight times during the first 24 h, and then every 6 h until two consecutive slides were negative for parasites (3). In the Bandarban, Bangladesh, trial, parasitemia was evaluated twice daily (approximately every 12 h).

Parasite clearance half-lives were estimated using the parasite clearance estimator developed by the WorldWide Antimalarial Resistance Network (www.wwarn.org/research/parasite-clearance-estimator) (7). The estimator calculates the parasite clearance rate based on the linear portion of the $\log_e$ parasitemia time curve, and half-life is estimated as $\log_e(2)$/clearance rate (7). Due to the less-frequent assessment of parasitemia in the Bandarban, Bangladesh, study, some patients had only two parasite counts available before clearance (i.e., patients who cleared their parasites within 24 h). To get an estimate of clearance for these patients, the zero parasitemia at 24 h was replaced by the level of detection for microscopy (40 parasites per microliter), and a straight line was fitted to the $\log_e$-transformed parasitemia values. As a result, the absolute value of the slope likely underestimated the clearance rate (and overestimated the half-life) for patients at this site.

Dihydroartemisinin IC_{50}. In the Tasanh, Cambodia, and Bandarban, Bangladesh, studies, fresh samples, without prior freezing or preculturing, were assayed in the histidine-rich protein 2 drug sensitivity ex vivo assay for susceptibility to DHA (8). In the Pailin, Cambodia, and Wang Pha, Thailand, studies DHA IC_{50} was determined from frozen isolates using a tritiated hypoxanthine-based assay (9).

Data Analysis. ***Genotype calling and quality control.*** Raw allele intensity data from the SNP array were normalized using GTGS. The following quality control (QC) metrics were modified to retain allele intensity data for 41 samples that barely failed the software's default thresholds: QC call rate was decreased from 80% to 65%, QC half-rate was increased from 10% to 15.5%, and the controls coefficient of variation was increased from 30% to 34%. To assess the robustness of SNP calls, genotypes were called from normalized intensities using three algorithms: (*i*) GTGS, (*ii*) Illuminus (1), and (*iii*) a heuristic algorithm based on discrete cutoffs of intensity strength (total signal intensity between 120 and 10,142 required to make a call) and contrast (homozygous −1.5 to −0.5, heterozygous −0.5 to 0.5, homozygous 0.5–1.5, otherwise missing call), with cutoffs being established from an analysis of empirical distributions. SNP calls from the heuristic algorithm were used in the analysis because this algorithm allows for heterozygous calls at a wider range of allele frequencies than algorithms designed to genotype diploid individuals, and may therefore more accurately represent mixtures of haploid organisms observed in *P. falciparum* field samples.

Uninformative SNPs that were invariant or had very low minor allele frequency (MAF; <1%) were excluded from the analysis, as were SNPs with an extreme number of undetermined or heterozygous calls (>15%). Samples with an extreme number of undetermined SNP calls were also excluded from the analysis (>20%). The thresholds used to determine the SNPs and samples included in the analysis are detailed above and in Fig. S2. The genotype data used in the analysis is available on the PlasmoDB database (www.plasmodb.org/).

Phenotype heritability. We used ANOVA to assess heritability (H^2) of the two clinical phenotypes in identical parasite clones identified from among the 331 samples included in this study (10) Clones were identical at all of 8,079 SNPs on the array, excepting those that were either heterozygous or missing in all members of the clone group. Identical clones were only observed among the Cambodian parasites in the study; no Cambodian clones were observed at the study sites in Thailand or Bangladesh, and no clones overlapped between the sites in Thailand and Bangladesh. Confounding factors that were significantly associated with the phenotype were included in final ANOVA models [e.g., age (clearance half-life) and log-transformed parasitemia at diagnosis (clearance time)] to yield adjusted heritability estimates. ANOVA was performed using SAS v.9.2.

Regression and PCA. Covariates for regression models were selected based on bivariate association with either clinical phenotype. Potential confounding factors that were considered include patient age, study site, parasitemia at diagnosis, and drug dose/treatment arm. Two regression approaches were used to estimate the association between each SNP and the clinical phenotypes while adjusting for confounding factors and population structure. The first approach used linear regression models of the effect of each SNP on the phenotype, with patient age, study site, parasitemia at diagnosis, and the first four principal components from a PCA included as fixed effects. Principal components were estimated from a matrix of pairwise sample identity-by-state metrics (11). The second approach used linear mixed-regression models, adjusting for the same covariates as above, except for the principal components, which were replaced by an identical-by-state allele-sharing matrix included as a random effect to account for the correlation between genetically similar individuals that results from population structure. Drug dose/treatment arm was not significantly associated with either phenotype in regression models and was therefore excluded from the final models. The EMMA method was used to determine restricted maximum-likelihood estimates and *P* values (2). Because *Plasmodium* is a haploid organism, heterozygous SNP calls do not represent heterozygous individuals, but rather reflect polyclonal infections. In this sample set, ~5% of all SNP calls were heterozygous, and similar results were obtained from statistical analyses when heterozygous calls were included or excluded and when polyclonal infections were excluded from the analysis (i.e., samples with heterozygous calls were excluded). In the final analysis, heterozygous calls were excluded when estimating genotype/phenotype associations, but samples with heterozygous calls were retained in the dataset. A phenotype permutation approach was used to estimate the threshold for genome-wide significance. A total of 10,000 permuted datasets were generated for each phenotype, and the minimum *P* value for each permutation was recorded. The threshold was chosen as the fifth percentile of the minimum *P* values from the permuted datasets. The significance thresholds for clearance half-life and clearance time were 2.7E-05 and 3.6E-05, respectively, and were only slightly less stringent than a Bonferroni threshold (1.8E-05). Quantile–quantile plots for *P* values were used to assess the robustness of modeling approaches in minimizing false-positive results due to population structure. The quality of the genotype calling for significant SNPs was assessed by inspection of cluster plots (12). Regression analysis and PCA were performed using R statistical software (13).

Random Forest. Random Forest analyses (14) were done using the randomForest package (15) of the R statistical computing environment (13). The number of variables tried at each split was 1,031, and the number of trees in the forest was 10,000. The importance of each SNP in predicting the phenotype was assessed based on the percent increase in mean-squared error.

Odds ratios to evaluate SNPs as markers of delayed parasite clearance. Logistic regression was used to estimate odds ratios (ORs) comparing the log odds of an infection having clearance half-life >5 h in parasites with a given allele at each SNP compared with infections with the alternative allele. A clearance half-life of 5 h represents the median clearance half-life observed in our sample set. Covariates such as patient age and log-transformed parasitemia did not significantly change OR estimates, and were therefore not included in the final models. Logistic regression was performed in SAS v.9.2.

Definition of linkage disequilibrium windows around phenotype-associated SNPs. Linkage disequilibrium (LD) windows containing SNPs associated with clinical phenotypes were defined using Haploview (16) (www.broad.mit.edu/mpg/haploview/). LD windows were defined to include all SNPs (including invariable and low-MAF SNPs) upstream and downstream from the associated SNP up to but not including the next SNP with MAF >0.05 and $R^2 < 0.3$. LD windows were defined in parasites in each country separately to avoid false estimates of linkage disequilibrium due to population structure.

Signatures of selection. Parasites were grouped based on country of origin (Cambodia, Thailand, Bangladesh). To identify regions of the genome under recent positive selection in Cambodia, parasites from Cambodia were compared with those from Thailand or Bangladesh. Samples representing 323 clinical malaria infections (excluding recurrent parasitemias) were included in the analysis. SNPs and samples with a large proportion of undetermined calls were excluded, as were SNPs and samples with hyperheterozygosity, with thresholds determined based on manual inspection of inflection points on plotted data No MAF threshold was imposed. The final number of SNPs used in the analysis was 6,649.

Extended haplotype homozygosity (EHH) was calculated according to previously published methods (17) using the equation below, where G is the number of haplotypes detected, n_i is the number of individuals with a given haplotype, and N is the total number of individuals. Haplotypes used in the EHH analysis were composed of all genotyped alleles from a core SNP to a point X. Core SNPs included all SNPs with MAF >0.05, with point X representing genomic loci 500 kb in the forward and reverse directions from the core SNP. At every genotyped locus, integrated EHH (iHH) scores were calculated for each population and the cross-population (XP)-EHH score was calculated as ln(iHHpopA/iHHpopB) (17). All XP-EHH scores were binned and normalized, giving a normalized XP-EHH score for each SNP regardless of MAF (XP-EHH does not require the core to have a minimum MAF) and accounting for genome-wide differences in haplotype length between populations:

$$\mathrm{EHH} = \sum_{\mathrm{i}=1}^{\mathrm{G}} \frac{\binom{n_{\mathrm{i}}}{2}}{\binom{\mathrm{N}}{2}} = \frac{1}{\mathrm{N}(\mathrm{N}-1)} \sum n_{\mathrm{i}}(n_{\mathrm{i}}-1).$$

Wright's $\mathrm{F_{ST}}$ statistic was calculated for each SNP in pairwise population comparisons (equation below) (18). SNP $\mathrm{F_{ST}}$ values were combined within a 20-SNP sliding window, with a five-SNP overlap (19, 20). All $\mathrm{F_{ST}}$ values were normalized by dividing the average $\mathrm{F_{ST}}$ value of the window by the average $\mathrm{F_{ST}}$ of all SNPs on each chromosome in each population to account for differences in $\mathrm{F_{ST}}$ due to genetic drift:

$$\mathrm{F_{ST}} = 1 - \frac{(\mathrm{H_1} + \mathrm{H_2})/2}{\mathrm{H_0}}.$$

Genomic regions under recent selection in Cambodia were identified based on a combined XP-EHH–$\mathrm{F_{ST}}$ score, calculated within the same SNP windows that were used in the sliding-window $\mathrm{F_{ST}}$ analysis. The XP-EHH score of each window was determined as the average normalized XP-EHH score of all 20 SNPs, and the combined score was calculated as the normalized XP-EHH window score multiplied by the normalized $\mathrm{F_{ST}}$ window value. XP-EHH–$\mathrm{F_{ST}}$ scores were determined for both pairwise population comparisons independently (Cambodia/Thailand and Cambodia/Bangladesh). Windows representing regions under selection in Cambodia (Fig. S9) were selected based on the inflection point of a plot of the ranked XP-EHH–$\mathrm{F_{ST}}$ scores from both population comparisons (Fig. S8), and top-ranked signatures were selected from among these regions as the top 10% of windows from each population comparison. Windows were combined manually based on overlap of window boundaries. Thirty windows represented the 10 top-ranked signatures for the Cambodia/Thailand comparison, and 30 windows represented the 12 top-ranked signatures for the Cambodia/Bangladesh comparison.

1. Teo YY, et al. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* 23(20):2741–2746.
2. Kang HM, et al. (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178(3):1709–1723.
3. Bethell D, et al. (2011) Artesunate dose escalation for the treatment of uncomplicated malaria in a region of reported artemisinin resistance: A randomized clinical trial. *PLoS ONE* 6(5):e19283.
4. Dondorp AM, et al. (2009) Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med* 361(5):455–467.
5. Mu J, et al. (2010) Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet* 42(3):268–271.
6. Mu J, et al. (2007) Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat Genet* 39(1):126–130.
7. Flegg JA, Guerin PJ, White NJ, Stepniewska K (2011) Standardizing the measurement of parasite clearance in *falciparum* malaria: The parasite clearance estimator. *Malar J* 10:339.
8. Noedl H, Attlmayr B, Wernsdorfer WH, Kollaritsch H, Miller RS (2004) A histidine-rich protein 2-based malaria drug sensitivity assay for field use. *Am J Trop Med Hyg* 71(6):711–714.
9. Lim P, et al. (2010) Decreased in vitro susceptibility of *Plasmodium falciparum* isolates to artesunate, mefloquine, chloroquine, and quinine in Cambodia from 2001 to 2007. *Antimicrob Agents Chemother* 54(5):2135–2142.
10. Anderson TJ, et al. (2010) High heritability of malaria parasite clearance rate indicates a genetic basis for artemisinin resistance in western Cambodia. *J Infect Dis* 201(9): 1326–1330.
11. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
12. Jallow M, et al.; Wellcome Trust Case Control Consortium; Malaria Genomic Epidemiology Network (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet* 41(6):657–665.
13. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
14. Breiman L (2001) Random Forests. *Mach Learn* 45:5–32.
15. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2(1): 18–22.
16. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2):263–265.
17. Sabeti PC, et al.; International HapMap Consortium (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164): 913–918.
18. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74(1):175–195.
19. Oleksyk TK, et al. (2008) Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE* 3:, e1712.
20. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res* 15 (11):1468–1476.

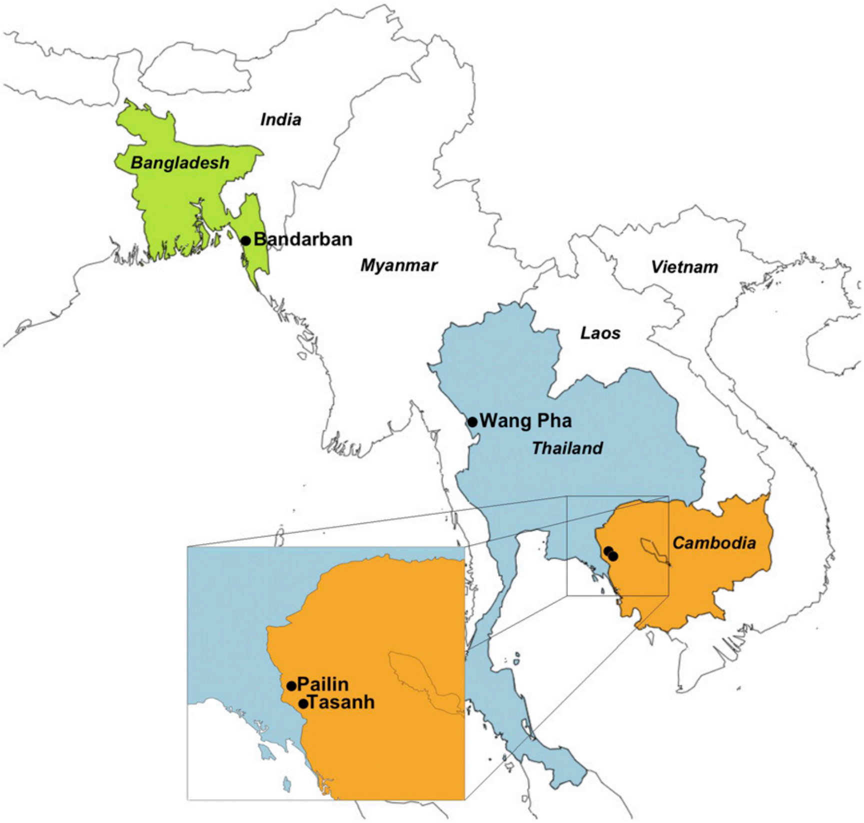

Fig. S1. Map showing ARC3 clinical trial sites. As part of the ARC3 pilot project, four clinical trials of artesunate curative efficacy were conducted at two sites in western Cambodia, in Thailand near the Myanmar border, and in Bangladesh. DNA extracted from *P. falciparum* parasites collected during these trials was genotyped using a SNP microarray specific to *P. falciparum*. Parasite genotypes were examined for signatures of recent positive selection and association with clinical resistance phenotypes to identify regions of the parasite genome associated with artemisinin resistance.

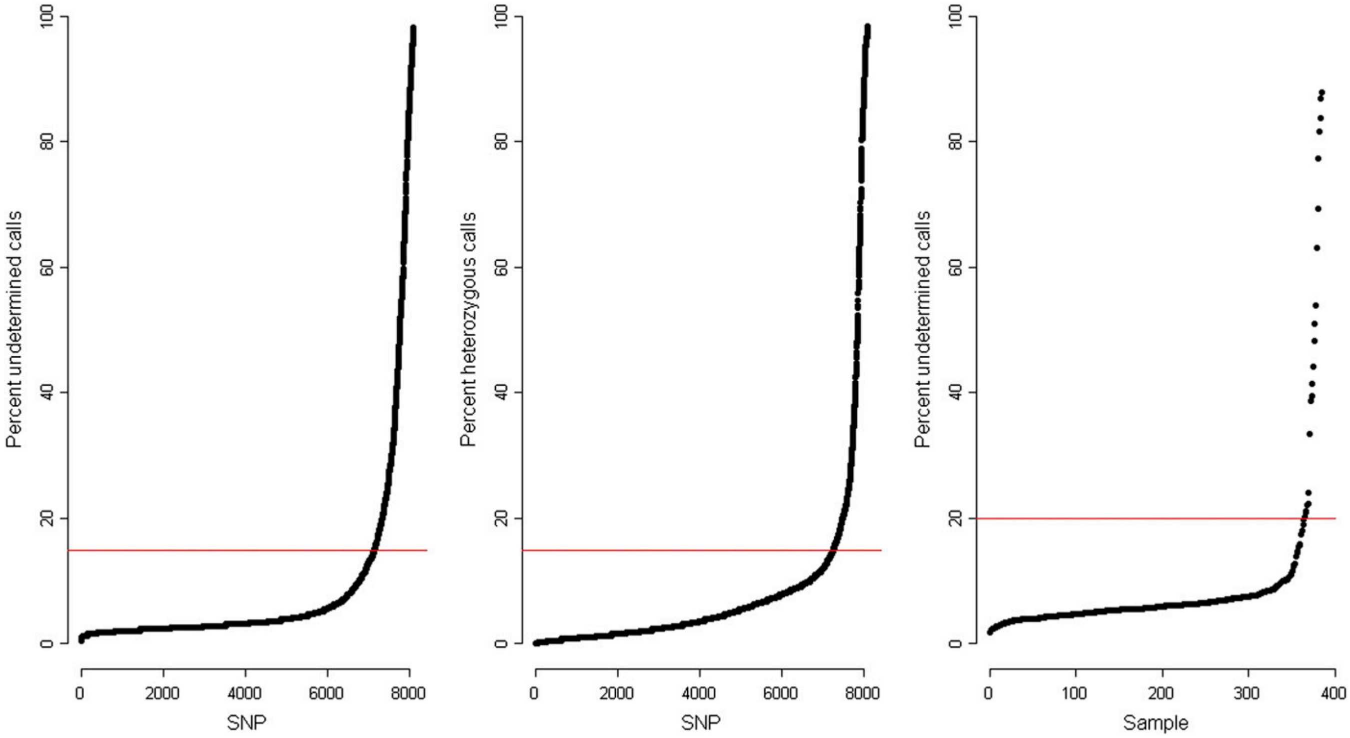

Fig. S2. Quality control thresholds for GWAS SNPs and samples. Plots of the proportion of undetermined calls per SNP (*Left*), heterozygous calls per SNP (*Center*), and missing calls per sample (*Right*). Horizontal red lines indicate the thresholds that were used in the study. Genotype/phenotype associations were robust to the thresholds that were chosen.

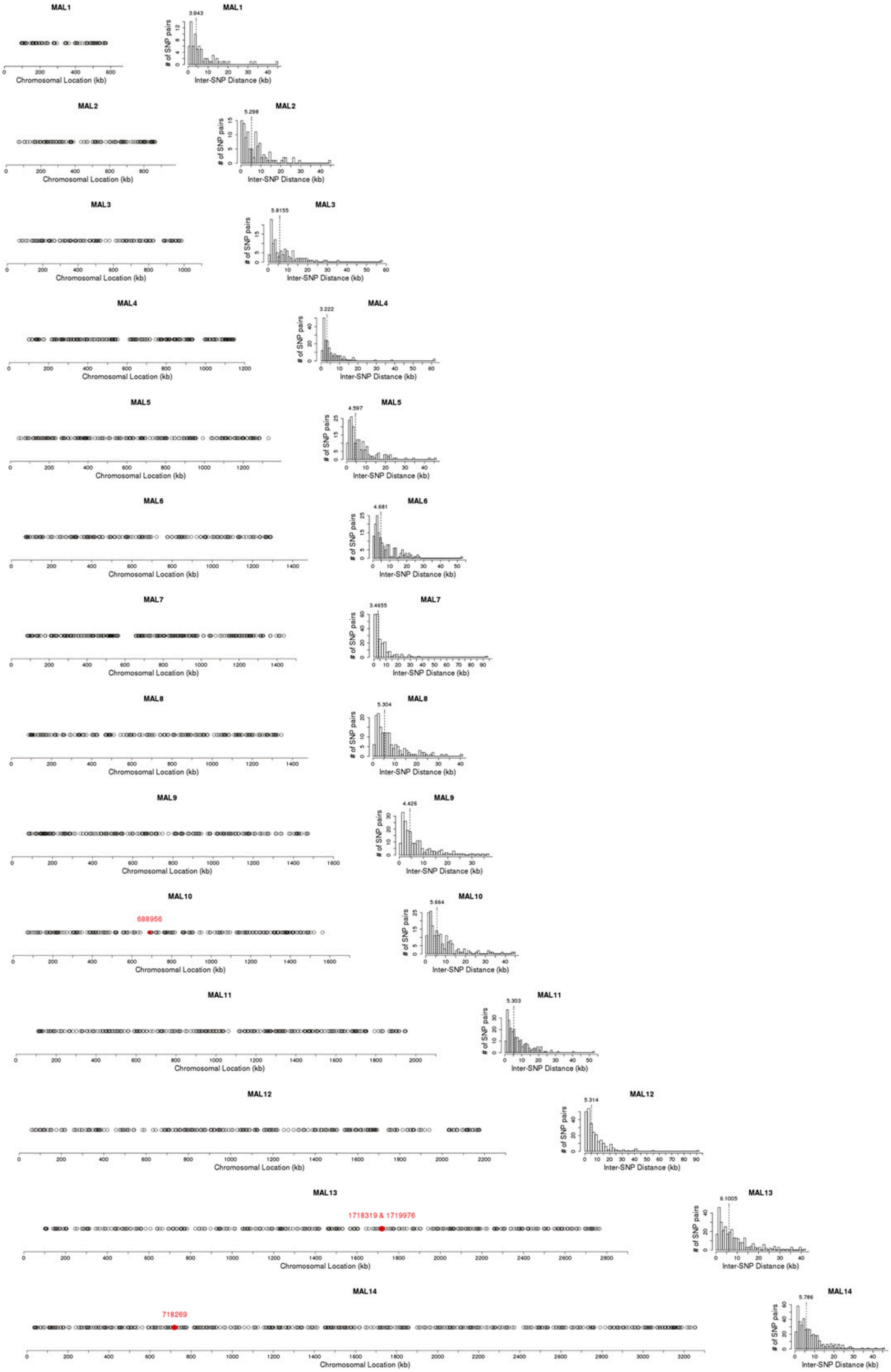

Fig. S3. Distribution of informative SNPs across the genome. The location of each of the 2,827 SNPs (circles) included in the GWAS is shown for each chromosome. Chromosomal location (in kb) is shown along the horizontal axis. A histogram showing the distribution of the distances (in kb) between pairs of adjacent SNPs is also shown for each chromosome, with the median distance indicated by a labeled dashed line. In most cases, large gaps in SNP coverage represent regions with highly variable *var* genes. The chromosomal locations of the top GWAS hits are shown in red on chromosomes 10 and 13.

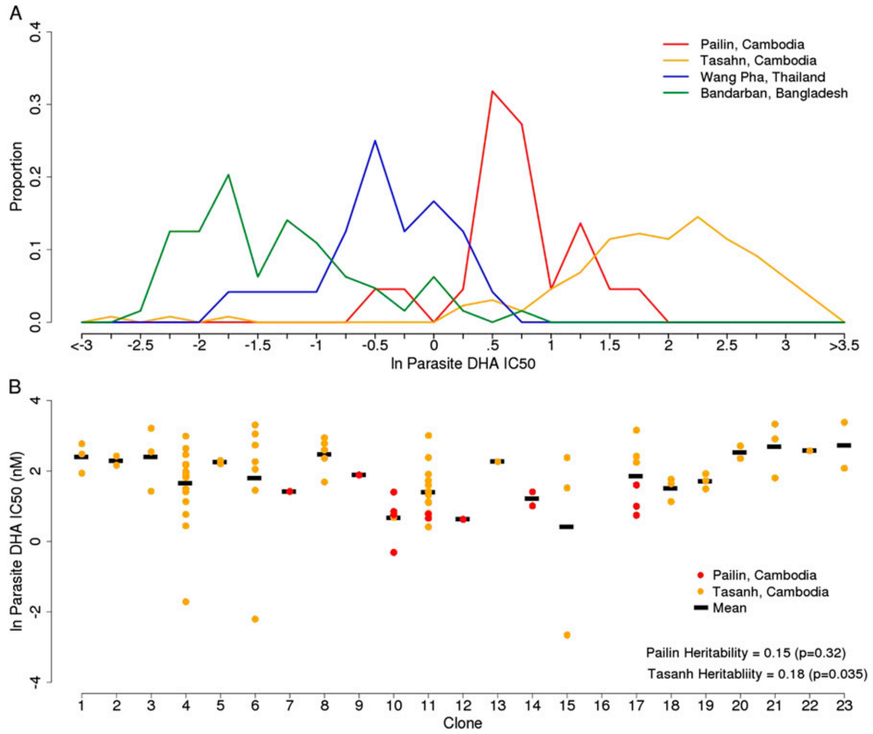

Fig. S4. Distribution and heritability of DHA IC_{50}. (*A*) Distribution of log-transformed DHA IC_{50} at the four study sites. (*B*) Distribution of log-transformed DHA IC_{50} in identical clones from the Tasanh, Cambodia, and Pailin, Cambodia, study sites. The heritability of DHA IC_{50} (adjusted for confounding factors) was assessed separately for the two study sites and is shown in *B Bottom Right*.

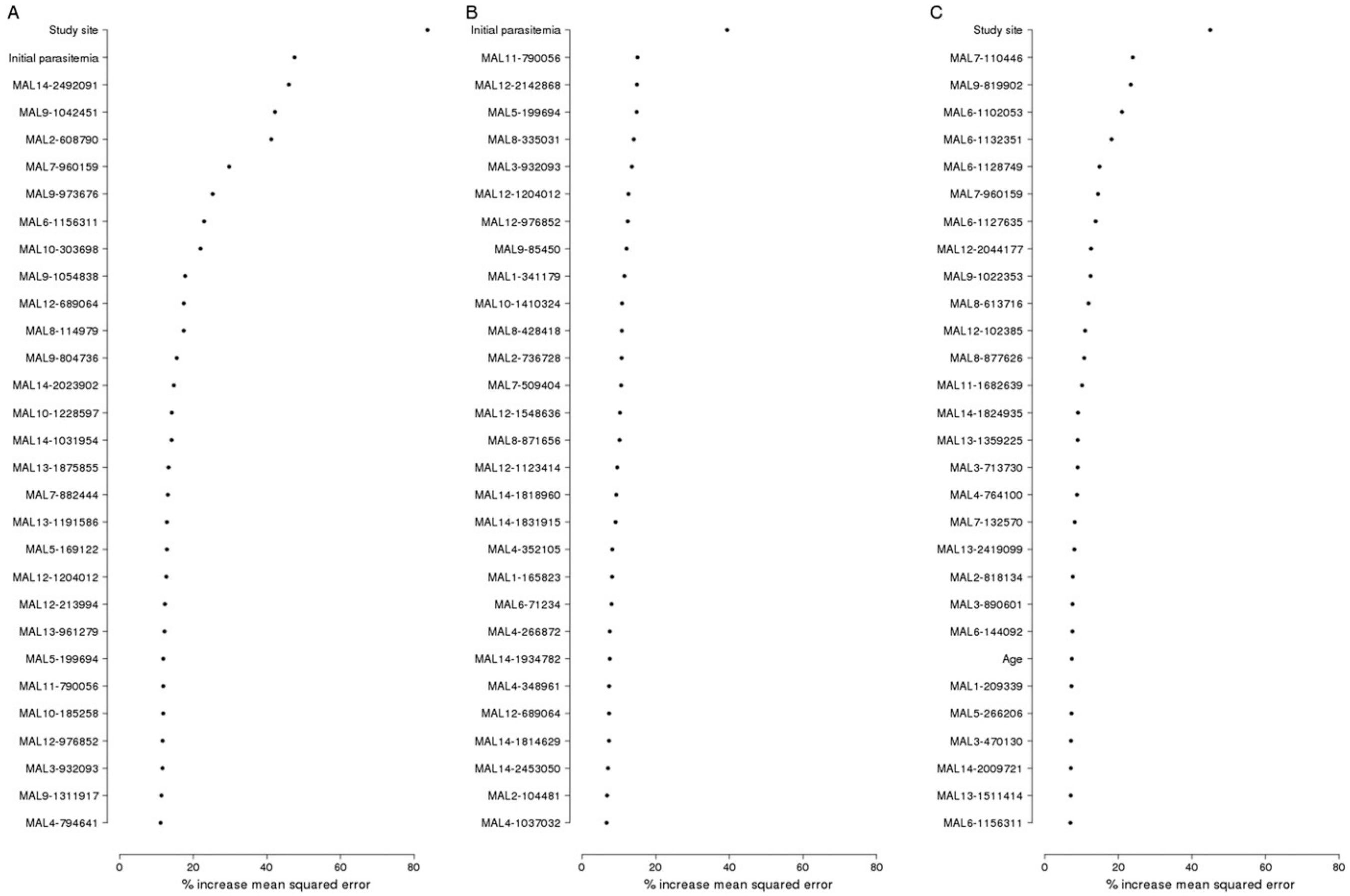

Fig. S5. SNPs associated with DHA IC_{50} using Random Forests. Plots showing the importance of specific SNPs and covariates in predicting DHA IC_{50} in parasites from (*A*) Tasanh, Cambodia, and Bandarban, Bangladesh; (*B*) Tasanh, Cambodia; and (*C*) Pailin, Cambodia, and Wang Pha, Thailand. The importance of each variable in predicting DHA IC_{50} was assessed as the percent increase in mean-squared error, with the best predictors having the largest percent increase in mean-squared error.

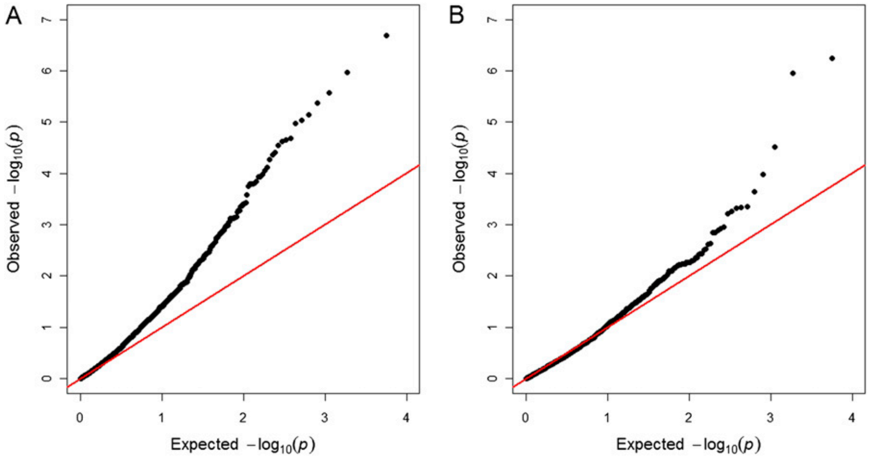

Fig. S6. Accounting for confounding due to population structure. Quantile–quantile (Q–Q) plots for *P* values used to evaluate the effectiveness of two approaches in accounting for population structure in regression models. (*A*) Q–Q plot for a PCA-based approach where regression models included the first four principal components as fixed effects in linear regression models. (*B*) Q–Q plot for the EMMA approach, which accounts for population structure by including a genetic similarity matrix as a random effect in regression models.

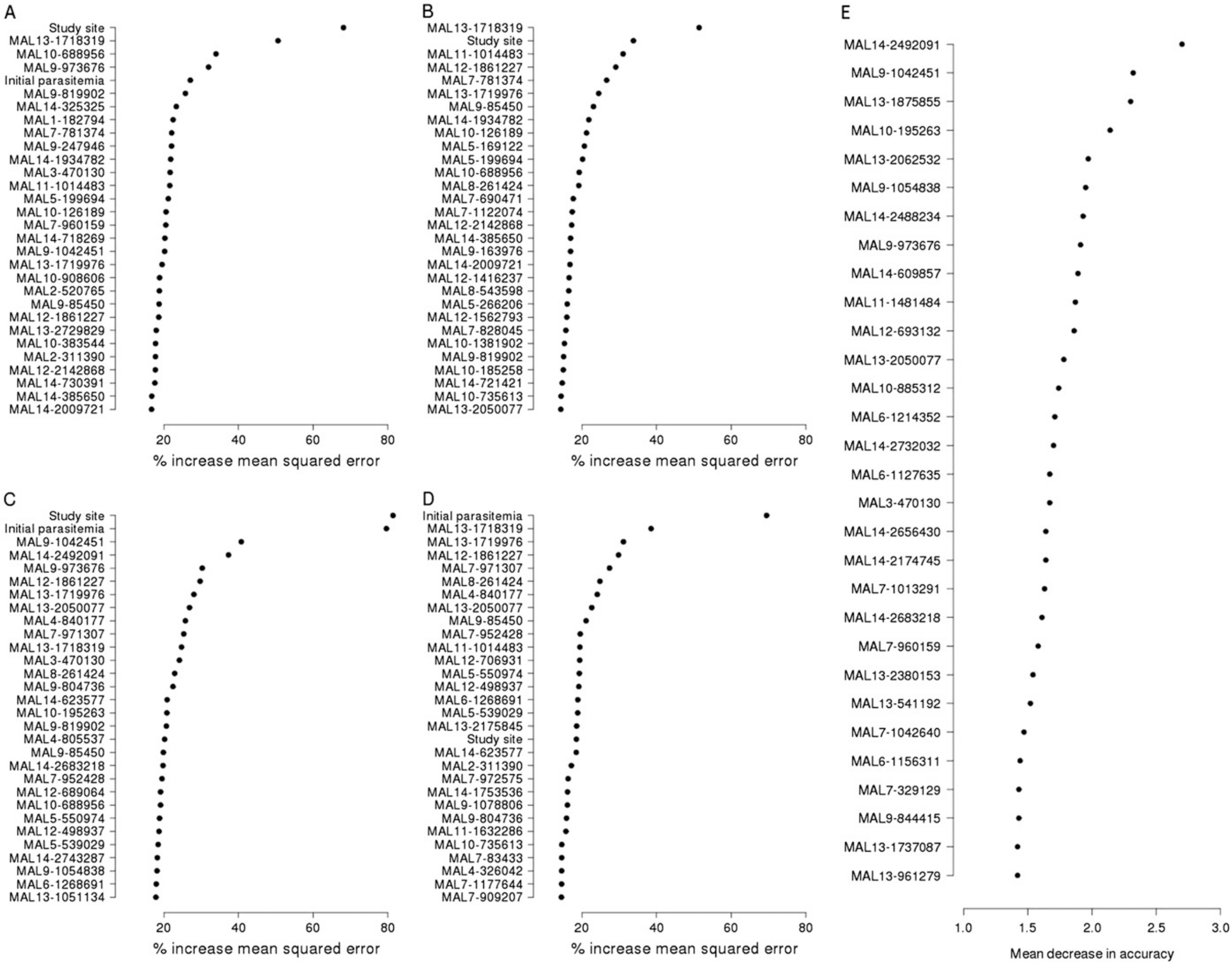

Fig. S7. SNPs associated with parasite clearance phenotypes using Random Forests. Plots showing the importance of specific SNPs and other covariates in predicting parasite clearance phenotypes. Predictors of parasite clearance half-life, including all parasites and excluding those from Bandarban, Bangladesh, are shown in *A* and *B*, respectively. Predictors of parasite clearance time, including all parasites and excluding those from Bandarban, Bangladesh, are shown in *C* and *D*, respectively. The importance of each variable in predicting the phenotype was assessed as the percent increase in mean-squared error, with the best predictors having the largest percent increase in mean-squared error. *E* shows plots showing the importance of specific SNPs for identifying parasites from Bandarban, Bangladesh, as determined by Random Forests. The importance of each SNP for correctly identifying parasites from Bandarban was assessed as the mean decrease in classification accuracy. The Random Forests algorithm was able to classify parasites from Bandarban with zero error.

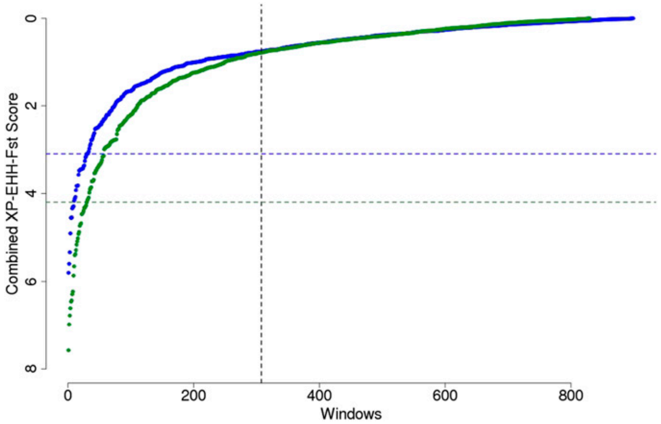

Fig. S8. Determination of cutoff values for genomic regions under recent selection in Cambodia. Plot of the ranked combined XP-EHH–F_{ST} scores, with blue points representing scores from the Cambodia vs. Thailand comparison, and green points representing scores from the Cambodia vs. Bangladesh comparison. The dashed black line indicates the inflection point used to select regions with strong evidence of selection in Cambodia. Top-ranked signatures were selected from among these regions as the top 10% of windows from each separate population comparison, indicated by dashed blue and green lines for the Thailand and Bangladesh comparator populations, respectively.

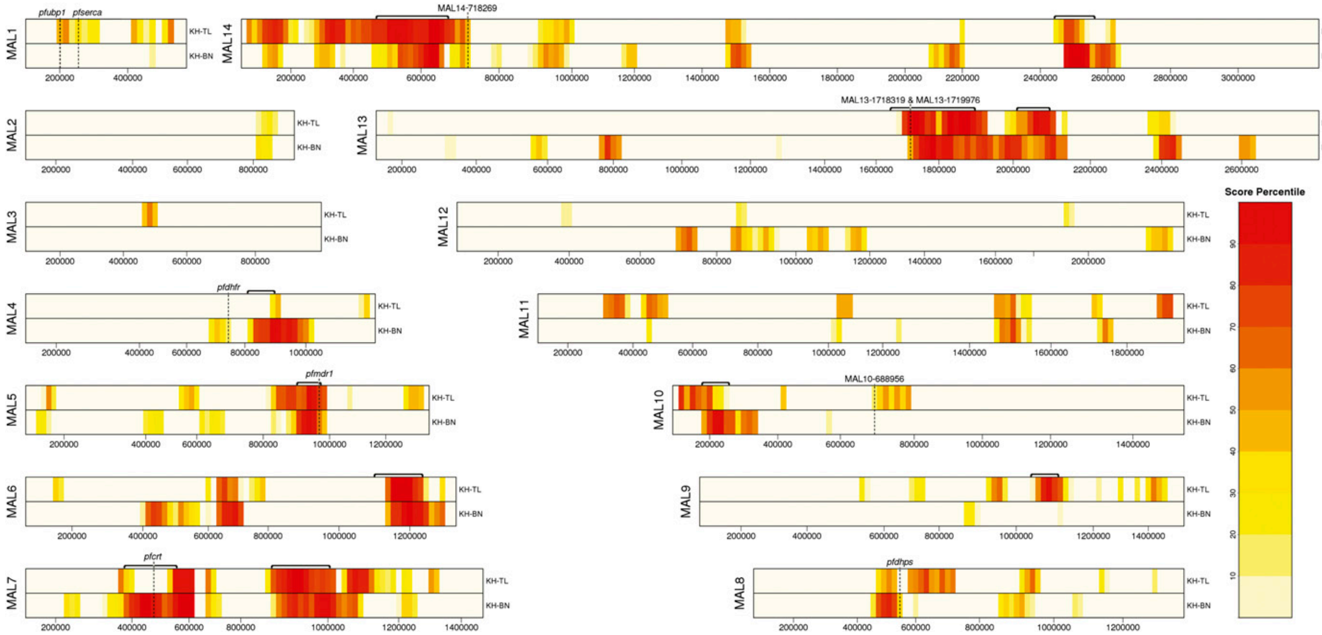

Fig. S9. Regions of the parasite genome under recent positive selection. Regions highlighted in yellow to red represent genomic regions under selection in western Cambodian parasites, and were selected based on an inflection point in the plot of ranked XP-EHH–F_{ST} scores (Fig. S9). Red hues represent regions with the highest XP-EHH–F_{ST} rank and the strongest evidence of selection. The top bar of each chromosome represents regions identified in comparisons of Cambodia and Thailand (KH-TL), and the bottom bar represents regions identified in comparisons of Cambodia and Bangladesh (KH-BN). Putative artemisinin resistance candidate genes, known resistance genes to other antimalarial drugs, and SNPs associated with parasite clearance phenotypes are represented by vertical dashed lines, and top-ranked signatures (top 10% of plotted data based on ranked XP-EHH–F_{ST} score) are shown with a horizontal bar corresponding to the length of the signature. The more recent the selection, the longer we would expect the signature to be. The signature on chromosome 13 that contains two of the GWAS "hits" is one of the largest observed in this analysis, suggesting that it could be due to a more recent selective event in Cambodia than those on other chromosomes around known drug resistance markers such as *dhfr*, *dhps*, and *pfmdr1*. Of note, the signature on chromosome 7 containing the chloroquine resistance transporter gene *pfcrt* also contains long haplotypes, which may be maintained by continued drug pressure due to use of this drug to treat *Plasmodium vivax* in the region.

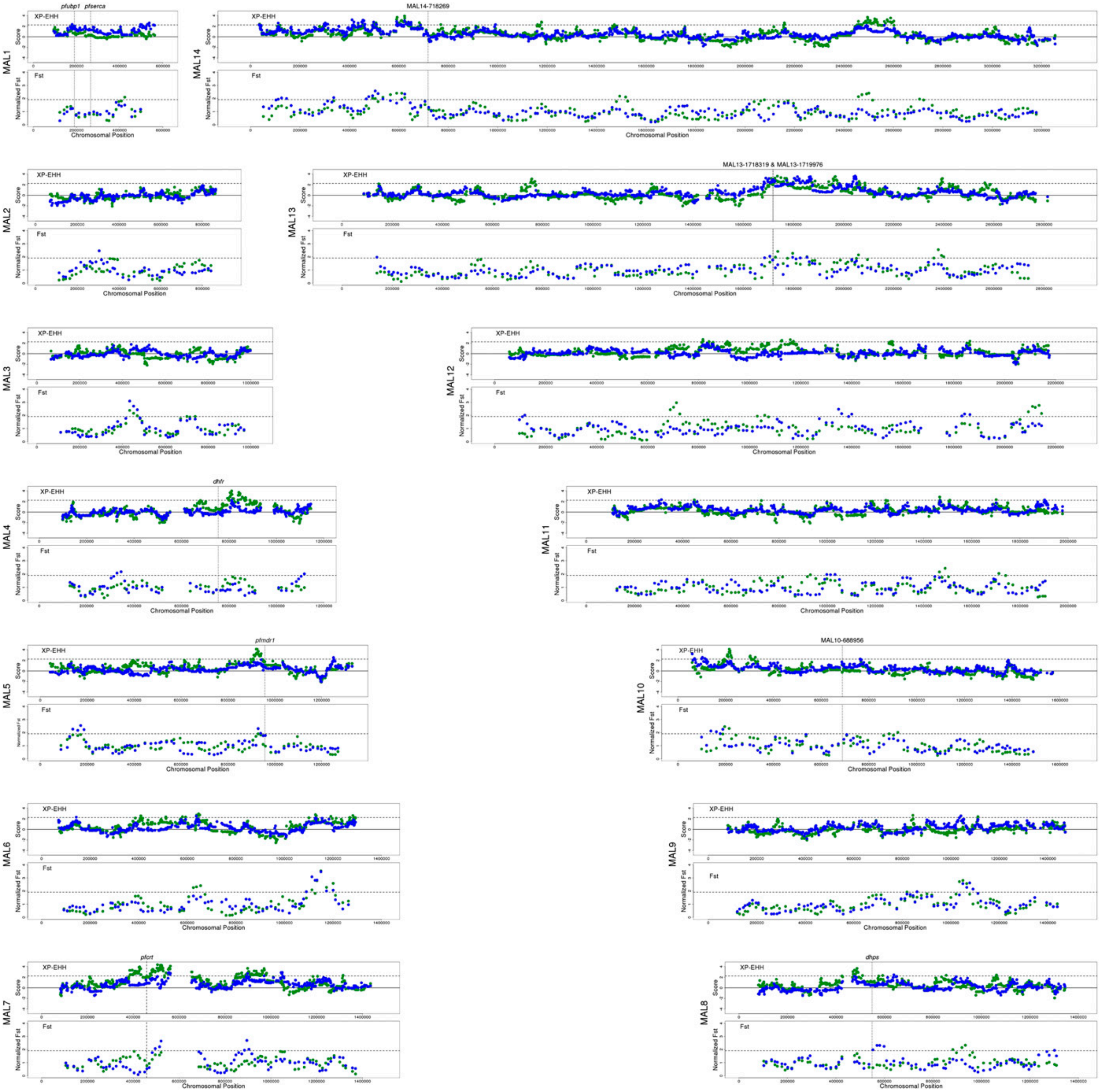

Fig. S10. Individual XP-EHH values by SNP and average F_{ST} window values across each chromosome. Plots showing individual SNP XP-EHH values (*Upper* panel of each chromosome) and average F_{ST} values for SNP windows (*Lower* panel of each chromosome). Values from the Cambodia vs. Thailand comparison are shown in blue, and those from the Cambodia vs. Bangladesh comparison are in green. Putative artemisinin resistance genes, known drug resistance genes, and SNPs associated with either parasite clearance half-life or parasite clearance time in the GWAS are all represented by vertical dotted lines.

Table S1. Treatment regimens and sample sizes for ARC3 clinical trials

Study	Treatment	No. of patients
Tasanh, Cambodia	2 mg/kg AS, 7 d	75
	4 mg/kg AS, 7 d	39
	6 mg/kg AS, 7 d	28
Pailin, Cambodia	6 mg/kg AS, 7 d	10
	6 mg/kg AS, 7 d, split dose	10
	8 mg/kg AS, 3 d, 15 mg/kg MQ day 3, 10 mg/kg MQ day 4	14
	8 mg/kg AS, 3 d, split dose, 15 mg/kg MQ day 3, 10 mg/kg MQ day 4	14
	6 mg/kg AS, 3 d, 15 mg/kg MQ on day 3, 10 mg/kg MQ day 4*	5
	6 mg/kg AS, 3 d, split dose, 15 mg/kg MQ day 3, 10 mg/kg MQ day 4*	5
Wang Pha, Thailand	2 mg/kg AS, 7 d	15
	4 mg/kg AS, 3 d, 15 mg/kg MQ day 3, 10 mg/kg MQ day 4	15
Bandarban, Bangladesh	2 mg/kg AS, 7 d	51
	4 mg/kg AS, 7 d	50

AS, artesunate; MQ, mefloquine.

*Seven-day regimens of 6 mg/kg with 3-d regimens in 2009 following reports of neutropenia in the Tasanh, Cambodia, trial (1).

1. Bethell D, et al. (2010) Dose-dependent risk of neutropenia after 7-day courses of artesunate monotherapy in Cambodian patients with acute *Plasmodium falciparum* malaria. *Clin Infect Dis* 51(12):e105–e114.

Table S2. Chromosomal locations of top-ranked signatures of selection

Chromosome	Region start	Region stop	Length (bases)	Comparator population
4	794641	849273	54632	Bangladesh
5	907206	960404	53198	Bangladesh
	907206	950912	43706	Thailand
6	1080126	1201811	121685	Thailand
	1114277	1224779	110502	Bangladesh
7	370564	460214	89650	Bangladesh
	460214	552590	92376	Bangladesh and Thailand
	860675	944740	84065	Thailand
	917890	990099	72209	Bangladesh
9	1022353	1078806	56453	Thailand
10	178161	244211	66050	Bangladesh
13	1648093	1738570	90477	Thailand
	1706965	1828443	121478	Bangladesh
	1760054	1886603	126549	Thailand
	2009961	2074874	64913	Thailand
14	446142	566686	120544	Thailand
	536297	655014	118717	Bangladesh
	566686	665380	98694	Thailand
	2433083	2536420	103337	Bangladesh

Table S3. Chromosomal locations of LD blocks containing SNPs associated with delayed parasite clearance

	LD block					
Chromosome	Start	Stop	Size, kb	No. of SNPs*	SNPs with MAF >0.05†	Associated SNP
10	632782	713023	80.2	21	1	MAL10-688956
13	1681358	1757919	76.6	28	3	MAL13-1718319, MAL13-1719976
14	629527	724053	94.5	39	4	MAL14-718269

*Number of SNPs of the 2,827 included in the analysis that fall within the LD blocks.

†Number of SNPs with MAF >0.05 among Cambodian parasites (*n* = 200) that were used to define LD blocks.

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)