Supporting Information

Park et al. 10.1073/pnas.1210585109

SI Results

We identified 32 regions under selection using the cross-population extended haplotype homozygosity (XP-EHH) test and identified top candidate mutations within each region associated with drug resistance using the efficient mixed-model association (EMMA) test (Dataset S1). Of the 163 mutations, 48 (29%) were intergenic; 6 (4%) were intronic; 33 (20%) conferred synonymous changes; and 76 (47%) conferred nonsynonymous changes. We evaluated the annotations for the 59 genes on this list using gene and protein prediction algorithms through PlasmoDB.org (1) and associated links, including predicted Gene Ontology function, pathway, interpro domains, or user comments, combined with any published literature for each gene. The great majority of the 59 genes (65%) can be collectively classified into the following categories: surface molecules or transporters (11/59 or 19%, of which 6/59 or 10% are transporters) including *pfcrt*; molecules involved in genome maintenance or transcriptional regulation (9/59 or 15%); metabolic enzymes (12/59 or 20%, of which 3/59 or 5% mediate lipid metabolism) including *dhfr*; and molecules involved in ubiquitination (6/59 or 10%). The remaining genes were determined as mediators of various other cellular functions including protein binding, invasion, and gamete fertilization (12/59 or 20%) or unclassified (9/59 or 15%).

We also analyzed all 35 genes within the chromosome 6 region (between position 1,117,269 and 1,390,662) found to be under selection in pyrimethamine-resistant parasites. This region contained a large stretch of intergenic mutations and it was difficult to localize the signal to any one gene. The region contains a large number of metabolic genes (12/35 or 34%, of which 3/35 or 9% participate in lipid metabolism; 2/35 or 6% mediate folate metabolism); chaperones and genes involved in ubiquitination (5/35 or 14%); with additional genes classified as genome maintenance or transcription regulation (8/35 or 23%); surface molecules or transporters (3/35 or 9%); other biological functions including structural proteins (2/35 or 6%); and the remainder (5/35 or 14%) as unclassified.

Molecules implicated in the ubiquitination cascade were mainly associated with resistance to pyrimethamine and include a putative E2 conjugating enzyme (PFL2100w), which likely acts as a ubiquitin E2 variant because of the lack of a catalytic cysteine and a HECT (homologous to the E6-AP carboxyl terminus) E3 ubiquitin ligase (MAL8P1.23). Within the chromosome 6 region there were several other molecules proposed to modulate ubiquitination including a HECT E3 (PFF1365c) and a Cullin-like E3 (PFF1445c). Two other molecules contain domains suggestive of a possible role in ubiquitination, including PF08_0080, which contains a PUB (peptide:*N*-glycanase/UBA or UBX) domain found in proteins linked to the ubiquitin proteasome system (2), and PFF1485w, which contains an ubiquitin-interacting motif. Also in this region are two putative chaperones, including a protein containing a Dna J domain (PFF1415c) associated with heat shock molecules (3) and a TRP (tetratricopeptide repeat protein) (PFF1505w) involved in RNA degradation (4), with a proposed chaperone function. Finally, there is a putative RING E3 (PFD0765w) in a region of selection associated with primaquine sensitivity.

Several genes putatively involved in lipid metabolism were identified in our regions of drug-associated selection, including an acyl-CoA synthetase, PfACS8 (PFB0695c) (5), and a putative phosphopantothenoylcysteine synthetase (PFD0610w, under selection in quinine-resistant parasites) proposed to be involved in CoA biosynthesis. Other lipid-metabolism-associated molecules in the chromosome 6 region include an acetyl-CoA synthetase (PFF1350c) (6), an ethanolaminephosphotransferase (PFF1375c-a/b) (7, 8), and a phosphatidylcholine-sterol acyltransferase precursor (PFF1420w). Finally, the PFD0350w gene, predicted to play a role in isoprenoid biosynthesis (9), is in a region under selection in artemisinin-sensitive parasites.

Folate pathway molecules in regions of selection specific to pyrimethamine-resistant parasites include the *dhfr* locus (PFD0830w), PF14_0487 (aminomethyltransferase), as well as PFF1360w (6-pyruvolytetrahydropterin synthetase, and PFF1490w (methenyltetrahydrofolate activity) found within the chromosome 6 region. Folate metabolism has been shown to be a target of pyrimethamine resistance mechanisms and, specifically, SNP changes in *dhfr* and *dhps*, as well as copy number variants in *gch1* (10), are associated with antifolate resistance (11).

There are three ABC transporters among the gene lists, including PF10_0049, MAL8P1.97, and PF08_0078, which are intriguing because these molecules have been shown to modulate drug responses in malaria (e.g., *pfmdr1*) and other organisms (12). Finally, we believe that there are a large number of molecules from among the genome maintenance or transcriptional regulation classification may be candidates for drug modulation through changes in gene expression (13), chromatin or histone structure (14, 15), or RNA binding (16).

SI Methods

Drug Assays. Drug assays were performed as previously described (17) with slight modifications for a 384-well format. Synchronized ring-stage parasites were cultured in the presence of serial dilutions of test compounds in 40 μL of RPMI supplemented with AlbuMAX II (Life Technologies; 1021-045) at 1.0% (by volume) hematocrit and an initial parasitemia of 1.0% (by cell count) in black clear-bottom plates (Greiner Bio-one; 781090). Following a 72-h incubation under standard culture conditions, SYBR Green I dye (Invitrogen; S7563) was added to a dilution of 1:5,000 and plates were stored at room temperature until the flourescence signal was read on a Spectramax M5 plate reader (Molecular Devices; ex 480 nm, em 530 nm). Raw flourescence data were analyzed using the Prism v5.0 software package (GraphPad). After background subraction and normalization, IC_{50} values were determined based on application of a nonlinear regression log(inhibitor)-response curve fit.

XP-EHH. Selection-association tests were run using the XP-EHH test (18). Replicate IC_{50} data were geometrically averaged [equivalently, $\log_{10}(IC_{50})$ data were arithmetically averaged] and then converted to binary phenotypes ("sensitive" vs. "resistant") according to cutoffs shown in Fig. S1. For drugs with a bimodal distribution, binary cutoffs were chosen at positions that clearly separated the sensitive and resistant populations. For drugs with a more unimodal distribution, cutoffs were manually placed at a distribution minimum near the median IC_{50}, because the XP-EHH test, like many other tests, loses power when either of the two populations becomes too small (when the cutoff is too far from the median). Although these may not represent samples that are especially sensitive or resistant in the traditional sense, it is common in studies of quantitative phenotypes to simply compare the upper part of a distribution against the lower part for binary tests (19).

The recombination map was constructed with LDhat v2.1 (20) using a block penalty of 5.0, 10 million reversible-jump Markov chain Monte Carlo (rjMCMC) iterations, a missing data cutoff of

20%, minimum minor allele frequency of 8%, and otherwise default parameters. Because the XP-EHH test does not tolerate missing data, SNPs with data in at least 80% of individuals were imputed with PHASE 2.1.1 (21). Because PHASE requires "diploid" data, we dropped the sample with the lowest call rate (SenP60.02) to create an even number of haploid individuals, randomly paired together; 83,540 fully imputed SNPs were polymorphic among the remaining 44 individuals. We then filtered out singleton SNPs and used only 29,605 SNPs that had at least two samples with a minor allele (minor allele frequency of 4%).

The XP-EHH test calculates haplotype decay separately for the resistant population and sensitive population using the EHH. The test then integrates these values with respect to genetic distance and computes a log ratio of these areas for the resistant population over the sensitive population. These log ratios, after normalization, are called XP-EHH scores or *z*-scores, as they are found to correspond to a normal distribution, with the exception of the tails that diverge from the null expectation (Fig. S4*B*). Other than its application to a phenotypically divided population (instead of a geographically divided population), the original algorithms were used as published (18) without any modifications. Significantly positive *z*-scores are indicative of positive selection among resistant parasites. Negative scores are indicative of selection in sensitive parasites. We used a two-sided conversion of *z*-scores to *P* values, but generally focused our attention on positive *z*-scores. It would be equally valid to do a one-sided, left- or right-tailed conversion for studies that are interested in specific selection scenarios.

We attempt to localize the signal in these regions by searching for the strongest EMMA signal in that window for that phenotype. We use this SNP to suggest a causal gene for the region (Dataset S1). We do not require significance from the EMMA test, as the region has already been identified as genome-wide significant by the XP-EHH test. We do not combine the results from these or any other statistics.

1. Aurrecoechea C, et al. (2009) PlasmoDB: A functional genomic database for malaria parasites. *Nucleic Acids Res* 37(Database issue):D539–D543.
2. Allen MD, Buchberger A, Bycroft M (2006) The PUB domain functions as a p97 binding module in human peptide N-glycanase. *J Biol Chem* 281:25502–25508.
3. Cyr DM, Langer T, Douglas MG (1994) DnaJ-like proteins: Molecular chaperones and specific regulators of Hsp70. *Trends Biochem Sci* 19:176–181.
4. Das AK, Cohen PW, Barford D (1998) The structure of the tetratricopeptide repeats of protein phosphatase 5: Implications for TPR-mediated protein-protein interactions. *EMBO J* 17:1192–1199.
5. Bethke LL, et al. (2006) Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. *Mol Biochem Parasitol* 150:10–24.
6. Prigge ST, He X, Gerena L, Waters NC, Reynolds KA (2003) The initiating steps of a type II fatty acid synthase in *Plasmodium falciparum* are catalyzed by pfACP, pfMCAT, and pfKASIII. *Biochemistry* 42:1160–1169.
7. Vial HJ, Thuet MJ, Philippot JR (1984) Cholinephosphotransferase and ethanolaminephosphotransferase activities in *Plasmodium knowlesi*-infected erythrocytes. Their use as parasite-specific markers. *Biochim Biophys Acta* 795:372–383.
8. Le Roch KG, et al. (2008) A systematic approach to understand the mechanism of action of the bisthiazolium compound T4 on the human malaria parasite, *Plasmodium falciparum*. *BMC Genomics* 9:513.
9. Jordão FM, Kimura EA, Katzin AM (2011) Isoprenoid biosynthesis in the erythrocytic stages of *Plasmodium falciparum*. *Mem Inst Oswaldo Cruz* 106(Suppl 1):134–141.
10. Kidgell C, et al. (2006) A systematic map of genetic variation in *Plasmodium falciparum*. *PLoS Pathog* 2:e57.
11. Nair S, et al. (2008) Adaptive copy number evolution in malaria parasites. *PLoS Genet* 4:e1000243.
12. Leprohon P, Légaré D, Ouellette M (2011) ABC transporters involved in drug resistance in human parasites. *Essays Biochem* 50:121–144.
13. Mok S, et al. (2011) Artemisinin resistance in *Plasmodium falciparum* is associated with an altered temporal pattern of transcription. *BMC Genomics* 12:391.
14. Cui L, Miao J (2010) Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. *Eukaryot Cell* 9:1138–1149.
15. Coleman BI, Duraisingh MT (2008) Transcriptional control and gene silencing in *Plasmodium falciparum*. *Cell Microbiol* 10:1935–1946.
16. Meng X, et al. (2012) Cytoplasmic Metadherin (MTDH) provides survival advantage under conditions of stress by acting as RNA-binding protein. *J Biol Chem* 287: 4485–4491.
17. Plouffe D, et al. (2008) In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc Natl Acad Sci USA* 105: 9059–9064.
18. Sabeti PC, et al.; International HapMap Consortium (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.
19. Bacanu SA, Nelson MR, Whittaker JC (2011) Comparison of methods and sampling designs to test for association between rare variants and quantitative traits. *Genet Epidemiol* 35:226–235.
20. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–1241.
21. Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73:1162–1169.

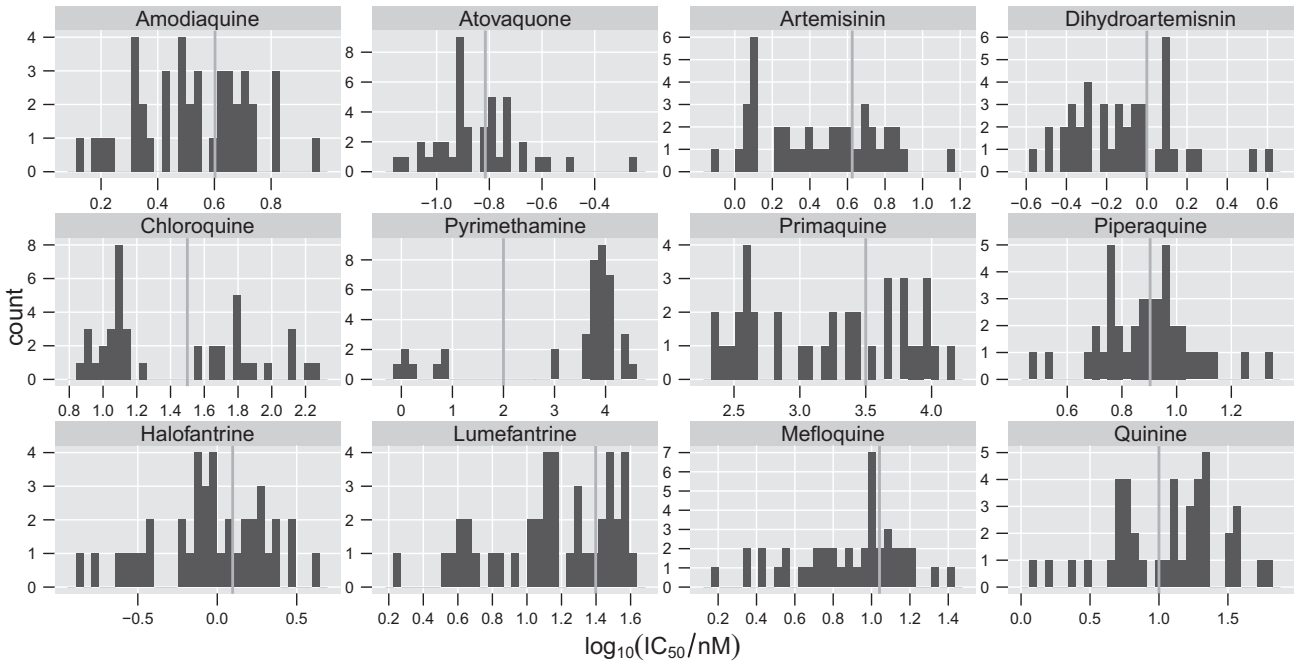

Fig. S1. Drug response distributions for the 12 drugs used in this study: histograms of $\log_{10}$ (IC_{50}) across 45 strains. Dark gray lines indicate binary cutoff values used for the XP-EHH test.

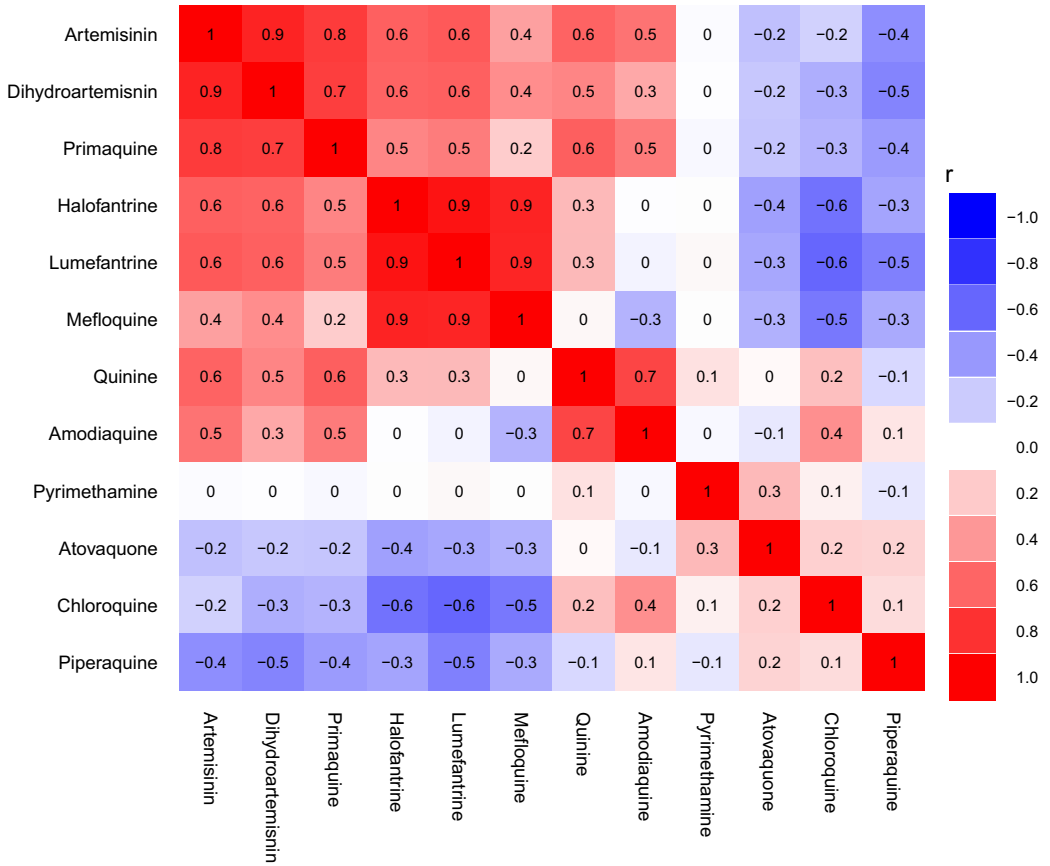

Fig. S2. Drug response correlation heat map for the 12 drugs used in this study. Pearson correlations are rendered for each pair of drugs, based on $\log_{10}$ (IC_{50}) values for each strain.

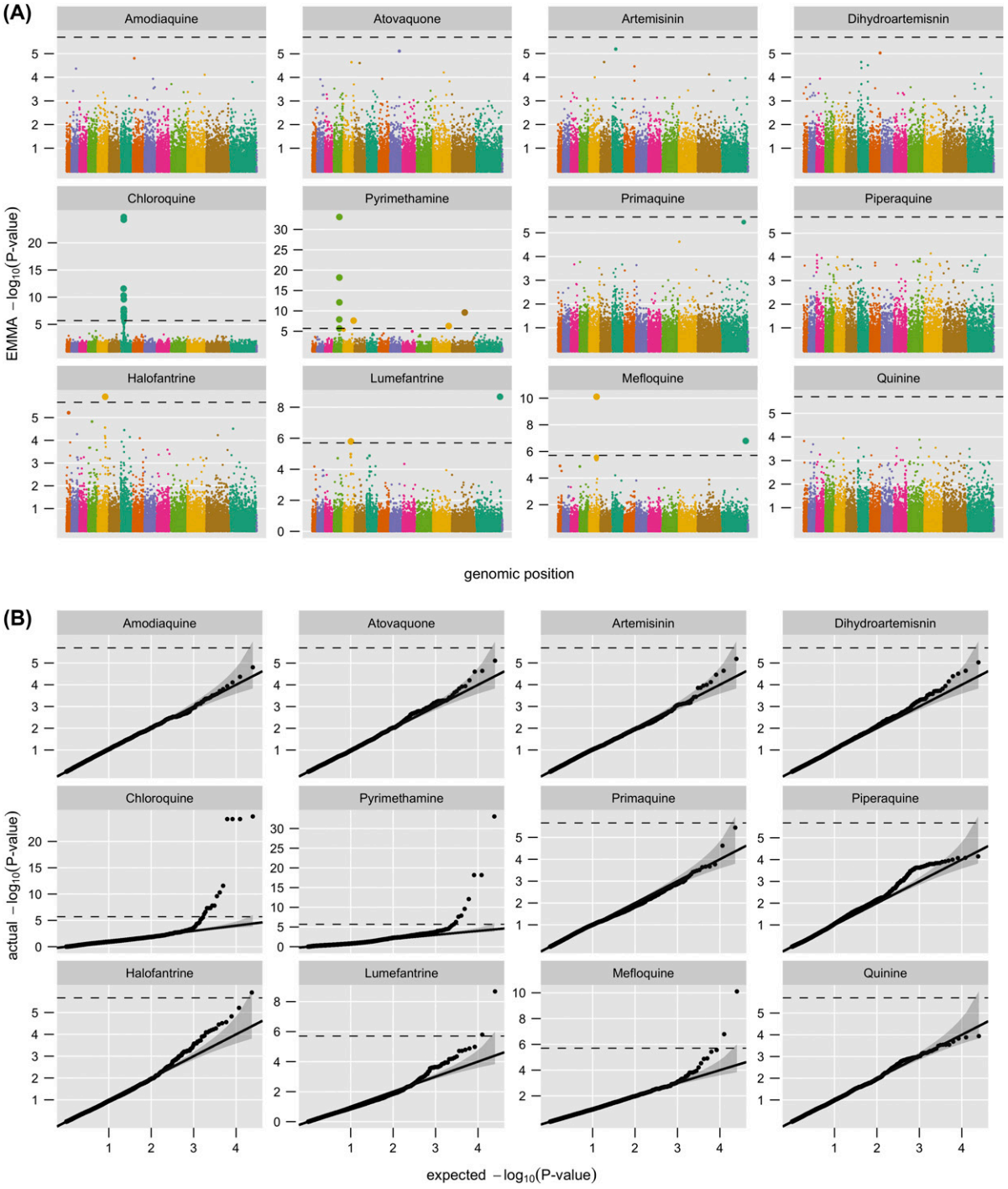

Fig. S3. EMMA genome-wide association study (GWAS) plots (sequence data, 45 samples). (*A*) EMMA Manhattan plots of $-\log_{10}$ (*P*) against genomic position. (*B*) EMMA P-P plots against the expected uniform distribution. Dashed line indicates Bonferroni-corrected significance of 5%. The shaded area indicates a 95% confidence interval around the null.

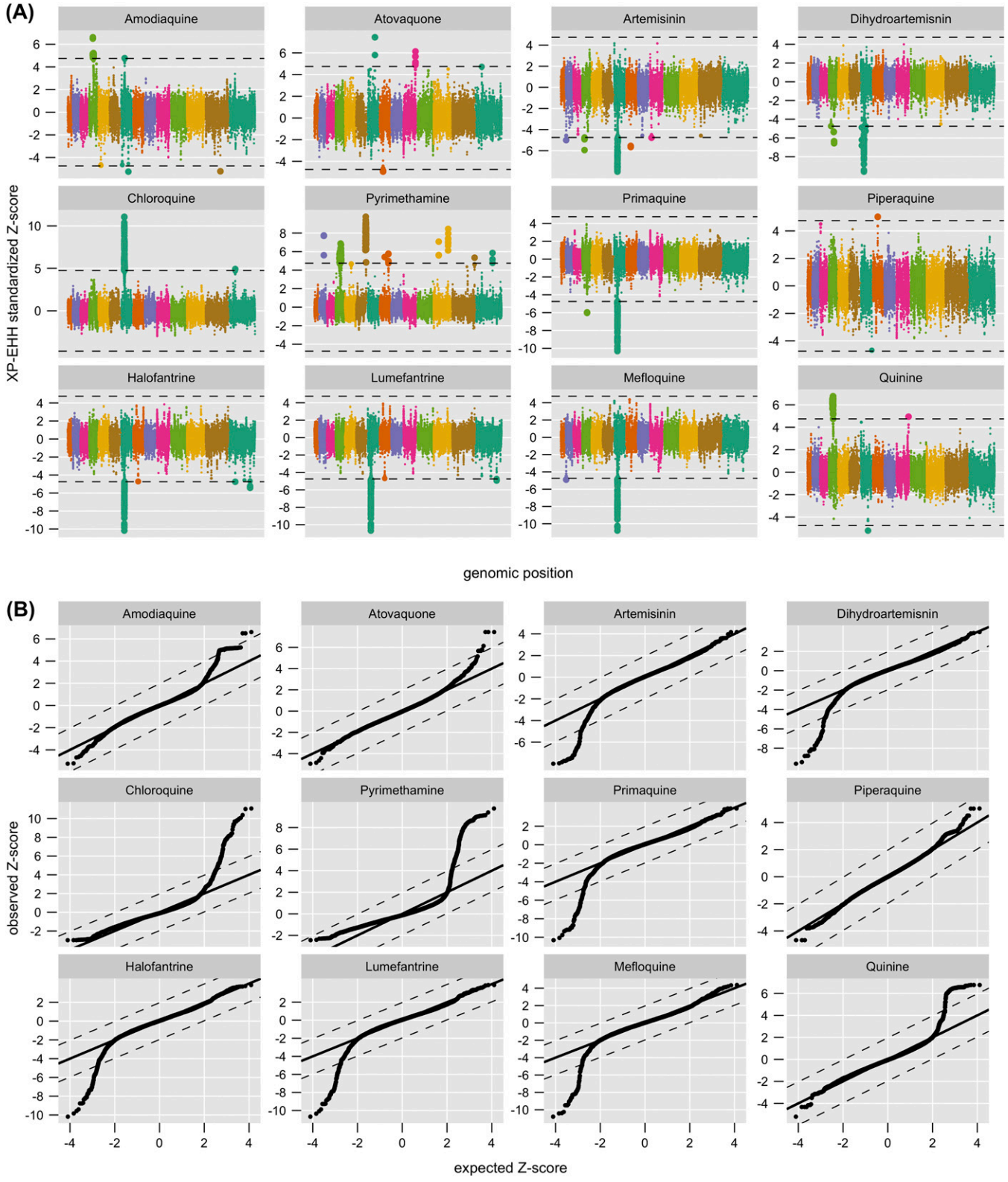

Fig. S4. XP-EHH GWAS plots (sequence data, 45 samples). (*A*) XP-EHH Manhattan-like plots of *z*-scores against genomic position. (*B*) XP-EHH Q-Q plots of XP-EHH *z*-scores against the expected normal distribution. Dashed lines indicate a 95% confidence interval around the null.

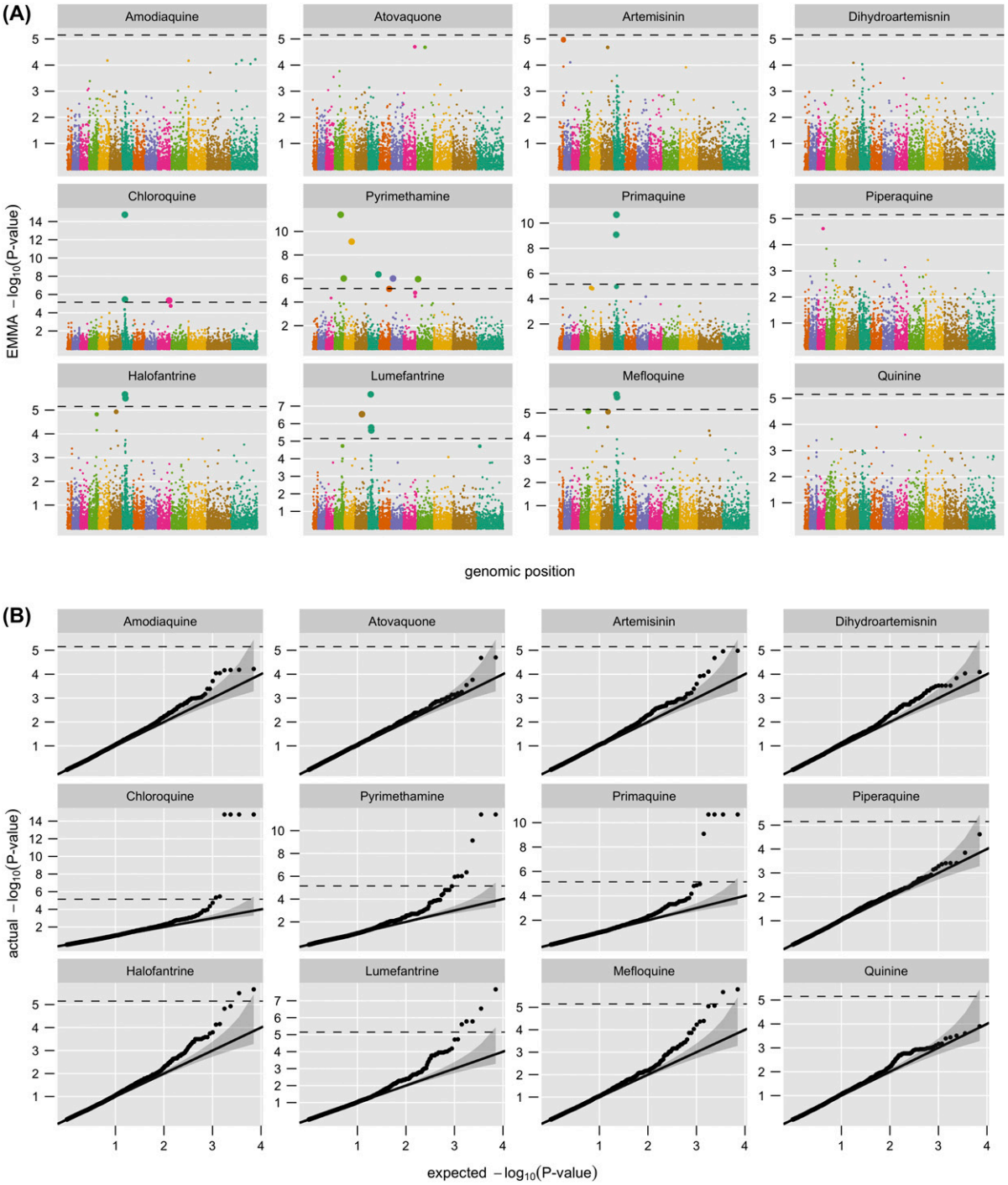

Fig. S5. EMMA GWAS plots (array data, 24 samples). (*A*) EMMA Manhattan plots of $-\log_{10}$ (*P*) against genomic position. (*B*) EMMA P-P plot against the expected uniform distribution. Dashed line indicates Bonferroni-corrected significance of 5%. The shaded area indicates a 95% confidence interval around the null.

Other Supporting Information Files

Dataset S1 (XLSX)