Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing

Malaria elimination strategies require surveillance of the parasite population for genetic changes that demand a public health response, such as new forms of drug resistance. Here we describe methods for the large-scale analysis of genetic variation in Plasmodium falciparum by deep sequencing of parasite DNA obtained from the blood of patients with malaria, either directly or after short-term culture. Analysis of 86,158 exonic single nucleotide polymorphisms that passed genotyping quality control in 227 samples from Africa, Asia and Oceania provides genome-wide estimates of allele frequency distribution, population structure and linkage disequilibrium. By comparing the genetic diversity of individual infections with that of the local parasite population, we derive a metric of within-host diversity that is related to the level of inbreeding in the population. An open-access web application has been established for the exploration of regional differences in allele frequency and of highly differentiated loci in the P. falciparum genome.

The genetic diversity and evolutionary plasticity of P. falciparum are major obstacles for malaria elimination. New forms of resistance against antimalarial drugs are continually emerging, and new forms of antigenic variation are a critical point of vulnerability for future malaria vaccines. Effective tools are needed to detect evolutionary changes in the parasite population and to monitor the spread of genetic variants that affect malaria control.

Here we describe the use of deep sequencing to analyse P. falciparum diversity, using blood samples from patients with malaria. The P. falciparum genome has several unusual features that greatly complicate sequence analysis, such as extremely AT-rich, large tracts of non-unique sequence and several large families of intensely polymorphic genes. Our aim was therefore not to determine the entire genome sequence of individual field samples—which would be prohibitively expensive with current technologies—but to define an initial set of single nucleotide polymorphisms (SNPs) distributed across the P. falciparum genome, whose genotype can be ascertained with confidence in parasitized blood samples by deep sequencing.

An additional complication in the analysis of P. falciparum genome variation is that the billions of haploid parasites that infect a single individual can be a complex mixture of genetic types. Previous studies have largely focused on laboratory-adapted parasite clones, but the within-host diversity of natural infections is of fundamental biological interest. Parasites in the blood replicate asexually, but when they are taken up in the blood meal of an Anopheles mosquito they undergo sexual mating. If the parasites in the blood are of diverse genetic types, this process of sexual mating can generate novel recombinant forms. Deep sequencing provides new ways of investigating within-host diversity and the role of sexual recombination in parasite evolution.

P. falciparum DNA was obtained from blood samples collected from 290 patients with malaria at clinics in Burkina Faso, Cambodia, Kenya, Mali, Papua New Guinea and Thailand (Supplementary Table 1). For 149 samples we used the conventional method of growing the parasites in short-term blood culture before extracting the P. falciparum DNA. For 141 samples we used a new method by which P. falciparum DNA is extracted directly from venous blood samples after the removal of leukocytes. We refer to these as cultured and direct samples, respectively.

Paired-end sequence reads were generated (median 7 × 10⁸ base pairs per sample) by using the Illumina Genome Analyzer platform. Sequence analysis was divided into stages of SNP discovery, quality control filtering, genotyping and validation (see Supplementary Methods and Supplementary Fig. 1). After alignment to the 3D7 reference genome, non-coding regions had a much lower read depth than coding regions (Supplementary Fig. 2); this can be ascribed to their high AT content (non-coding 87% AT, coding 70% AT). Read depth was also low in the highly polymorphic var, rifin and stever coding regions (Supplementary Fig. 3). For the purposes of this study, to decrease genotyping errors due to low coverage or copy number variation we excluded all non-coding regions, as well as coding regions at the extremes of the read depth distribution. After these exclusions we were left with 70% of all exonic positions across the genome, with
more than 50% of exonic positions for 71% of genes, and more than 70% for 54% of genes (Supplementary Table 2).

Within-host diversity complicates the process of excluding sequencing and alignment errors that are manifested as false heterozygous genotypes. Two approaches were identified to address this problem (see Supplementary Methods). We scored each position in the reference genome for its degree of uniqueness, and this was found to be a strong predictor of false heterozygous genotypes. We also observed a relationship between the population allele frequency of a SNP and its average level of within-sample heterozygosity, analogous to the Hardy–Weinberg relationship in diploid organisms. This enabled us to exclude SNPs that had excessive levels of within-sample heterozygosity relative to their population frequency.

After applying the above filters, and excluding SNPs and samples with high levels of missing data, we obtained a final data set of 86,158 SNPs genotyped in 227 samples (120 direct and 107 cultured) in which a median of 98% samples had valid genotyping data for each SNP, and a median of 98% SNPs had valid genotyping data for each sample (Supplementary Fig. 4). This set of 86,158 SNPs (here referred to as the 86k SNP set) represents 10% of the SNPs discovered at the initial stage of sequence alignment. Comparison with the PlasmoDB 5.5 database indicates that 77,283 (89%) of these SNPs are novel, but it should be noted that previous genome-wide SNP discovery efforts have largely been based on low-coverage capillary sequencing, and the overall error rate is unknown.

The accuracy of genotype calls in the 86k SNP set was evaluated by five independent approaches (see Supplementary Methods). We examined the evidence for 275 putative novel SNPs using independent data from PCR-based capillary sequencing and Sequenom primer-extension mass spectrometry: the existence of the novel allele was confirmed for 270 of the 275 loci. The genotype concordance rate with Sequenom was 99.9% and with capillary sequencing it was 98.6%, excluding heterozygotes (Supplementary Tables 3 and 4). In the case of heterozygous genotypes, deep sequencing gives the allelic ratio, whereas most other P. falciparum SNP typing methods give the majority allele or return a missing genotype. The observation of heterozygosity by deep sequencing was correlated with Sequenom’s failing to call a majority allele, but when Sequenom made a majority allele call it agreed with deep sequencing data in 94.8% of cases (Supplementary Fig. 5). Capillary sequencing data do not allow allelic ratios to be quantified precisely, but visual inspection of capillary sequence traces was consistent with heterozygous genotype calls in the deep sequencing data (Supplementary Fig. 6). In a separate study to be reported elsewhere, we sequenced 90 laboratory-adapted parasite clones derived from three genetic crosses of P. falciparum and determined that the rate of Mendelian errors in the 86k SNP set was 0.05%.

Population genetic analyses were conducted with the 86k SNP set typed in 227 samples as described above. The allele frequency spectrum was dominated by low-frequency variants (Fig. 1 and Supplementary Fig. 7) even when synonymous sites alone were considered, which is consistent with recent population expansion (Supplementary Table 5). Samples from Africa had a greater number of low-frequency variants than samples from Southeast Asia or Papua New Guinea with or without correction for sample size. Multiple lines of evidence indicate that P. falciparum originated in Africa, and loss of low-frequency variation might have occurred as a result of population bottlenecks during migration out of Africa, as in human populations.

The most likely ancestral state of each SNP was determined from the P. reichenowi genome sequence but is difficult to estimate with confidence, because P. reichenowi might have diverged from P. falciparum relatively recently, and its genome sequence has been determined for only one individual (refs 6, 12 and T. D. Otto, unpublished observations). There seem to be more SNPs with low-frequency derived (non-ancestral) alleles in Africa than in Southeast Asia or Papua New Guinea (Supplementary Figs 8 and 9). Focusing on SNPs that are private to one continent, those with high derived allele frequency show a considerable excess of non-synonymous substitutions, suggesting that these are largely the result of directional selection (Fig. 1b and Supplementary Fig. 10).

Many SNPs (64%) were observed in only one continent, but most were low-frequency variants and larger sample sizes are needed to determine how many of these are truly private. Corrected for sample size, the number of private SNPs was greatest in East Africa and least in Southeast Asia, both of which comprised cultured samples (Supplementary Fig. 11). Intermediate numbers were observed in West Africa and Papua New Guinea, both of which comprised direct samples. Thus the effect of culturing on SNP ascertainment seems to be relatively small in comparison with the effect of geographical location.

The global population structure of P. falciparum shows a clear division by continent (Fig. 2a). Mean fixation index (Fst) values between continents ranged from 0.19 to 0.28 (Supplementary Table 6). Population structure within continents is evident from Fst values, principal-components analysis (Supplementary Fig. 12) and a neighbour-joining tree (Fig. 2b). All of these methods show greater degree of population structure in Southeast Asia than in West Africa; that is, samples from Cambodia and Thailand form separate clusters, whereas samples from Mali and Burkina Faso are intermixed. These data are consistent with previous evidence that parasite population structure tends to be increased in regions of low or patchy malaria transmission.

To understand the hierarchical population structure of P. falciparum, methods are needed to quantify the genetic diversity of individual infections relative to the genetic diversity of the parasite population as a whole. With deep sequencing data, we can estimate
levels of heterozygosity both within an individual sample ($H_w$) and within the local parasite population ($H_s$). For a biallelic SNP, we define $H_w$ as $2p_wq_w$, where $p_w$ and $q_w$ denote the proportions of the two alleles in the sequence reads of an individual sample, and $H_s$ as $2p_sq_s$ where $p_s$ and $q_s$ denote the corresponding population allele frequencies at that geographical location. We observe a strong linear relationship between $H_w$ and $H_s$ when data for all 86k SNPs are aggregated for an individual sample (Fig. 3a and Supplementary Fig. 13). More specifically, each sample shows a linear relationship between $H_w$ and $H_s$ but the gradient of the line varies considerably between samples. This gradient is essentially a genome-wide estimate of $H_w/H_s$ for the sample in question. Thus for each sample we can derive the metric $F_{ws}$, where

$$F_{ws} = 1 - H_w/H_s$$

This is closely related to Wright’s inbreeding coefficient $F_{is}$ which can be formulated as

$$F_{is} = 1 - H_i/H_s$$

where $H_i$ is the heterozygosity of the individual and $H_s$ is that of the local population\(^\text{14}\). Estimation of $F_{is}$ is of practical relevance for malaria

---

**Figure 2 | Representations of a pairwise distance matrix between the 227 samples analysed.** a, Principal-components analysis. b, Unrooted neighbour-joining tree. Leaf branches are coloured (as in a) according to the country of origin of the sample.
control, because high rates of inbreeding are thought to favour the emergence of multigenic drug resistance. F_w is conventionally measured at the oocyst stage of infection—that is, after the parasites have undergone sexual mating within the mosquito and before they develop into separate haploid forms—but this is technically demanding and difficult to implement on a large scale. Because parasites undergo sexual mating shortly after the mosquito has ingested blood from an infected person, the level of within-host diversity determines the potential for inbreeding or outcrossing in the next generation. Thus F_w values observed in blood samples provide a proxy indicator of inbreeding rates in the population. The precise relationship to inbreeding rates quantified in oocysts merits further investigation. We report elsewhere a study of how F_w relates to standard methods of estimating multiplicity of infection.

We observe marked differences in F_w between locations (Fig. 3b). High levels of F_w (0.95 or more) were much more common in Papua New Guinea (89% of samples) than in West Africa (38%), with intermediate rates in Southeast Asia (67%) and East Africa (63%). Culturing might affect F_w estimation, but the samples from Papua New Guinea and West Africa were not cultured. In general, high levels of inbreeding tend to be associated with low transmission intensity, and these data are therefore somewhat surprising because the entomological inoculation rate has been estimated to lie in the range 45–293 in Madang in Papua New Guinea, where the Papua New Guinea samples were collected, in contrast with 140–389 in Burkina Faso, about 6 in rural areas of Cambodia and about 1 on the Thailand–Burma border. Although the entomological inoculation rate can be highly variable within a locality and these estimates are indicative, it seems unlikely that the high levels of F_w in Papua New Guinea are primarily due to low transmission intensity. An alternative explanation is that, in this geographical region, people tend to live in small isolated communities, which might reduce the likelihood of infection with parasites of different genetic types. The small size of the Papua New Guinea sample provides limited information about local parasite population structure (Supplementary Fig. 14), but previous studies indicate that this is very high in some villages within this area of Papua New Guinea.

These data allow linkage disequilibrium in the P. falciparum genome to be estimated with greater precision than has previously been possible. In particular, we can begin to distinguish linkage disequilibrium due to haplotype structure, which decays with distance in the genome, from linkage disequilibrium due to population structure, which is independent of distance in the genome (see Supplementary Methods, Supplementary Tables 7 and 8 and Supplementary Figs 15–17). Averaged across the genome, after correcting for population structure and other confounders, we find that r^2 decays to less than 0.1 within 1 kilobase in all populations studied here, whereas D' decays to less than 0.1 within about 1 kb in West Africa and East Africa, and within 50 kb in Southeast Asia and Papua New Guinea (Supplementary Fig. 18). These findings imply that high levels of haplotypic diversity exist at all of these locations, despite low transmission intensity and high rates of inbreeding at some locations. This might be partly due to the high rate of meiotic recombination in P. falciparum, previously estimated to be about 17 kb per centimorgan. It is also possible that much of the haplotypic diversity seen in contemporary P. falciparum populations has ancient origins, and arose in Africa before P. falciparum was spread around the world by human migration. This would be analogous to the situation that is seen in human populations, in which migration out of Africa was associated with a series of population bottlenecks, which have led to a reduction in haplotypic diversity in descendant populations around the world. The higher levels of linkage disequilibrium observed in Southeast Asia and Papua New Guinea than in West Africa and East Africa are consistent with both of these possibilities.

A web application is provided for browsing, querying and downloading information about all of the SNPs genotyped in this study and their allele frequencies in different geographical regions (http://www.malarialogen.net/resource/10). It can be used, for example, to view regional patterns of variation in known antimalarial drug resistance genes: from these data it is immediately apparent that the pfcrf K76T allele has markedly different haplotypic backgrounds in Southeast Asia and in Papua New Guinea, consistent with previous evidence that chloroquine resistance has evolved independently in multiple locations (Supplementary Table 9). It can also be used to search for genes that are highly differentiated between geographical regions (Supplementary Tables 10 and 11). For example, two genes that affect the fertility of gametocytes, Pf6230 and Pf47, are among the most highly differentiated loci in this data set. Two SNPs in Pf6230 codon 1566 result in three amino-acid variants: N (widespread), T (private to Southeast Asia, frequency 0.87) and K (private to Africa, frequency 0.79). Codon variant T236I of Pf47 has a fixed difference between Africa and other populations. These data lend weight to previous reports of extreme differentiation in Pf47 and the related gene Pfs48/45 (ref. 26), which is suggested to be due to evolutionary selection of gene-specific recognition and compatibility. Another example is codon variant F368S of the putative transporter gene PFA0245w (ref. 27), which has a fixed difference between Papua New Guinea and other populations, raising the question of whether this has a function in drug resistance; it is also noteworthy that the Plasmodium berghei orthologue of this gene is critical for sexual development of the parasite.

These data are the first stage in the development of methods for population-based genome sequencing of P. falciparum. Work is ongoing to increase the number of SNPs that can be reliably genotyped, and to develop accurate methods for typing indels, copy number polymorphisms and large structural variations. Future studies will benefit from new methods to reduce the effects of AT bias on sequencing library preparation, and the increasing length and accuracy of sequencing...
reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY
Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture samples after leukocyte depletion to minimize contamination with human Blood samples from malaria patients were collected with informed consent after antimalarial drug resistance and to monitor evolutionary changes in logical strain selection.

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.