Genetic architecture of artemisinin-resistant *Plasmodium falciparum*

Olivo Miotto[1–3,25], Roberto Amato[1,2,4,25], Elizabeth A Ashley[3,5], Bronwyn MacInnis[1,2], Jacob Almagro-Garcia[1,2,4], Chanaki Amaratunga[6], Pharath Lim[6,7], Daniel Mead[1], Samuel O Oyola[1], Mehul Dhorda[5,8,9], Mallika Imwong[10], Charles Woodrow[3,5], Magnus Manske[1,2], Jim Stalker[1,2], Eleanor Drury[1], Susana Campino[1,2], Lucas Amenga-Etego[2,11], Thuy-Nhien Nguyen Thanh[12], Hien Tinh Tran[5,12], Pascal Ringwald[13], Delia Bethell[14], Francois Nosten[3,5,15], Aung Pyae Phyo[3,5,15], Sasithon Pukrittayakamee[10], Kesinee Chotivanich[10], Char Meng Chuor[7], Chea Nguon[7], Seila Suon[7], Sokunthea Sreng[7], Paul N Newton[5,16], Mayfong Mayxay[5,16,17], Maniphone Khanthavong[18], Bouasy Hongvanthong[18], Ye Htut[19], Kay Thwe Han[19], Myat Phone Kyaw[19], Md Abul Faiz[20], Caterina I Fanello[3,5], Marie Onyamboko[5,21], Olugbenga A Mokuolu[22], Christopher G Jacob[8], Shannon Takala-Harrison[8], Christopher V Plowe[8,23], Nicholas P Day[3,5], Arjen M Dondorp[3,5], Chris C A Spencer[2,4], Gilean McVean[2,4,24], Rick M Fairhurst[6], Nicholas J White[3,5] & Dominic P Kwiatkowski[1,2,4]

We report a large multicenter genome-wide association study of *Plasmodium falciparum* resistance to artemisinin, the frontline antimalarial drug. Across 15 locations in Southeast Asia, we identified at least 20 mutations in *kelch13* (PF3D7_1343700) affecting the encoded propeller and BTB/POZ domains, which were associated with a slow parasite clearance rate after treatment with artemisinin derivatives. Nonsynonymous polymorphisms in *fd* (ferredoxin), *arps10* (apicoplast ribosomal protein S10), *mdr2* (multidrug resistance protein 2) and *crt* (chloroquine resistance transporter) also showed strong associations with artemisinin resistance. Analysis of the fine structure of the parasite population showed that the *fd*, *arps10*, *mdr2* and *crt* polymorphisms are markers of a genetic background on which *kelch13* mutations are particularly likely to arise and that they correlate with the contemporary geographical boundaries and population frequencies of artemisinin resistance. These findings indicate that the risk of new resistance-causing mutations emerging is determined by specific predisposing genetic factors in the underlying parasite population.

Preventing the spread of *P. falciparum* resistance to artemisinin derivatives, the frontline drugs for severe and uncomplicated malaria, is an urgent priority for global health[1]. Although artemisinin derivatives remain effective, the rate at which they clear malaria parasites from the blood has progressively declined in Southeast Asia for at least 5 years, threatening control strategies based on the overall efficacy of artemisinin combination therapies (ACT)[2–6]. This is of serious concern, as the spread across Asia to Africa of resistance to previous frontline drugs, chloroquine and sulfadoxine-pyrimethamine

(Fansidar), originated in exactly the same region[7]. It is particularly important to prevent the development of high-level artemisinin resistance (loss of artemisinin efficacy) and the spread of resistance to Africa, where most malaria deaths occur.

To develop an effective strategy to combat drug resistance and contain it within Southeast Asia, it is imperative to understand the genetic factors that determine how it emerges and spreads. Artemisinin resistance in *P. falciparum* was recently associated with multiple SNPs in a gene on chromosome 13, referred to here as *kelch13*, mapping to

the β-propeller domain of the encoded kelch-like protein, PF3D7_1343700 (ref. 8), but many questions remain. What other genes are involved? Is the spread of resistance due mainly to parasite migration or to the emergence of new mutations? Why do new resistance-conferring mutations emerge in some places more often than others? How does this observation relate to the underlying genetic structure of the parasite population?

Here we report new insights into the genetic architecture of the artemisinin-resistant *P. falciparum* parasites that are spreading throughout Southeast Asia, based on analysis of 1,612 samples from 15 locations in Cambodia, Vietnam, Laos, Thailand, Myanmar and Bangladesh. *P. falciparum* genome sequencing and genotype calling at >600,000 SNP positions was performed on all samples. We first conducted a genome-wide association study (GWAS) with clinical data on 928 individuals with malaria recruited during 2011–2013 by the Tracking Resistance to Artemisinin Collaboration (TRAC) and 135 affected individuals recruited during 2009–2010 by the National Institute of Allergy and Infectious Diseases,US National Institutes of Health (**Table 1**). We then performed a population genetics analysis, incorporating additional samples collected by multiple projects in the same region (**Supplementary Table 1**) to translate our GWAS findings into an understanding of genetic architecture.

RESULTS

Genome-wide association study

We analyzed data on 1,063 individuals receiving artemisinin derivative treatment for *P. falciparum* malaria at 13 locations in Cambodia, Vietnam, Laos, Thailand, Myanmar, Bangladesh, Democratic Republic of the Congo and Nigeria. Parasite response to the drug was evaluated by determining parasite densities in blood samples collected every 6 h after admission (Online Methods). Clinical phenotype was expressed as the parasite clearance half-life (PC $t_{1/2}$), representing the time taken for artesunate to reduce parasite density by half during the log-linear decline in parasite densities; the findings at each location have been reported elsewhere[6,9].

We performed whole-genome sequencing on pretreatment blood samples, after enrichment of parasite DNA by leukocyte depletion[10], using an Illumina sequencing platform. Sequence reads were aligned against the *P. falciparum* 3D7 reference genome and combined with a collection of worldwide samples to discover variants and perform quality control based on genome coverage and other metrics (Online Methods)[11]. This process identified 681,587 high-quality exonic SNPs in the global data set, from which we obtained a data set of 18,322 SNPs with minor allele frequency (MAF) > 0.01 that were well covered in a set of 1,063 samples used for GWAS analysis. The association between SNP genotypes and PC $t_{1/2}$ (treated as a continuous dependent variable) was analyzed using a linear regression mixed-model algorithm, implemented in FaST-LMM[12]. This algorithm corrected for the confounding effect of population structure by treating genetic similarity as a random effect, reducing the genomic inflation factor λ_{GC} from 14.24 to 1.003 (**Supplementary Fig. 1**). At each SNP, samples for which the genotype was missing (for example, owing to low coverage) or heterozygous (for example, owing to mixed infection) were excluded from the test.

The GWAS identified strong signals of association ($P < 1 \times 10^{-7}$) at nine independent loci (**Fig. 1** and **Table 2**). The significance threshold chosen is conservative if Bonferroni correction is applied to <20,000 SNP tests; **Supplementary Table 2** shows the results with a threshold of $P < 1 \times 10^{-5}$. The strongest signal of association ($P = 4 \times 10^{-26}$) was for a nonsynonymous SNP (referred to as k13-C580Y here) in *kelch13* producing a p.Cys580Tyr substitution in the encoded propeller domain. This gene is within a genomic region identified by previous association studies[13,14], and the SNP corresponds exactly to the most common *kelch13* variant found to be associated with artemisinin resistance[8,15].

There were strong signals of association with other SNPs causing nonsynonymous changes, including in *arps10* encoding a p.Val127Met substitution in apicoplast ribosomal protein S10 ($P = 1 \times 10^{-20}$; referred to as arps10-V127M here); in *fd* encoding a p.Asp193Tyr substitution in ferredoxin ($P = 3 \times 10^{-17}$; fd-D193Y); in *mdr2* encoding a p.Thr484Ile substitution in multidrug resistance protein 2 ($P = 2 \times 10^{-10}$; mdr2-T484I); in *pib7* encoding a p.Cys1484Phe substitution in putative phosphoinositide-binding protein ($P = 4 \times 10^{-10}$; pib7- C1484F); in *crt* encoding p.Ile356Thr and p.Asn326Ser substitutions in chloroquine resistance transporter ($P = 7 \times 10^{-10}$; crt-I356T and crt-N326S, respectively); and in *pph* encoding a p.Val1157Leu substitution in protein phosphatase ($P = 8 \times 10^{-8}$; pph-V1157L). Although previous GWAS did not identify these additional loci, two genes in close proximity to *arps10* were reported to be mutated along with *kelch13* during *in vitro* selection for artemisinin resistance[8]. In the following sections, we examine these loci and their interrelationships in more detail.

### Diversity and phenotypic effect of resistance alleles

Previous work has associated artemisinin resistance with multiple SNPs in *kelch13* corresponding to the propeller domain. Only the k13-C580Y variant showed significant association in our analysis, but GWAS are poorly powered to detect associations with low-frequency variants, particularly when there is allelic heterogeneity. Detailed analysis of sequence variation identified 33 nonsynonymous SNPs in *kelch13* for which at least one sample had a homozygous call (**Supplementary Table 3**). When these SNPs were tested individually for association with parasite clearance, a clear pattern emerged: at least 20 of the 25 nonsynonymous changes in *kelch13* affecting the

Table 1 Geographical distribution of the samples used in the GWAS

Contributor	Years	Country (region)	Code	Location	Samples
TRAC	2011–2013	Bangladesh	BD	Ramu	50
		Myanmar	MM	Bago division	59
		Thailand (western)	WTH	Mae Sot	103
		Thailand (southern)	STH	Ranong	20
		Thailand (eastern)	ETH	Sisakhet	21
		Vietnam	VN	Binh Phuoc	97
		Laos	LA	Attapeu	77
		Cambodia (western)	WKH	Pursat, Pailin	185
		Cambodia (northern)	NKH	Preah Vihear	106
		Cambodia (northeastern)	NEKH	Ratanakiri	95
		Democratic Republic of the Congo	CD	Kinshasa	112
		Nigeria	NG	Ilorin	3
NIAID/NIH	2009–2010	Cambodia (western)	WKH	Pursat	89
		Cambodia (northeastern)	NEKH	Ratanakiri	46
Total					1,063

This table includes samples for which both phenotype (PC $t_{1/2}$) and genotype data were available. Abbreviations used in this paper to indicate the respective geographical regions are shown in the "Code" column. NIAID/NIH, National Institute of Allergy and Infectious Diseases of the US National Institutes of Health.

Figure 1 Manhattan plot showing the significance of SNP association in the GWAS. Each point represents 1 of the 18,322 SNPs with MAF > 0.01 in a set of 1,063 samples, colored according to chromosome. The x axis represents genomic location, and the y axis represents the P value for the SNP's association calculated using a linear mixed model (Online Methods). SNPs with $P \le 1 \times 10^{-7}$ after Bonferroni correction (n = 24; above the horizontal red line) are represented by diamond symbols. The nine loci containing these SNPs are identified in the plot and listed in **Table 2**. Polymorphisms with association $P \le 1 \times 10^{-5}$ (n = 24) are shown in a larger size between the horizontal orange and red lines and are listed in **Supplementary Table 2**.
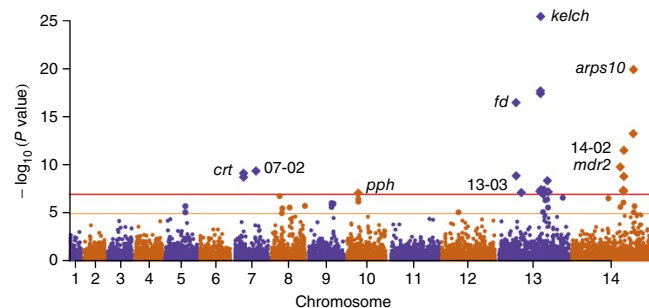
highly conserved BTB/POZ and propeller domains[8] were associated with prolonged PC $t_{1/2}$, whereas none of the variants in the upstream *P. falciparum*–specific portion of the gene were (**Supplementary Fig. 2** and **Supplementary Table 3**). Thirteen of these 20 mutations were previously observed to circulate in Cambodia[8]. On the basis of these findings, we classified samples into those with and without *kelch13* resistance alleles, defined here as nonsynonymous changes affecting the BTB/POZ and propeller domains with respect to the *P. falciparum* reference genome. The median PC $t_{1/2}$ was 6.5 h (interquartile range (IQR) = 5.4–7.8 h) for samples homozygous for a *kelch13* resistance allele and 2.6 h (2.0–3.3 h) for those without a *kelch13* resistance allele. No samples carried more than one resistance allele in the gene, apart from samples with clear evidence of mixed infection. The median PC $t_{1/2}$ was 5.9 h (4.4–7.1 h) for samples from Southeast Asia containing a mixture of parasites with and without *kelch13* resistance alleles. We found no homozygous *kelch13* mutations affecting the propeller domain in the African samples. Eight samples from Democratic Republic of the Congo carried different *kelch13* propeller domain–affecting mutations in the heterozygous state, but none of these was observed in more than three samples, and they were not associated with elevated PC $t_{1/2}$ (median = 1.9 h, range = 1.0–4.6 h).

We repeated the GWAS analysis treating *kelch13* resistance alleles collectively as a covariate. This adjustment resulted in a major reduction in the GWAS signals at other loci, but it did not abolish them completely. After correcting for the effect of *kelch13* variants, the estimated prolongation of PC $t_{1/2}$ was 0.54 h for mdr2-T484I ($P = 5 \times 10^{-5}$), 0.58 h for arps10-V127M ($P = 4 \times 10^{-5}$), 0.53 h for fd-D193Y ($P = 5 \times 10^{-4}$), 0.52 h for pph-V1157L ($P = 9 \times 10^{-5}$) and 0.47 h for crt-I356T ($P = 5 \times 10^{-4}$; **Supplementary Table 4**). A further iteration, in which the mdr2-T484I allele was added as a covariate along with *kelch13* variants, showed only a small residual effect for arps10-V127M and the other loci, implying that the phenotypic effects of these loci are not mutually independent.

In summary, we find that multiple mutations in *kelch13* affecting the BTB/POZ and propeller domains are the strongest predictors of prolonged PC $t_{1/2}$ across the genome. Other genomic loci are associated with PC $t_{1/2}$, largely owing to their population genetics relationship to *kelch13* resistance alleles.

Analysis of founder populations

To investigate the relationship between *kelch13* and other loci, we began by examining the founder effects previously observed in Cambodia, which appear to be due to the recent population expansion of multiple strains of artemisinin-resistant *P. falciparum*[16]. An iterative approach (Online Methods) was used to identify founder populations, defined here as distinct outlier clusters of samples showing loss of diversity and numerous polymorphisms with high F_{ST} values in comparison to the general population, consistent with founding events and recent population expansions. We identified seven founder populations strongly associated with artemisinin resistance: five in Cambodia and two in Vietnam (**Supplementary Figs. 3–5** and **Supplementary Tables 5–8**).

We identified SNP markers common to all artemisinin-resistant founder populations by performing a genome-wide analysis of the F_{ST} value between each artemisinin-resistant founder population and the artemisinin-sensitive core population in the same country and then combining the results across all founder populations (**Supplementary Table 9**). This analysis showed that fd-D193Y was the SNP most strongly associated with resistant founder populations and that other strong associations included crt-I356T, crt-N326S, arps10-V127M and mdr2-T484I (**Supplementary Table 9**). These findings coincide closely with the top signals of association in the GWAS (**Table 2**), with the notable exception of *kelch13* variants. We repeated this analysis to screen for genes containing multiple SNPs where each was a marker for a specific founder population (Online Methods). The *kelch13* locus ranked at the top, having five SNPs with

Table 2 Genomic loci most strongly associated with PC $t_{1/2}$ identified in the GWAS

Locus	Chr.	Top SNP					
		Position	Gene ID	Gene description	N/S	Alteration	P value
13-01 (*kelch*)	13	1,725,259	PF3D7_1343700	Kelch protein, putative	N	p.Cys580Tyr	4×10^{-26}
14-01 (*arps10*)	14	2,481,070	PF3D7_1460900.1	Apicoplast ribosomal protein S10 precursor, putative	N	p.Val127Met	1×10^{-20}
13-02 (*fd*)	13	748,395	PF3D7_1318100	Ferredoxin, putative	N	p.Asp193Tyr	3×10^{-17}
14-02	14	2,098,642	PF3D7_1451200	Conserved *Plasmodium* protein, unknown function	S	p.71Asn	3×10^{-12}
14-03 (*mdr2*)	14	1,956,225	PF3D7_1447900	Multidrug resistance protein 2+ (heavy metal transport family) (MDR2)	N	p.Thr484Ile	2×10^{-10}
07-02	7	896,660	PF3D7_0720700	Phosphoinositide-binding protein, putative	N	p.Cys1484Phe	4×10^{-10}
07-01 (*crt*)	7	405,600	PF3D7_0709000	Chloroquine resistance transporter (CRT)	N	p.Ile356Thr	7×10^{-10}
13-03	13	958,469	PF3D7_1322700	Conserved *Plasmodium* protein, unknown function	N	p.Thr236Ile	7×10^{-8}
10-01 (*pph*)	10	490,720	PF3D7_1012700	Protein phosphatase, putative	N	p.Val1157Leu	8×10^{-8}

This table shows nine loci identified across the genome that contained one or more SNPs significant with a Bonferroni-corrected threshold ($P \le 1 \times 10^{-7}$), ordered by increasing P value. For each locus, the SNP within the locus exhibiting the strongest signal is listed; for information about other results within these loci, refer to **Supplementary Table 2**. For each SNP, we show locus name; chromosome number (Chr.); nucleotide position; gene ID and description; whether the SNP is nonsynonymous (N) or synonymous (S); encoded alteration; and association P value. For ease of reference, some loci are named after their highest scoring gene.

Table 3 Allele frequencies for SNPs that are highly associated with artemisinin resistance

Locus	Chr.	Position	Gene ID	Alteration	GH	CD	BD	MM	TH	LA	VN-C	KH-C	VN-F01	VN-F04	WKH-F01	WKH-F02	WKH-F03	WKH-F04	NKH-F02
kelch	13	1725259	PF3D7_1343700	p.Cys580Tyr	0.00	0.00	0.00	0.11	0.14	0.00	0.00	0.00	0.00	0.14	**0.96**	0.00	**1.00**	0.00	**0.98**
kelch	13	1725340	PF3D7_1343700	p.Prp553Leu	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.00	**0.66**	0.00	0.00	0.00	0.00	0.00
kelch	13	1725370	PF3D7_1343700	p.Ile543Thr	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
kelch	13	1725382	PF3D7_1343700	p.Arg539Thr	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	**1.00**	0.00
kelch	13	1725521	PF3D7_1343700	p.Tyr493His	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00
fd	13	748395	PF3D7_1318100	p.Asp193Tyr	0.00	0.01	0.02	**0.65**	**0.90**	0.02	0.06	0.02	**1.00**	**0.82**	**0.98**	**1.00**	**1.00**	**1.00**	**0.98**
mdr2	14	1956225	PF3D7_1447900	p.Thr484Ile	0.00	0.00	0.06	**0.79**	**0.79**	0.22	0.28	0.22	**1.00**	**0.89**	**1.00**	**1.00**	**1.00**	**1.00**	**0.97**
arps10	14	2481070	PF3D7_1460900.1	p.Val127Met	0.00	0.00	0.31	0.49	**0.70**	0.12	0.13	0.08	**1.00**	0.07	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
crt	7	405362	PF3D7_0709000	p.Asn326Ser	0.01	0.29	**0.84**	**1.00**	**1.00**	0.12	0.14	0.06	**1.00**	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
crt	7	405600	PF3D7_0709000	p.Ile356Thr	0.02	0.29	0.01	**0.99**	**0.99**	0.13	0.15	0.05	**1.00**	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**
pph	10	490720	PF3D7_1012700	p.Val1157Leu	0.00	0.00	0.01	0.32	0.22	0.04	0.31	0.09	**1.00**	**0.92**	**1.00**	**1.00**	**1.00**	**0.80**	**0.99**

Allele frequencies in the seven artemisinin-resistant founder populations (columns on the right) are compared to those in the core populations of Vietnam (VN-C) and Cambodia (KH-C) and in other populations (GH, Ghana; CD, Democratic Republic of the Congo; BD, Bangladeshi; MM, Myanmar; TH, Thailand; LA, Laos). For each SNP, we show locus name; chromosome number (Chr.); nucleotide position; gene ID; encoded alteration; and frequency of the mutant or non-ancestral allele in various populations (bold indicates that the mutant allele is the majority allele).
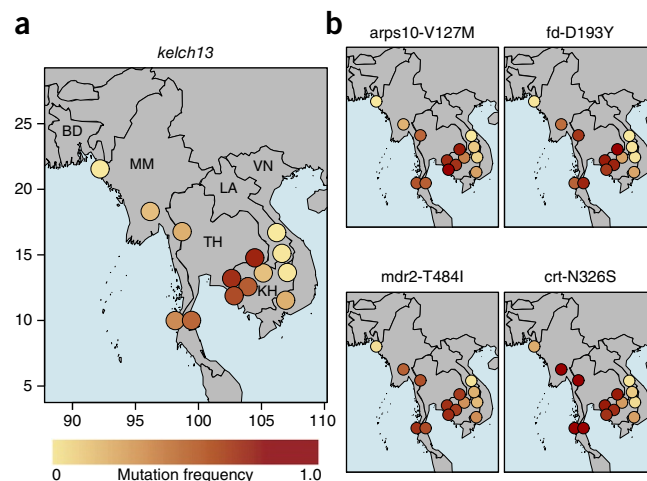

Figure 2 Distributions of genetic background mutations associated with artemisinin resistance across Southeast Asian sites. Mutant allele frequencies at each site are represented by colored circles, where deeper shades of red denote higher allele frequencies. Country codes correspond to those listed in **Table 1**. The two Bangladeshi sites, which are relatively close to each other, are combined because of the low sample sizes. (**a**) Distribution of *kelch13* resistance mutations (defined as any nonsynonymous mutation in *kelch13* affecting the BTB/POZ or propeller domains). (**b**) Distribution of four mutations identified as potential background mutations in artemisinin-resistant parasites: arps10-V127M, fd-D193Y, mdr2-T484I and crt-N326S. Representative geographical coordinates are included on the axes in **a**.

close to 100% frequency in specific resistant founder populations and close to 0% frequency in the artemisinin-sensitive core population (**Supplementary Fig. 6**, and **Supplementary Tables 10** and **11**). All seven founder populations were associated with a mutant *kelch13* allele: three carried the common k13-C580Y variant, and the remaining four had the k13-R539T, k13-Y493H, k13-I542T and k13-P553L variants (encoding p.Arg539Thr, p.Tyr493His, p.Ile542Thr and p.Pro553Leu substitutions, respectively).

In summary, each artemisinin-resistant founder population was strongly associated with a specific *kelch13* resistance allele. In addition, most founder populations in Cambodia and Vietnam shared the same alleles at *fd*, *crt*, *mdr2*, *aprs10* and other loci identified by conventional GWAS analysis (**Table 3**). By analogy with cancer genetics, these findings suggest a model in which *kelch13* mutations act as driver mutations for the emergence of the artemisinin-resistant parasite strains that have recently undergone population expansion in Cambodia and Vietnam. Independent *kelch13* mutations often but not invariably arise in combination with specific alleles at other positions in the genome, which we refer to here as 'background' alleles.

Geographical and genetic compartments of emerging resistance

To understand how artemisinin resistance is spreading across Southeast Asia, we constructed a map of *kelch13* resistance allele frequencies across all 15 sampling locations (**Fig. 2** and **Supplementary Table 12**). The overall picture is suggestive of diffusion from a central hotspot of high-frequency resistance in the area of western Cambodia, with intermediate-frequency resistance in Vietnam, Thailand and Myanmar and very low-frequency resistance in Laos and Bangladesh. However, there are areas of sharp discontinuity in the frequency of *kelch13* resistance alleles, for example, at the junction between

Figure 3 Population structure and distribution of *kelch13* mutants in Southeast Asia. (**a**) Neighbor-joining tree showing population structure across the 15 Asian sampling sites. Branches with colored tip symbols indicate that the samples are *kelch13* mutants, whereas those without tip symbols are wild type for *kelch13*. The circular subpanel shows a magnified view of the major branching points near the tree root. Mixed-infection samples (with a mixture of wild-type and mutant parasites) were omitted. (**b**) Map of the 15 sites showing the geographical location of samples in the tree. Bangladeshi (purple) and Thailand-Myanmar border region (WSEA; blue) samples form separate branches, whereas the lower Mekong region (ESEA) samples divide into two major groups, separating samples in high-resistance areas (red) from those in low-resistance areas (green); parasites in intermediate-resistance areas (yellow) are split between these two groups. Map colors indicate the endemicity of *P. falciparum* in 2010: light blue, high; darker blue, low; gray, absent (the map was adapted from http://www.map.ox.ac.uk/)[34].
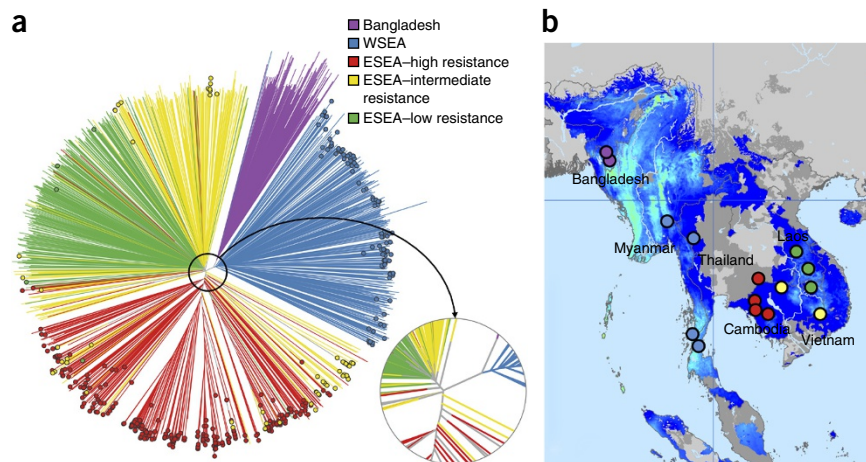
Cambodia, Thailand and Laos, raising the possibility that the spread of resistance is compartmentalized.

To investigate how the geographical distribution of resistance alleles relates to the genetic structure of the parasite population, we constructed a neighbor-joining tree grouping samples according to genome-wide genetic similarity (**Fig. 3a**). The tree showed three major compartments of population structure, corresponding to the western (WSEA) and eastern (ESEA) parts of Southeast Asia and to Bangladesh (BD). WSEA comprised Myanmar and western Thailand, and ESEA comprised Cambodia, Vietnam, Laos and eastern Thailand; these two major compartments of parasite population structure were separated by a malaria-free corridor running through the center of Thailand (**Fig. 3b**). In ESEA, there was considerable variation in the frequency of *kelch13* resistance alleles, with an area of high resistance comprising three locations in western Cambodia and eastern Thailand and an area of low resistance comprising three sites in Laos and northeastern Cambodia. Between the high-resistance and low-resistance areas were two locations in northern Cambodia and southern Vietnam with intermediate levels of resistance. Parasites from the high-resistance and low-resistance areas clearly separated in the neighbor-joining tree (**Fig. 3a**), whereas parasites from the intermediate-resistance area fell into two groups, one aligned with high resistance and the other with low resistance.

These findings raise a key question: what are the differences between the parasite populations residing on either side of the geographical boundary between high and low resistance? This data set provided two examples for such a comparison—the boundaries between WSEA and BD and between the high-resistance and low-resistance areas of ESEA. We performed a genome-wide screen for

SNPs showing high F_{ST} between resistant and non-resistant parasite population compartments (**Fig. 4** and **Supplementary Table 13**). In the WSEA-BD comparison, the strongest signals of differentiation included fd-D193Y ($F_{ST} = 0.64$), mdr2-T484I ($F_{ST} = 0.53$), crt-N326S ($F_{ST} = 0.53$) and arps10-V127M ($F_{ST} = 0.44$). In ESEA, when comparing areas of high- and low-resistance, the strongest signals of differentiation included fd-D193Y ($F_{ST} = 0.85$), crt-N326S ($F_{ST} = 0.71$), crt-I356T ($F_{ST} = 0.7$), arps10-V127M ($F_{ST} = 0.64$), pph-Y1133N ($F_{ST} = 0.5$) and mdr2-T484I ($F_{ST} = 0.44$). Combining these findings, the SNPs most clearly marking the geographical boundary of resistance were fd-D193Y, mdr2-T484I, crt-N326S and arps10-V127M. Allele frequency maps showed that the patterns of geographical variation in these alleles were remarkably similar to that for *kelch13* resistance alleles (**Fig. 2b**). We found that the arps10-V127M, fd-D193Y and mdr2-T484I alleles were rare or absent in 2 African populations (113 parasites from the Democratic Republic of the Congo and 475 parasites from Ghana), suggesting that they are the product of evolutionary selection within Southeast Asia (**Table 3**).

The WSEA and ESEA parasite populations were clearly distinct, both geographically and genetically, and they displayed some differences in the genetic features of artemisinin resistance. PC $t_{1/2}$ was somewhat longer in WSEA than in ESEA for parasites without resistance alleles ($P = 2 \times 10^{-5}$) and shorter for parasites with resistance alleles ($P = 4 \times 10^{-3}$), such that *kelch13* resistance alleles prolonged PC $t_{1/2}$ by a median of 3.2 h in WSEA in comparison to the time of 3.9 h in ESEA. In ESEA, where there was extreme heterogeneity in allele frequency with marked founder effects, background alleles were strongly associated with the presence of resistance alleles in individual samples, whereas in WSEA, where there was less evidence of founder effects, background alleles were present at high frequency but were weakly associated with resistance alleles in individual
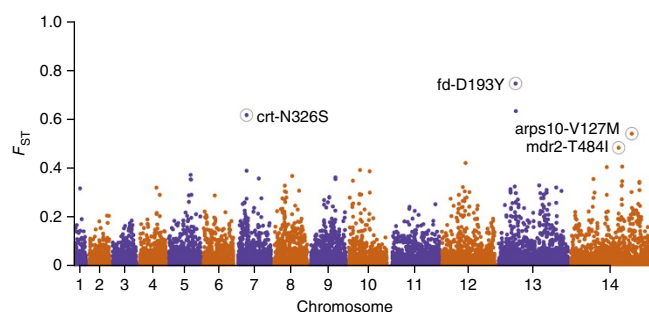

Figure 4 Genome-wide analysis of SNP differentiation between resistant and non-resistant geographical compartments. We selected two pairs of geographical compartments, each consisting of a resistant and a proximal non-resistant compartment: Bangladesh versus WSEA and high-resistance versus low-resistance sites in ESEA. For each pair, we estimated F_{ST} at every SNP across the genome and computed the mean of the two SNP estimates. In this Manhattan plot, mean F_{ST} is plotted at the corresponding genomic position. Polymorphisms at four loci associated with the artemisinin resistance genetic background are among the highest scorers and are highlighted by labeled circles.
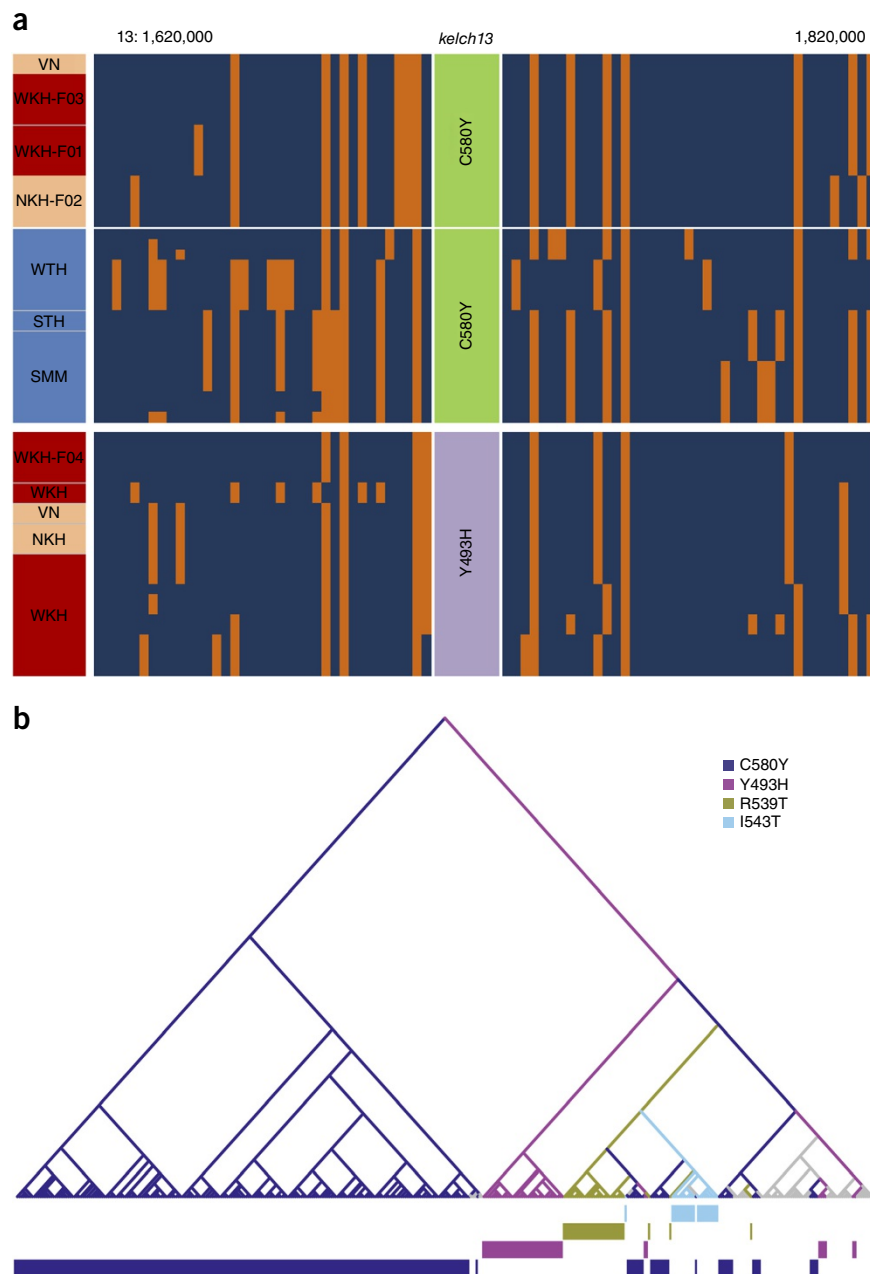
Figure 5 Analyses of the haplotypes surrounding *kelch13* in resistant mutants. (**a**) Haplotype diagram. Each vertical column represents a SNP (MAF ≥ 0.1 in our data set; within 100 kb on either side of the *kelch13* gene), and each horizontal line represents a sample; at the intersection, the color represents the genotyped allele in the sample: blue, reference; orange, alternative. Haplotypes are shown for a selection of samples carrying the mutations k13-C580Y (top) and k13-Y493H (bottom), organized by geographical compartment, as shown in the left-hand column (blue, WSEA; red, ESEA high resistance; pink, ESEA intermediate resistance). The samples' regions or founder populations are indicated. (**b**) Tree based on the LCHLs for samples containing *kelch13* mutations: samples with long shared haplotypes cluster closely. Tips represent samples and have been colored by the alteration present in the sample. Internal branches have been colored by the most frequent alteration present in the subtrees to which they lead. The bars at the bottom offer visual aid for tracking how different mutations cluster and segregate across the tree. Branches have been colored by the most frequent mutation present in the subtree they subtend. Colored bars at the bottom show which alteration corresponds to the tree tip directly above. Details of the method are discussed in the Online Methods.

samples (**Supplementary Table 14**). A possible contributory factor is the higher level of malaria transmission in WSEA, which tends to increase the likelihood of recombination between parasites of different genetic types and thus to decouple resistance alleles from the genetic background on which they originated.

Spread of resistance between population compartments

To what extent is the spread of artemisinin resistance across Southeast Asia due to the geographical migration of resistant parasites as opposed to multiple origins of resistance in different locations? Multiple origins are clearly an important factor, as we observed a wide repertoire of *kelch13* resistance alleles, most of which appeared to be localized in their geographical distribution (**Supplementary Table 15**). A notable exception was the k13-C580Y allele, present in 16% of the 1,612 samples in this study and observed at 3 locations in WSEA and 7 locations in ESEA. A simple reconstruction of haplotypes in the genomic flanking regions extending 100 kb on either side of the *kelch13* sequence encoding the propeller region showed a wide variety of haplotypic backgrounds surrounding the different resistance alleles (**Fig. 5a**). In our data set, most resistance alleles were associated with one or more unique haplotypes, suggesting that they have originated from separate mutational events. In some cases, the same *kelch13* allele was accompanied by different haplotypes, which may indicate that some mutations might have emerged independently multiple times, as recently suggested by another analysis[17].

To assess whether parasite migration has had a major role alongside independent emergence, we performed a more detailed demographic analysis. For each pair of *kelch13*-mutant samples, we estimated the longest common haplotype length (LCHL), that is, the nucleotide distance on either side of the *kelch13* gene over which the samples' haplotypes were identical. These estimates were then used to cluster samples that were likely to share the same recent demographic history (**Fig. 5b**). Samples carrying less common *kelch13* mutations tended to cluster by allele, as expected if each of these clusters originated from a different recent evolutionary event. In contrast, the majority of samples carrying the most common allele (k13-C580Y) formed a large branch comprising several clusters, consistent with a common origin of the k13-C580Y alleles shared by different subpopulations in Cambodia. However, we observed two separate clusters of Cambodian k13-C580Y mutants, whose flanking haplotypes were similar to those of parasites carrying other *kelch13* mutations (k13-R529T and k13-I543H, encoding p.Arg529Thr and p.Ile543His substitutions, respectively), suggesting that this allele might have

emerged independently multiple times in ESEA. We also found that k13-C580Y mutants in western Thailand occupied a separate branch from those in ESEA and shared core haplotypes with other WSEA mutants, suggestive of an independent mutational event, a conclusion supported by the short length of the haplotype shared by ESEA and WSEA parasites (**Supplementary Fig. 7**). For a number of the most common alleles (k13-C580Y, k13-I543T, k13-Y493H and k13-R539T), we found clusters containing ESEA parasites from more than one country (Cambodia, Vietnam and eastern Thailand), indicating that mutants have crossed international borders, at least within this region (**Supplementary Fig. 8**). However, we found no evidence that k13-C580Y mutants from ESEA might have migrated to the WSEA region or vice versa, consistent with the observation that k13-C580Y mutants are genetically more similar to samples from their own geographical region than to k13-C580Y mutants in other regions (**Supplementary Fig. 9**).

DISCUSSION

This large multicenter GWAS shows that the major genomic locus controlling *P. falciparum* resistance to artemisinin in Southeast Asia at the present time is *kelch13*. We identify at least 20 distinct resistance alleles—alleles associated with artemisinin resistance—arising from multiple independent mutations in the *kelch13* sequences encoding the propeller and BTB/POZ domains. Most resistance alleles appear to be localized to a relatively small geographical area, and we find that the most widespread resistance allele, k13-C580Y, has originated independently in multiple locations. We conclude that the spread of artemisinin resistance across several countries in Southeast Asia is primarily due to the proliferation of newly emerging mutations in the *kelch13* sequences encoding the propeller and BTB/POZ domains.

Understanding the factors that lead to the emergence of new *kelch13* mutations is therefore central to the problem of containing and controlling artemisinin resistance. Geographical location is clearly a major risk factor: *kelch13* resistance alleles are well established in Cambodia, Vietnam, Thailand and Myanmar but are absent or found at much lower frequency in Laos and Bangladesh. Mutations in *kelch13* affecting the propeller and BTB/POZ domains were also found to be rare in African samples. Thus, the key question is what differentiates the *P. falciparum* in regions in Southeast Asia from those in neighboring countries and other parts of the world. A wide range of epidemiological factors could potentially be involved, including the intensity of transmission, the vector species and antimalarial drug usage, but the fact that the problem has emerged independently in several countries with different levels of malaria intensity and different treatment policies makes it somewhat unlikely that such factors are the sole cause.

We find that mutations in *kelch13* mapping to the propeller and BTB/POZ domains are mostly likely to arise on a particular genetic background that is common in parts of Southeast Asia, and we identify the strongest markers of this genetic background to be nonsynonymous variants of *arps10* on chromosome 14 and *fd* on chromosome 13. Other background markers include nonsynonymous mutations in *mdr2* on chromosome 14 and *crt* on chromosome 7. The association between background markers and *kelch13* mutations operates at multiple levels. Local variations in the proportion of samples carrying *kelch13* resistance alleles are correlated with background marker frequency. Background markers show high levels of genetic differentiation at the geographical boundary between areas of high and low resistance and are absent or at much lower frequency in other parts of the world. Different artemisinin-resistant founder populations with independent *kelch13* driver mutations tend to share the

same background alleles. Background markers are strongly associated with slow parasite clearance rates after correcting for population structure using a linear mixed-model GWAS analysis, and these associations are greatly attenuated when *kelch13* resistance alleles are collectively treated as a covariate. Thus, the background markers can be regarded as markers of the risk that an artemisinin-sensitive parasite in Southeast Asia will acquire a *kelch13* mutation that makes it artemisinin resistant.

There are several ways in which genetic background could influence risk of the emergence of a new resistance-conferring mutation. The polypeptide sequences of kelch propeller domains are highly conserved across *Plasmodium* species, raising the possibility that *kelch13* mutations carry a biological fitness cost, as is the case for many drug resistance mutations. Genetic background might reduce the fitness cost through compensatory mutations elsewhere in the genome. Alternatively, these background mutations might boost the selective advantage of *kelch13* mutations by enhancing their phenotypic effects on artemisinin resistance. It is certainly plausible that different components of the background have different roles: the observation that mutations in the close neighborhood of *arps10* emerged alongside *kelch13* mutations during the *in vitro* development of artemisinin resistance[8] points to a possible interaction between *arps10* and *kelch13*, whereas we estimated a marginal contribution to PC $t_{1/2}$ by the *fd* mutation. We also considered the hypothesis that the genetic background might have been selected for conferring resistance to an ACT partner drug, but, given that it is present across a region where at least three different partner drugs (piperaquine, mefloquine and amodiaquine) have been used, this seems unlikely. A further possibility is that the genetic background is simply a marker for a particular evolutionary niche in which the biological fitness costs and benefits are altered from the norm, for example, where most parasites already have high levels of multidrug resistance. This raises the questions of why the background markers have grown to high frequency in Southeast Asia and what their functional role is. It is clear that they emerged considerably before *kelch13* mutations and that they have at best a small effect on artemisinin resistance.

Two prominent background markers affect the apicoplast proteins encoded in the nuclear genome, apicoplast ribosomal protein S10 and ferredoxin. The apicoplast is a relict plastid with a range of metabolic functions, and the apicoplast ribosomal protein complex is the target of clindamycin and tetracycline, which have both been used as antimalarial drugs[18]. Ferredoxin is a key component of the apicoplast electron transport chain[19] and might therefore affect the parasite's ability to withstand the oxidant stress created by artemisinin. The ferredoxin pathway has been implicated in the mode of action of primaquine[20], and halofantrine resistance has been associated with a SNP in the *fd* locus, PF3D7_1318300 (ref. 21). Two other background markers affect transporter genes localized in the digestive vacuole of the parasite, the multidrug resistance transporter 2 and chloroquine resistance transporter[22,23]. The role of *mdr2* in antimalarial drug resistance remains unclear, but it has been implicated in resistance to antifolates[24,25] and tolerance to heavy metals[26]. In *crt*, there are two closely linked background markers, both nonsynonymous variants affecting the same part of the membrane-spanning structure, some distance from the p.Lys76Thr variant that is the major determinant of chloroquine resistance[27].

In the past 60 years, the lower Mekong region has been the epicenter of drug resistance. Repeatedly, resistance has emerged here before spreading to the rest of the world, successively undermining the effectiveness of chloroquine, sulfadoxine-pyrimethamine and mefloquine[28]. The vast majority of parasites in this region

are multidrug resistant, possessing both *crt* mutations encoding p.Lys76Thr and *dhfr* mutations encoding p.Ser108Asn, the critical determinants of chloroquine and pyrimethamine resistance, respectively. Several factors have been proposed to explain such a propensity for drug resistance, such as low transmission rates promoting high rates of parasite inbreeding[16], a checkered history of public health interventions[29] and the widespread availability of poor-quality antimalarials[30]. It has also been hypothesized that *P. falciparum* in Southeast Asia may have a genetic predisposition to develop resistance-causing mutations[31]; thus far, this hypothesis has been backed by inference from experimental systems[32,33]. Here we describe the first clear epidemiological evidence, to our knowledge, for such an effect. Scientifically, these data provide a foundation for investigating the complex multistage processes by which antimalarial drug resistance evolves in natural parasite populations. Practically, these findings raise concern that, unless efforts to eliminate malaria from the greater Mekong subregion are rapidly and successfully implemented, further mutations may lead to high-level resistance both to artemisinin and to ACT partner drugs. Although artemisinin resistance appears to have made little progress in other parts of the world, the situation might change as the phenotype continues to evolve, and understanding the multiple genetic factors involved in this process may provide vital clues about how to prevent its spread.

URLs. MalariaGEN *Plasmodium falciparum* Community Project, http://www.malariagen.net/projects/parasite/pf; European Nucleotide Archive (ENA), http://www.ebi.ac.uk/ena/; *P. falciparum* 3D7 reference sequence V3, ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.latest_version/version3/; Parasite Clearance Estimator, Worldwide Antimalarial Resistance Network (WWARN), https://www.wwarn.org/toolkit/data-management/parasite-clearance-estimator; R language, http://www.r-project.org/; R language ape package, http://ape-package.ird.fr/; R language stats package, http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html; ADMIXTURE program, http://www.genetics.ucla.edu/software/admixture/; PLINK toolset, http://pngu.mgh.harvard.edu/~purcell/plink/.

METHODS
Methods and any associated references are available in the online version of the paper.

Accession codes. A document containing lists of ENA accession codes for all samples used in the present study is available from http://www.malariagen.net/data/pf-sample-info.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS
C.A., P.L., M.D., M.I., L.A.-E., T.-N.N.T., H.T.T., D.B., F.N., A.P.P., S.P., K.C., C.M.C., C.N., S. Suon, S. Sreng, P.N.N., M. Mayxay, M.K., B.H., Y.H., K.T.H., M.P.K., M.A.F., C.I.F., M.O. and O.A.M. carried out field and laboratory work to obtain *P. falciparum* samples for sequencing. E.A.A., C.A., P.L., T.-N.N.T., H.T.T., D.B., F.N., A.P.P., S.P., K.C., C.M.C., C.N., P.N.N., M. Mayxay, M.K., B.H., Y.H., K.T.H., M.P.K., M.A.F., C.I.F., M.O. and O.A.M. carried out clinical studies to obtain parasite clearance data. D.M., S.O.O., E.D., S.C. and B.M. developed and implemented methods for sample processing and sequencing library preparation. J.S. and M. Manske managed data production pipelines. E.A.A., C.W., P.R., C.G.J., S.T.-H., C.V.P., N.P.D., A.M.D., R.M.F., N.J.W., O.M., B.M. and D.P.K. contributed to study design and management. O.M., R.A., J.A.-G., C.C.A.S., G.M. and D.P.K. performed data analyses. O.M., R.A., J.A.-G., C.C.A.S., G.M. and D.P.K. performed data analyses. O.M., R.A. and D.P.K. drafted the manuscript, which was reviewed by all authors.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Dondorp, A.M. *et al.* The threat of artemisinin-resistant malaria. *N. Engl. J. Med.* **365**, 1073–1075 (2011).
2. Dondorp, A.M. *et al.* Artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* **361**, 455–467 (2009).
3. Phyo, A.P. *et al.* Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. *Lancet* **379**, 1960–1966 (2012).
4. Hien, T.T. *et al. In vivo* susceptibility of *Plasmodium falciparum* to artesunate in Binh Phuoc Province, Vietnam. *Malar. J.* **11**, 355 (2012).
5. Kyaw, M.P. *et al.* Reduced susceptibility of *Plasmodium falciparum* to artesunate in southern Myanmar. *PLoS ONE* **8**, e57689 (2013).
6. Amaratunga, C. *et al.* Artemisinin-resistant *Plasmodium falciparum* in Pursat province, western Cambodia: a parasite clearance rate study. *Lancet Infect. Dis.* **12**, 851–858 (2012).
7. Mita, T., Tanabe, K. & Kita, K. Spread and evolution of *Plasmodium falciparum* drug resistance. *Parasitol. Int.* **58**, 201–209 (2009).
8. Ariey, F. *et al.* A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature* **505**, 50–55 (2014).
9. Ashley, E.A. *et al.* Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* **371**, 411–423 (2014).
10. Venkatesan, M. *et al.* Using CF11 cellulose columns to inexpensively and effectively remove human DNA from *Plasmodium falciparum*–infected whole blood samples. *Malar. J.* **11**, 41 (2012).
11. Manske, M. *et al.* Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487**, 375–379 (2012).
12. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).
13. Takala-Harrison, S. *et al.* Genetic loci associated with delayed clearance of *Plasmodium falciparum* following artemisinin treatment in Southeast Asia. *Proc. Natl. Acad. Sci. USA* **110**, 240–245 (2013).
14. Cheeseman, I.H. *et al.* A major genome region underlying artemisinin resistance in malaria. *Science* **336**, 79–82 (2012).
15. Ashley, E.A. *et al.* Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* **371**, 411–423 (2014).
16. Miotto, O. *et al.* Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nat. Genet.* **45**, 648–655 (2013).
17. Takala-Harrison, S. *et al.* Independent emergence of artemisinin resistance mutations among *Plasmodium falciparum* in Southeast Asia. *J. Infect. Dis.* doi:10.1093/infdis/jiu491 (1 September 2014).
18. Dahl, E.L. & Rosenthal, P.J. Apicoplast translation, transcription and genome replication: targets for antimalarial antibiotics. *Trends Parasitol.* **24**, 279–284 (2008).
19. Kimata-Ariga, Y., Saitoh, T., Ikegami, T., Horii, T. & Hase, T. Molecular interaction of ferredoxin and ferredoxin–NADP+ reductase from human malaria parasite. *J. Biochem.* **142**, 715–720 (2007).
20. Vásquez-Vivar, J. & Augusto, O. Hydroxylated metabolites of the antimalarial drug primaquine. Oxidation and redox cycling. *J. Biol. Chem.* **267**, 6848–6854 (1992).
21. Van Tyne, D. *et al.* Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in *Plasmodium falciparum*. *PLoS Genet.* **7**, e1001383 (2011).

22. Fidock, D.A. *et al.* Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* **6**, 861–871 (2000).

23. Zalis, M.G., Wilson, C.M., Zhang, Y. & Wirth, D.F. Characterization of the *pfmdr2* gene for *Plasmodium falciparum. Mol. Biochem. Parasitol.* **62**, 83–92 (1993).

24. Martinelli, A., Henriques, G., Cravo, P. & Hunt, P. Whole genome re-sequencing identifies a mutation in an ABC transporter (*mdr2*) in a *Plasmodium chabaudi* clone with altered susceptibility to antifolate drugs. *Int. J. Parasitol.* **41**, 165–171 (2011).

25. Briolant, S. *et al.* The F423Y mutation in the *pfmdr2* gene and mutations N51I, C59R, and S108N in the *pfdhfr* gene are independently associated with pyrimethamine resistance in *Plasmodium falciparum* isolates. *Antimicrob. Agents Chemother.* **56**, 2750–2752 (2012).

26. Rosenberg, E. *et al. pfmdr2* confers heavy metal resistance to *Plasmodium falciparum. J. Biol. Chem.* **281**, 27039–27045 (2006).

27. Millet, J. *et al.* Polymorphism in *Plasmodium falciparum* drug transporter proteins and reversal of *in vitro* chloroquine resistance by a 9,10-dihydroethanoanthracene derivative. *Antimicrob. Agents Chemother.* **48**, 4869–4872 (2004).

28. Wongsrichanalai, C., Pickard, A.L., Wernsdorfer, W.H. & Meshnick, S.R. Epidemiology of drug-resistant malaria. *Lancet Infect. Dis.* **2**, 209–218 (2002).

29. Payne, D. Did medicated salt hasten the spread of chloroquine resistance in *Plasmodium falciparum? Parasitol.Today* **4**, 112–115 (1988).

30. Tabernero, P., Fernandez, F.M., Green, M., Guerin, P.J. & Newton, P.N. Mind the gaps—the epidemiology of poor-quality anti-malarials in the malarious world—analysis of the WorldWide Antimalarial Resistance Network database. *Malar. J.* **13**, 139 (2014).

31. Rathod, P.K., McErlean, T. & Lee, P.C. Variations in frequencies of drug resistance in *Plasmodium falciparum. Proc. Natl. Acad. Sci. USA* **94**, 9389–9393 (1997).

32. Beez, D., Sanchez, C.P., Stein, W.D. & Lanzer, M. Genetic predisposition favors the acquisition of stable artemisinin resistance in malaria parasites. *Antimicrob. Agents Chemother.* **55**, 50–55 (2011).

33. Sen, S. & Ferdig, M. QTL analysis for discovery of genes involved in drug responses. *Curr. Drug Targets Infect. Disord.* **4**, 53–63 (2004).

34. Gething, P.W. *et al.* A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar. J.* **10**, 378 (2011).

ONLINE METHODS

Ethics statement. All samples in this study were derived from blood samples obtained from patients with *P. falciparum* malaria, collected with informed consent from the patient or a parent or guardian. At each location, sample collection was approved by the appropriate local ethics committee: Ethical Committee, Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam; Ethics Committee for Biomedical Research of the Ministry of Health, Institute of Malariology-Parasitology-Entomology, Ho Chi Minh City, Vietnam; National Ethics Committee for Health Research, Ministry of Health, Phnom Penh, Cambodia; Ethics Committee, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand; Tak Province Community Ethics Advisory Board (T-CAB), Tak, Thailand; Government of the Republic of the Union of Myanmar, Ministry of Health, Department of Medical Research (lower Myanmar); National Ethics Committee for Health Research, Ministry of Health, Lao Peoples' Democratic Republic; National Research Ethics Committee, Bangladesh Medical Research Council; Comité d'Ethique, Ecole de Santé Publique, Université de Kinshasa, Ministère de l'Enseignement Superieur, Universitaire et Recherche Scientifique, Democratic Republic of the Congo; Ethical Review Committee, University of Ilorin Teaching Hospital, Ilorin, Nigeria; Navrongo Health Research Centre Institutional Review Board, Navrongo, Ghana; Institutional Review Board, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA; Ethics Review Committee, World Health Organization, Geneva, Switzerland; and Oxford Tropical Research Ethics Committee (OxTREC), Oxford, UK.

Phenotype estimation. Full details of clinical studies that contributed phenotypes for the GWAS analysis, including treatment regimens and clearance rate estimation, can be obtained from the trial registrations at ClinicalTrials.gov: NCT01350856 for the TRAC study and NCT00341003 and NCT01240603 for the US National Institutes of Health study.Clinical study reports also detail in depth the methods used in these studies[6,9].

During treatment, parasite densities were estimated by counting parasitized erythrocytes in blood smears from peripheral blood samples taken at 0, 4, 6, 8 and 12 h after patient admission and then every 6 h until two consecutive counts were negative. $PCt_{1/2}$ estimates were computed from these parasite counts, by fitting a statistical model[35] using the Parasite Clearance Estimator developed by WWARN.

Sample preparation, sequencing and genotyping. DNA was extracted directly from blood samples taken from patients at admission, after leukocyte depletion to minimize contamination from human DNA. Leukocyte depletion was achieved by CF11 filtration for most samples[10] or alternatively by Lymphoprep density gradient centrifugation (Axis-Shield) followed by Plasmodipur filtration (Euro-Diagnostica)[36] or by Plasmodipur filtration alone. Genomic DNA was extracted using the QIAamp DNA Blood Midi or Maxi kit (Qiagen), and the quantities of human and *Plasmodium* DNA were determined by fluorescence analysis using a Qubit instrument (Invitrogen) and multispecies quantitative PCR(qPCR) using the Roche LightCycler 480 II system, as described previously[11]. Samples with >50 ng of DNA and <80% human DNA contamination were selected for sequencing on the Illumina HiSeq platform following the manufacturer's standard protocols[37]. Paired-end sequencing reads of 200–300bp in length were obtained, generating approximately 1Gb of read data per sample.

Polymorphism discovery, quality control and sample genotyping followed a process described elsewhere[11]. Short sequence reads from 3,281 *P. falciparum* samples included in the MalariaGEN *Plasmodium falciparum* Community Project were aligned against the *P. falciparum* 3D7 reference sequence V3 using the bwa program[38] as previously described[11], to identify an initial global set of 3,373,632 potential SNPs. This list was then used to guide stringent realignment using the SNP-o-matic algorithm[39], to reduce misalignment errors. Stringent alignments were then examined by a series of quality filters, with the aim of removing alignment artifacts and their sources. In particular, the following were removed: (i) noncoding SNPs; (ii) SNPs where polymorphisms had extremely low support (<10 reads in 1 sample); (iii) SNPs with more than 2 alleles, with the exception of loci known to be important for drug resistance, which were manually verified to not have artifacts; (iv) SNPs where coverage across samples was lower than the 25th percentile or higher than the 95th percentile of

coverage in coding SNPs (these thresholds were determined from an analysis of artifact incidence); (v) SNPs located in regions of relatively low uniqueness[11]; (vi) SNPs where heterozygosity levels were found to be inconsistent with the heterozygosity distribution at the SNP's allele frequency; and (vii) SNPs where the genotype could not be established in at least 70% of samples. These analyses produced a final list of 681,587 high-quality SNPs in the 14 chromosomes of the nuclear genome, whose genotypes were used for analysis in this study.

All samples were genotyped at each high-quality SNP by a single allele, on the basis of the number of reads observed for the two alleles at that position in the sample. At positions with fewer than five reads, the genotype was set to undetermined (no call was made). At all other positions, the sample was determined to be heterozygous if both alleles were each observed in more than two reads; otherwise, the sample was called as homozygous for the allele observed in the majority of reads. For the purposes of estimating allele frequencies and genetic distances, a within-sample allele frequency (f_w) was also assigned to each valid call. For heterozygous calls, f_w was estimated as the ratio of the non-reference read count to the reference read count; homozygous calls were assigned $f_w = 0$ when called with the reference allele and $f_w = 1$ when called with the non-reference allele.

For specific analyses that required no genotype missingness in our data set, we produced a set of genotypes where missing calls (with coverage <5 reads) were assigned a genotype by simple imputation. First, we considered missing calls where the two flanking positions (on each side) had valid genotypes, imputing with the allele that most frequently appeared at the same position between the same flanking alleles in the full sample set. Finally, remaining samples with missing genotypes were assigned with the most common allele at that position in their population.

Genotype-phenotype association. Genotype-phenotype association analysis (GWAS) and correction for population structure was performed using a linear mixed model, implemented in FaST-LMM v2.06. We tested 17,395 SNPs with MAF > 0.01 where genotype was encoded as the number of non-reference alleles (0 or 1). At each SNP, heterozygous calls were excluded from the test, to minimize the confounding effect of mixed infections. $PCt_{1/2}$, estimated as described above, was used as the continuous dependent variable. A relationship matrix was calculated using a subset of 11,785 unlinked SNPs with MAF > 0.01, extracted using PLINK v1.07 (options: –indep-pairwise 100 10 0.3 –maf 0.01). In estimating the relationship matrix, we found that the exclusion of proximal SNPs (either within 10 kb or 100 kb of the tested variant) had no significant effect on the results (data not shown). Given the number of independent SNPs used, we applied Bonferroni correction to define for all GWAS analyses a significance threshold of $P \le 1 \times 10^{-7}$. In addition, we defined a 'suggestive' threshold at $P \le 1 \times 10^{-5}$ to help define high-ranking loci.

In a later analysis aimed at disentangling the residual effect of high-ranking loci after discounting the effect of *kelch13* mutations, we performed a conditional analysis, including the *kelch13* allele genotype as a fixed effect, as implemented in FaST-LMM.

Population genetics analysis. For a given population P, we estimated the non-reference allele frequency (NRAF) at a given SNP as the mean of the within-sample allele frequency (f_w) for all samples in P that had a valid genotype at that SNP. The MAF was computed as min(NRAF, (1 – NRAF)).

As input to population structure analyses, we computed an $N \times N$ pairwise distance matrix, where N was the number of samples. Each cell of the matrix contained an estimate of the genetic distance between the relevant pair of samples, obtained by summing the pairwise distance at each SNP, estimated from within-sample allele frequency (f_w). When comparing a pair of samples s_A and s_B at a single SNP i where a genotype could be called in each sample, with within-sample allele frequencies f_A and f_B, respectively, the distance d_{AB} was estimated as $d_{AB} = f_A(1 - f_B) + f_B(1 - f_A)$. The genome-wide distance D_{AB} between the two samples was then calculated as:

$$D_{AB} = \frac{\alpha}{n_{AB}} \sum_i w_i d_{AB}$$

where n_{AB} is the number of SNPs where both samples could be genotyped, w_i is a linkage disequilibrium (LD) weighting factor and α is a scaling constant,

equal to 70% of the number of coding positions in the genome (included because our genotyping covers approximately 70% of the coding genome). The exact value of α did not influence the analyses conducted in this study. The LD weighting factor, which corrects for the cumulative contribution of physically linked polymorphisms, was computed at each SNP i with MAF $\geq$ 0.1 in our sample set by considering a window of m SNPs ($j = 0, 1, \ldots, m$) centered at i. For each value of j, we computed the squared correlation coefficient r_{ij}^2 between SNPs i and j. Ignoring positions j where $r_{ij}^2 < 0.1$, the weighting factor w_i was computed by:

$$w_i = \frac{1}{1 + \sum_j r_{ij}^2}$$

Principal-coordinate analysis (PCoA) of pairwise distance matrices was performed using the classical multidimensional scaling (MDS) method[40]. PCoA is a computationally efficient variant of principal-component analysis (PCA) in which a pairwise distance matrix is used as input, rather than a table of genotypes. For each PCoA of a subset of N samples, we used an $N \times N$ pairwise distance matrix. The matrix was supplied as input to the MDS algorithm, using the cmdscale function in the R language stats package. The same pairwise distance matrix was also used to produce a neighbor-joining tree[41] using the nj implementation in the R ape package.

To estimate the F_{ST} value between two populations at a given SNP, we used $F_{ST} = 1 - (\hat{\pi}_s / \hat{\pi}_t)$, where $\hat{\pi}_s$ is the expected average probability that two samples chosen at random from the same population carry a different allele at the SNP and $\hat{\pi}_t$ is the expected average probability that two samples chosen at random from the joint population carry different alleles. Estimates for F_{ST} were obtained by using the NRAF values for the two populations (p_1 and p_2) to compute:

$$\hat{\pi}_s = \frac{1}{2}\big(2 p_1 (1 - p_1) + 2 p_2 (1 - p_2)\big)$$

and

$$\hat{\pi}_t = 2 \times \frac{(p_1 + p_2)}{2} \times \left(1 - \frac{(p_1 + p_2)}{2}\right)$$

Heteroallelic association analysis. We considered only samples from the two core populations (namely, KH-C and VN-C) and the seven artemisinin-resistant founder populations (WKH-F01, WKH-F02, WKH-F03, WKH-F04, NKH-F02, VN-F01 and VN-F04). Each founder population was compared to the respective core population (KH-C for WKH-F01, WKH-F02, WKH-F03, WKH-F04 and NKH-F02; VN-C for VN-F01 and VN-F04) to calculate the pairwise F_{ST} of all SNPs. Only SNPs with $F_{ST} \geq 0.3$ in at least one comparison were then considered (8,699 of 681,546). For each gene i containing at least one SNP meeting the above criterion ($n = 3,482$), we calculated for each pairwise comparison j ($1 \leq j \leq 7$) a score S_{ij} equal to the maximum F_{ST} across all qualifying SNPs. The score S_i for the gene was then calculated as the arithmetic mean of S_{ij} across the seven comparisons.

***kelch13* allele genotyping.** In analyses that required samples to be assigned a *kelch13* genotype, this was derived from the read counts at nonsynonymous SNPs in *kelch13*, using a procedure aimed at minimizing missing calls. At each position, the sample was assigned the reference allele if supported by ≥ 1 read for the reference allele and the alternative allele if supported by ≥ 2 reads for the alternative allele (≥ 3 reads where coverage exceeded 50 reads). Positions with no assigned allele were classified as missing, and those with both alleles were defined as heterozygous. Samples with ≥ 1 missing position were labeled with a missing genotype; samples with ≥ 1 heterozygous position were labeled as heterozygous; samples in which a single position carried exclusively the alternative allele were classified as single mutant and labeled with the mutation name; and the remainder of the samples were labeled as wildtype. No multiple mutants were identified in this data set. After the initial assessment of mutation phenotype (**Fig. 2**), we repeated genotyping using the same procedure but only considering nonsynonymous SNPs in *kelch13* encoding the resistance domains (the BTB/POZ and propeller domains); samples carrying mutations only mapping outside these domains were labeled as wild type.

Identification of populations and classification of samples. To identify core and founder populations, we used the following multistep method, aimed at a conservative classification of samples, applied separately to different geographical regions (**Supplementary Figs. 10–13** and **Supplementary Note**). To minimize ascertainment biases, we only used samples that were included in the SNP discovery phase. We applied the ADMIXTURE V1.23 program to estimate ancestry proportions for the selected samples, applying a model-based approach[42] that used majority-allele, imputed genotypes as input. All SNPs with extremely low MAF (MAF ≤ 0.01) were discarded owing to their low informative value in the inference process. Because the ADMIXTURE model requires low LD between SNPs, we also excluded SNPs belonging to highly linked pairs, selected by the PLINK toolset by scanning each chromosome in turn with a sliding window of 100 SNPs in size and removing any SNP with a correlation coefficient of ≥ 0.02 with any other SNP within the window.

For each run of ADMIXTURE, a hypothetical number K of ancestral populations was chosen, and each sample was assigned a fraction for each ancestral population. We ran the algorithm for multiple values of $K \geq 2$. To avoid fitting to local minima, for each K we ran the algorithm 50 times with different random seeds and assessed the distributions of cross-validation errors and loglikelihood for all runs. For our data sets, the cross-validation error distributions presented large plateaus where solutions showed only marginal improvement as K increased (**Supplementary Fig. 14**), accompanied by an increase in variance, making it problematic to choose an optimum value of K. Therefore, we followed a conservative iterative process for robustly assigning samples to populations. First, we used a published method for identifying a value of K that showed the uppermost level of structure[43]. Starting with this value, we gradually increased K to capture structure at a finer resolution. For each K, we chose the solution with the lowest cross-validation error. On the basis of the proportions estimated in this solution, we assigned each sample to one of K groups corresponding to the putative ancestral populations. A sample was assigned to a group if the proportion estimated for the corresponding ancestor was >0.5 and at least four times higher than the second highest proportion. Samples not meeting these criteria were assigned to an 'unclassified' group. We then used group labels to identify clusters of samples that consistently grouped together at different values of K. One classification mismatch at most (where the group assignment was different for one value of K) was allowed for cluster assignment. The value of K was incremented until newly identified clusters were deemed to be too small ($n < 5$) or unstable (where the cluster separated as an independent group at one value of K and then merged with a different group at a higher value of K). Samples assigned to the unclassified group for more than a single value of K were not assigned to clusters.

Core populations were identified by genetic similarity to populations previously found to be sensitive and representative of wild-type genotypes[16]. Putative founder clusters were characterized by a pairwise count of highly differentiated SNPs ($F_{ST} \geq 0.5$) with respect to a reference core population, expected to be much higher in a founder population than in a core population. We also performed PCoA to visualize the clustering of the putative founders and their separation along at least one component from the core populations and from each other. Populations with insufficient support for founder effects or with low numbers ($n < 5$) were discarded (samples were assigned to the unclassified group).

Demography of *kelch13* mutations. We based our analysis on the length of the longest haplotypes shared by a pair of samples. In both strand directions, starting from the mutation of interest, we found the closest positions where the two sequences differed (breakpoints). The distance, in base pairs, between two breakpoints was the LCHL for the pair. To account for heteroallelism, we collapsed all *kelch13* mutations into a single one and used position 580 as the notional locus of the mutation of interest. To minimize the influence of possible artifacts, we defined a breakpoint only if the mismatching allele had a frequency of >5% among samples carrying the same *kelch13* mutation or among samples wildtype for *kelch13*. Variants arising from the same recent evolutionary event will be embedded within identical haplotypes, whereas mutations originating separately will share shorter haplotypes; LCHL is also expected to decrease with age, owing to the effects of recombination. Hence, clustering by LCHL is expected to group together samples with common recent

demographic history. Accordingly, we constructed a pairwise matrix of the inverse of LCHL, from which a tree was constructed using a standard hierarchical clustering method (hclust) implemented in the R stats package using the Ward's minimum variance criterion.

35. Flegg, J.A., Guerin, P.J., White, N.J. & Stepniewska, K. Standardizing the measurement of parasite clearance in falciparum malaria: the parasite clearance estimator. *Malar. J.* **10**, 339 (2011).
36. Auburn, S. *et al.* An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS ONE* **6**, e22213 (2011).
37. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Manske, H.M. & Kwiatkowski, D.P. SNP-o-matic. *Bioinformatics* **25**, 2434–2435 (2009).
40. Gower, J.C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–328 (1966).
41. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
42. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
43. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).