Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi*

Samuel Assefa[a], Caeul Lim[b], Mark D. Preston[a], Craig W. Duffy[a], Mridul B. Nair[c], Sabir A. Adroub[c], Khamisah A. Kadir[d], Jonathan M. Goldberg[b], Daniel E. Neafsey[e], Paul Divis[a,d], Taane G. Clark[a], Manoj T. Duraisingh[b,d], David J. Conway[a,d,1], Arnab Pain[c,d,f,1], and Balbir Singh[d,1]

[a]Pathogen Molecular Biology Department, London School of Hygiene and Tropical Medicine, London, WC1E 7HT United Kingdom; [b]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA 02115; [c]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Jeddah, Kingdom of Saudi Arabia; [d]Malaria Research Centre, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia; [e]Broad Institute of MIT and Harvard, Cambridge, MA 02142; and [f]Center for Zoonosis Control, Global Institution for Collaborative Research and Education, Hokkaido University, N20 W10 Kita-ku, Sapporo, Japan

Edited by Xin-zhuan Su, National Institutes of Health, Rockville, MD, and accepted by the Editorial Board September 3, 2015 (received for review May 21, 2015)

Malaria cases caused by the zoonotic parasite *Plasmodium knowlesi* are being increasingly reported throughout Southeast Asia and in travelers returning from the region. To test for evidence of signatures of selection or unusual population structure in this parasite, we surveyed genome sequence diversity in 48 clinical isolates recently sampled from Malaysian Borneo and in five lines maintained in laboratory rhesus macaques after isolation in the 1960s from Peninsular Malaysia and the Philippines. Overall genomewide nucleotide diversity ($\pi = 6.03 \times 10^{-3}$) was much higher than has been seen in worldwide samples of either of the major endemic malaria parasite species *Plasmodium falciparum* and *Plasmodium vivax*. A remarkable substructure is revealed within *P. knowlesi*, consisting of two major sympatric clusters of the clinical isolates and a third cluster comprising the laboratory isolates. There was deep differentiation between the two clusters of clinical isolates [mean genomewide fixation index (F_{ST}) = 0.21, with 9,293 SNPs having fixed differences of F_{ST} = 1.0]. This differentiation showed marked heterogeneity across the genome, with mean F_{ST} values of different chromosomes ranging from 0.08 to 0.34 and with further significant variation across regions within several chromosomes. Analysis of the largest cluster (cluster 1, 38 isolates) indicated long-term population growth, with negatively skewed allele frequency distributions (genomewide average Tajima's D = −1.35). Against this background there was evidence of balancing selection on particular genes, including the circumsporozoite protein (*csp*) gene, which had the top Tajima's D value (1.57), and scans of haplotype homozygosity implicate several genomic regions as being under recent positive selection.

population genomics | *Plasmodium* diversity | reproductive isolation | zoonosis | adaptation

The zoonotic malaria parasite *Plasmodium knowlesi* is a significant cause of human malaria, with a wide spectrum of clinical outcomes including high parasitemia and death (1–4). Long known as a malaria parasite of long-tailed and pig-tailed macaques (5), the first large focus of human cases was described only in 2004 in the Kapit Division of Sarawak in Malaysian Borneo (6). Since then infections have been described from almost all countries in Southeast Asia (2, 7). Travelers to the region from Europe, North America, and Australasia also have recently acquired *P. knowlesi* malaria (7, 8). Until the application of molecular assays for specific detection, human *P. knowlesi* malaria was largely misdiagnosed as *Plasmodium malariae,* a morphologically similar but distantly related species (1, 6, 9, 10). Studies in the Kapit Division of Sarawak in Malaysian Borneo have indicated that *P. knowlesi* malaria is primarily a zoonosis with macaques as reservoir hosts (11) and that the forest-dwelling mosquito species *Anopheles latens* is the local vector for *P. knowlesi* (12). Other members of the *Anopheles leucosphyrus* group are vectors in different parts of Southeast Asia and may determine the geographical distribution of transmission (13, 14).

Although *P. knowlesi* malaria is regarded as an emerging infection, there clearly have been increased efforts in detection made since its existence as a significant zoonosis was discovered, and specific detection has been enhanced by the declining numbers of human cases caused by other malaria parasites in Southeast Asia (15). Aside from the first two human cases described several decades ago (5), there is direct evidence of human *P. knowlesi* infections from ∼20 y ago in Malaysian Borneo and Thailand obtained by retrospective molecular analysis of material from archived blood spots and slides (10, 16), and molecular population genetic evidence indicates the zoonosis has been in existence for a much longer time (11). The genetic diversity of *P. knowlesi* is high within humans as well as macaques, with sequence data on three loci [the circumsporozoite protein (*csp*) gene, 18S rRNA, and mtDNA genome] indicating extensive shared polymorphism and no fixed differences between *P. knowlesi* parasites from humans and monkeys sampled in the same area in Sarawak, Malaysian Borneo (6, 11). Analysis of samples from a smaller number of humans and monkeys in Thailand showed alleles of the *P. knowlesi* merozoite surface protein 1 (*msp1*) gene to be similarly diverse in both hosts (16), and there were shared

Significance

Genome sequence analysis reveals that the zoonotic malaria parasite *Plasmodium knowlesi* consists of three highly divergent subpopulations. Two, commonly seen in sympatric human clinical infections in Malaysian Borneo, were identified in a previous study as corresponding to parasites seen in long-tailed and pig-tailed macaque hosts, respectively. A third type has been detected in a few laboratory-maintained isolates originally derived in the 1960s elsewhere in Southeast Asia. Divergence between the subpopulations varies significantly across the genome but overall is at a level indicating different subspecies. Analysis of the diversity within the most common type in human infections shows strong signatures of natural selection, including balancing selection and directional selection, on loci distinct from those under selection in endemic human malaria parasites.

Author contributions: D.J.C., A.P., and B.S. designed research; S.A., C.L., M.B.N., S.A.A., K.A.K., D.E.N., M.T.D., D.J.C., A.P., and B.S. performed research; S.A., M.D.P., C.W.D., J.M.G., D.E.N., P.D., T.G.C., M.T.D., D.J.C., A.P., and B.S. analyzed data; and S.A., D.J.C., A.P., and B.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. X.S. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the European Nucleotide Archive (accession no. PRJEB10288) and NCBI database (accession no. SRP063014).

[1]To whom correspondence may be addressed. Email: david.conway@lshtm.ac.uk, arnab.pain@kaust.edu.sa, or bskhaira55@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1509534112/-/DCSupplemental.

GENETICS

polymorphisms of the *csp* gene in parasites from a few infections examined in humans and macaques in Singapore (17). Recent multilocus microsatellite analysis has indicated a deep population subdivision in *P. knowlesi* associated with long-tailed and pig-tailed macaques; both major types infect humans and occur sympatrically at most sites in Malaysia, but the two types show some additional geographical differentiation across sites (18).

Human populations have grown very rapidly in the Southeast Asian region and encroach on most of the wild macaque habitats, so it is vital to know if *P. knowlesi* parasites are adapting to human hosts or to anthropophilic mosquito vector species, either of which could cause human–mosquito–human transmission. Initial analysis of the *P. knowlesi* reference genome sequence (strain H) highlighted some unique features of the genome of this species (19), namely, schizont infected cell agglutination variant (*SICAvar*) and knowlesi interspersed repeat (*KIR*) variant antigen genes, which were widely dispersed instead of being predominantly localized in subtelomeric regions as seen in large gene families in *Plasmodium falciparum* and *Plasmodium vivax* (20). Here, we analyzed genomewide diversity in *P. knowlesi* and conducted scans for signatures of balancing and directional selection, revealing extremely high genetic diversity and significant structuring of this species into subpopulation clusters that appear to be reproductively isolated as well as loci that show evidence of recent strong selection.

Results

Sequencing and Nucleotide Diversity of Clinical and Laboratory *P. knowlesi* Isolates. Paired-end short-read sequencing produced high-quality data for 53 *P. knowlesi* isolates (48 clinical infections and five laboratory-maintained lines), with genomewide average mapping depth to the *P. knowlesi* H strain reference genome (19) having a minimum and median fold coverage of 36× and 120×, respectively (Table S1). Another three clinical isolates had average fold coverage below 30× and were not used for further analysis. All isolates contained sequences that potentially encode amino acid motif groups that may be involved in molecular mimicry of macaque CD99, as previously seen in the *P. knowlesi* H strain reference genome (19).

For analysis of polymorphism, we excluded subtelomeric and repeat regions (as defined in *Materials and Methods*) and members of large multigene families (*SICAvar* and *KIR*, $n = 373$ including fragments), allowing 92% of the genome to be analyzed (21.7 Mb of the 23.5-Mb reference genome), including 93% of all predicted protein-coding genes (4,860 genes). Overall, 975,642 biallelic high-quality SNPs were identified among the 53 isolates, indicating a very high level of nucleotide diversity [genomewide $\pi = 6.03 \times 10^{-3}$ (genic $\pi = 4.84 \times 10^{-3}$; noncoding $\pi = 7.50 \times 10^{-3}$)]. Almost half of the SNPs (423,160; 43.4% of the total) had a minor allele that occurred in only one isolate, and 21.8% ($n = 213{,}375$) had minor allele frequencies of above 10% in the overall sample of isolates.

All individual isolates were very distinct from one other (Fig. 1) except for two of the laboratory lines [labeled "H(AW)" and "Malayan"], which were virtually identical to each other and to the published H strain reference genome sequence (19). Only 251 SNPs were identified among these three sequences [numbers of pairwise differences: H reference vs. Malayan = 226; H reference vs. H(AW) = 221; Malayan vs. H(AW) = 57], 18 of which were in the coding regions of nine genes (Table S2) and two of which (PKH_112660 and PKH_133910) are members of the Apetala 2 (*AP2*) gene family of transcription factors. Interestingly, the SNP on one of the *AP2* genes (PKH_133910) led to the only stop codon mutation identified in the laboratory isolates, in line H(AW). It is notable that the MR4-H strain sequence [obtained from the Malaria Research and Reference Reagent Resource Center (MR4) repository as supposedly the reference H strain genome] was very different from the published H strain reference sequence. This difference, together with the observation that the line called "Malayan" was virtually identical to the H strain reference sequence, suggests historical mislabeling or mix up of laboratory lines. It is likely that most or all of the SNPs differentiating the three nearly identical genome sequences (including those in the *AP2* genes) were derived by mutations during laboratory passage in rhesus macaques or more recently in culture. For subsequent analysis of diversity between different sampled isolates, only one of these nearly identical H strain-like sequences (Malayan) was included.

Major Substructure Within the *P. knowlesi* Species. Analysis of pairwise nucleotide differences among all the isolates revealed three main clusters. The largest two represent subgroups of the clinical isolates (cluster 1, $n = 38$; cluster 2, $n = 10$), and a third comprises the laboratory isolates (cluster 3, $n = 5$ with $n = 4$ being distinct) (Fig. 1). Principal components analysis (PCA) confirmed the presence of the three groups, with the first two principal components

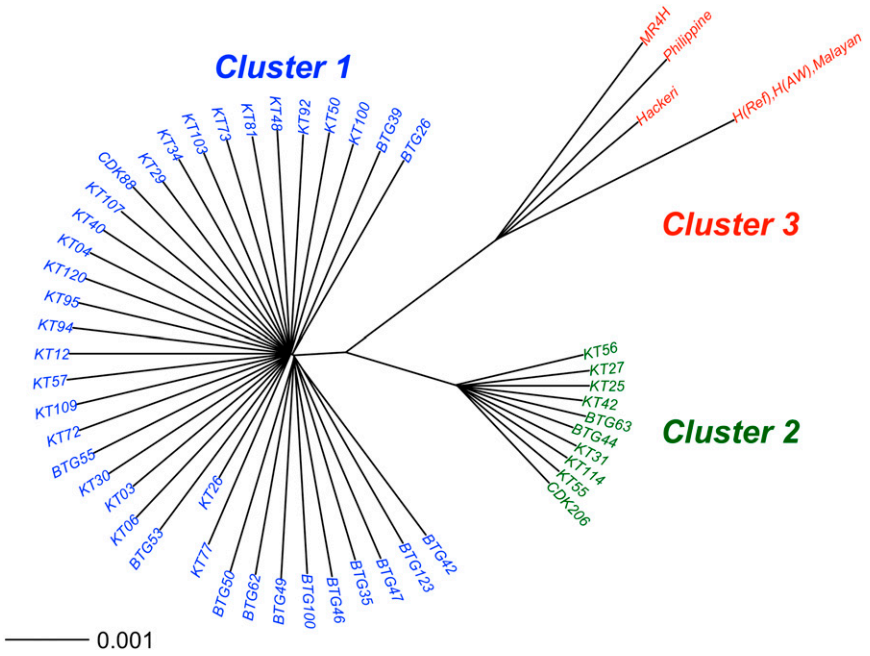

Fig. 1. Deep genomic population substructure in *P. knowlesi*. Neighbor-joining tree based on pairwise nucleotide diversity (π) between isolates using high-quality SNPs from 53 samples (48 clinical isolates from human patients and five laboratory samples maintained in rhesus macaques). This tree shows three major clusters representing two subgroups of clinical isolates (cluster 1, $n = 38$; cluster 2, $n = 10$) and a third cluster of laboratory isolates (cluster 3, $n = 5$) together with the reference genome sequence H(Ref). Clusters 1 and 2 occurred sympatrically at both of the sampling sites (in Kapit, 25 and 8 in each cluster respectively; in Betong, 13 and 2 respectively). Two of the cluster 3 laboratory isolates, labeled "H(AW)" and "Malayan," were nearly identical to each other and to the reference genome sequence. The isolate labeled here as "MR4H" was received from the MR4 reagent repository labeled as the "H" strain.

accounting for 26% of the variation (Fig. S1). The nucleotide divergence between cluster 3 and the others [for clusters 1 vs. 3, genetic distance (D_{XY}) = 8.4 × 10^{-3}; for cluster 2 vs. 3 D_{XY} = 8.1 × 10^{-3}) was slightly higher than between the clinical isolate clusters 1 and 2 (D_{XY} = 6.3 × 10^{-3}) (Fig. S2). The assignment of the clinical isolates into two distinct clusters shown here also was compared with an independent assignment recently performed using a STRUCTURE analysis of 10-locus microsatellite genotype data for 40 of the 48 isolates (18). All isolates had the same cluster assignment by both methods, except for one cluster 1 isolate, which had an intermediate cluster assignment in the microsatellite analysis (Fig. S3).

The average pairwise nucleotide diversity was higher among the cluster 1 clinical isolates (π = 5.28 × 10^{-3}) and among the cluster 3 laboratory isolates (π = 5.67 × 10^{-3}, n = 4 excluding nearly identical isolates) than among the cluster 2 clinical isolates (π = 3.15 × 10^{-3}). This result is not caused by genetic complexity within infections. Generally, there are low levels of complexity per infection in clinical *P. knowlesi* isolates, most containing one predominant haploid genotype each (Table S3). The estimates of within-infection genomic complexity (1 − F_{WS}) were similar for cluster 1 and cluster 2 isolates [mean within-infection fixation index (F_{WS}) values of 0.94 and 0.95 respectively; Mann–Whitney test; P = 0.40] and correlated highly with indices independently derived by counting the number of distinct alleles in a previously performed 10-locus microsatellite genotypic analysis (Pearson's r^2 = 0.76, P = 6.4 × 10^{-10}) (Table S3) (18).

Despite the differences in overall nucleotide diversity in the different clusters, the relative proportions of synonymous, nonsynonymous, intronic, and intergenic SNPs observed were similar in each (respectively 14%, 16%, 13%, and 57% for cluster 1; 14%, 17%, 13%, and 56% for cluster 2; and 13%, 18%, 14%, and 55% for cluster 3). A genomewide scan shows that lower nucleotide diversity among the cluster 2 isolates is observed in most chromosomes; exceptions with similar levels of diversity include chromosome 5 and regions of chromosomes 8, 10, and 11 (Fig. 2). In contrast, clusters 1 and 3 show similar variation in levels of diversity across the genome, so it is clear that cluster 2 differs in this particular respect.

Comparison of cluster 1 and 2 clinical isolates revealed a very high level of differentiation across the genome [genomewide average fixation index (F_{ST}) = 0.21] (Fig. 3). Moreover, a large number of SNP positions (n = 9,293) showed complete fixation of alternative alleles between these clusters (F_{ST} = 1.0), and these SNP positions with complete fixation occurred on all chromosomes (Fig. 3*B* and Fig. S4). However, there was marked heterogeneity in the level of differentiation between clusters 1 and 2 across the genome, with chromosomal mean F_{ST} scores ranging from 0.08 (chromosome 5) to 0.34 (chromosome 7) (Fig. 3 and Fig. S4). Analysis of windows of 500 consecutive SNPs revealed that chromosomes 5, 10, and 11 had the most windows below the genomewide average F_{ST} (Fisher's exact tests, P = 3.6 × 10^{-7}, 1.4 × 10^{-5}, and 2.5 × 10^{-5}, respectively), whereas chromosomes 7, 12, and 13 had the most windows above the genomewide average (3.7 × 10^{-16}, 4.0 × 10^{-6}, and 2.1 × 10^{-11}, respectively). Chromosomes 7, 12, and 13 also had high D_{XY} cluster 1 versus cluster 2 divergence values of 7.64 × 10^{-3}, 6.64 × 10^{-3}, and 7.08 × 10^{-3}, respectively (Fig. S2), indicating that the elevated F_{ST} values are not caused entirely by the lower diversity within cluster 2. There also was heterogeneity across different regions within chromosomes (Fig. 3 and Fig. S4), with window mean F_{ST} values ranging from <0.05 to >0.40. Regions with high F_{ST} values contained many SNPs that had completely fixed differences between the clusters (F_{ST} = 1.0) (Fig. 3*B* and Fig. S4). Application of Moran's I test of autocorrelation indicated that chromosomes 7, 8, 11, 12, and 13 had particularly significant nonrandom intrachromosomal clustering of F_{ST} (Fig. S4).

Scan for Genes Showing Signatures of Balancing Selection. Because of the overall deep population structuring and the small numbers of isolates belonging to clusters 2 and 3, scans for signatures of selection were focused on the main cluster 1 of clinical isolates (n = 38). A total of 759,409 biallelic SNPs were identified within cluster 1, 230,398 (30%) of which were found in coding regions (2,399 genes). A genomewide scan of allele frequency distributions conducted on genes that had a minimum of three SNPs (2,381 genes) showed a negatively skewed distribution, with a mean Tajima's D value of −1.35 (Fig. 4*A*). Only 16 of the 2,381 genes (0.67%) had a positive Tajima's D value (Fig. 4 and Dataset S1), with the highest being for the *csp* gene (Tajima's D = 1.57). Most other genes close to the top of the list encode putative proteins that have not been studied previously, although orthologs of particular antigen genes previously shown to be under balancing selection in other malaria parasite species (21) have quite high values here (e.g., the *trap* and *msp1* genes, with ranks of 24 and 61, respectively). However not all antigen genes have such high values; for example, the *P. knowlesi* apical membrane antigen 1 (*ama1*) and *DBP-α* genes had Tajima's D values of −1.68 and −1.31, respectively, close to the genomewide average and in contrast with the elevated values for their orthologs in population studies of other species (21).

Informative tests of allele frequency distributions by Tajima's D require larger numbers of isolates than were sequenced for clusters 2 and 3. To explore the allele frequency distribution in the isolates in cluster 2 (n = 10), Tajima's D values were calculated for 2,117 genes with three or more SNPs, yielding an overall mean value of −0.26, which was much less negative than seen for the larger sample of cluster 1 isolates. To test if this

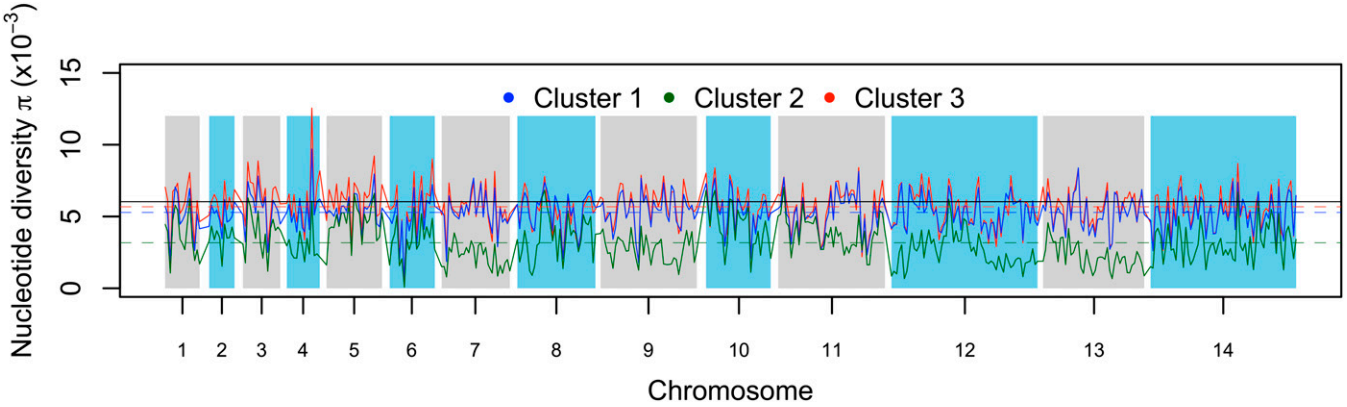

Fig. 2. Distribution of the average nucleotide diversity (π) for sliding windows of 50-kb regions within the three main *P. knowlesi* subpopulation clusters: cluster 1 (n = 38, blue line), cluster 2 (n = 10, green line), and cluster 3 (n = 4, red line). The dotted lines represent the genomewide mean values for the three respective clusters. The solid black line represents the overall nucleotide diversity across all samples.

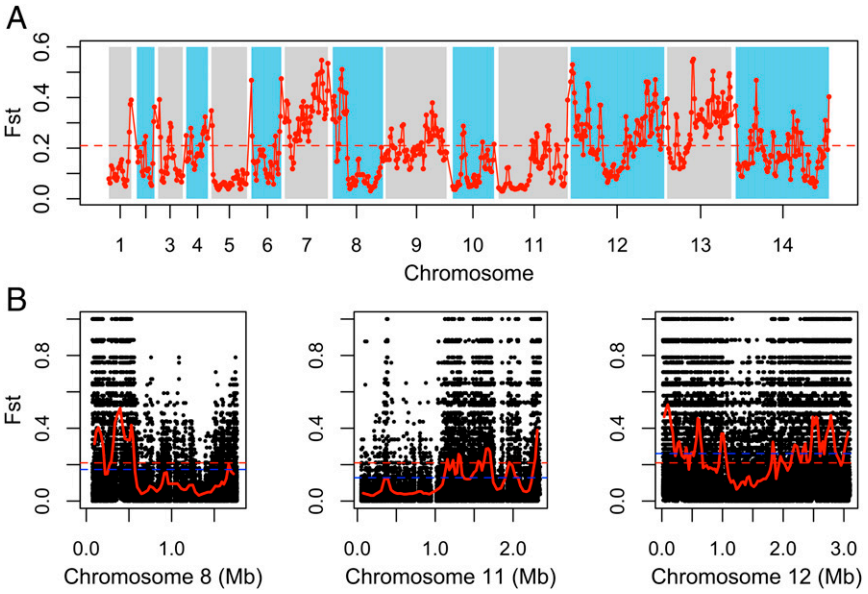

Fig. 3. Genomewide F_{ST} scans between the cluster 1 and 2 subpopulations of *P. knowlesi* clinical isolates. (*A*) Sliding window plot of mean F_{ST} scores for windows of 500 consecutive SNPs shows within-chromosome variation of F_{ST} scores. The blue and gray shades represent alternating chromosomes. The dashed red lines show the genomewide mean F_{ST} value of 0.21. (*B*) Heterogeneity within individual chromosomes illustrated by plots of F_{ST} values for chromosomes 8, 11, and 12. Black dots show values for individual SNPs, and red lines show mean F_{ST} values for consecutive windows of 500 SNPs. Dashed blue lines represent chromosome-wide mean F_{ST} values, and dashed red lines show the genomewide mean F_{ST} value of 0.21.

difference from cluster 1 was an artifact of the small sample size, Tajima's D values were computed from randomly generated subsamples of 10 isolates from cluster 1, yielding a mean value of –0.22, which was not significantly different from that of cluster 2. This result illustrates the low information content of this index with very small numbers of samples and the need for more extensive sampling of clusters 2 and 3 to enable allele frequency-based tests for selection within each of these clusters.

Evidence of Recent Positive Directional Selection. A genomewide scan using both coding and noncoding SNPs from 38 isolates of the main cluster 1 identified 235 positions with integrated haplotype score (|iHS|) values of >4.89 (Fig. 5), which clustered into 57 windows of contiguous elevated haplotype homozygosity covering a total of 1,978 kb. To focus on the most significant regions, we identified nine of these windows that had at least one SNP with |iHS| >6.0 and that did not have gaps of more than 20 kb between consecutive genotyped SNPs, covering a combined total of 593 kb. Individual window sizes ranged from 10–161 kb (Table S4). These windows contained 146 genes (Table S5), 75 (51%) of which encoded conserved *Plasmodium* proteins of unknown function and 11 (8%) of which are annotated as members of the species-specific *SICAvar* gene family. Because SNPs in *SICAvar* genes were excluded from analysis, the five windows that contained *SICAvars* are caused by signals from SNPs in the surrounding regions. Other genes within elevated iHS regions include *csp* and two *P. knowlesi*-specific proteins (PKH_083491 and PKH_112405) with no reported orthologs in other *Plasmodium* species. As expected, given the assumptions that *P. knowlesi* cases in humans are primarily zoonotic and that the reservoir parasite population is not under antimalarial drug selection, no signals of positive selection were observed around five orthologs of known *P. falciparum* drug-resistance genes: the chloroquine resistance transporter (*crt*, PKH_010710), multidrug resistance-1 (*mdr1*, PKH_100920), dihydrofolate reductase (*dhfr*, PKH_052130), dihydropteroate synthase (*dhps*, PKH_142780), and kelch K13 (PKH_121080).

Discussion

This whole-genome population study of zoonotic malaria reveals much higher diversity in *P. knowlesi* than in either of the major

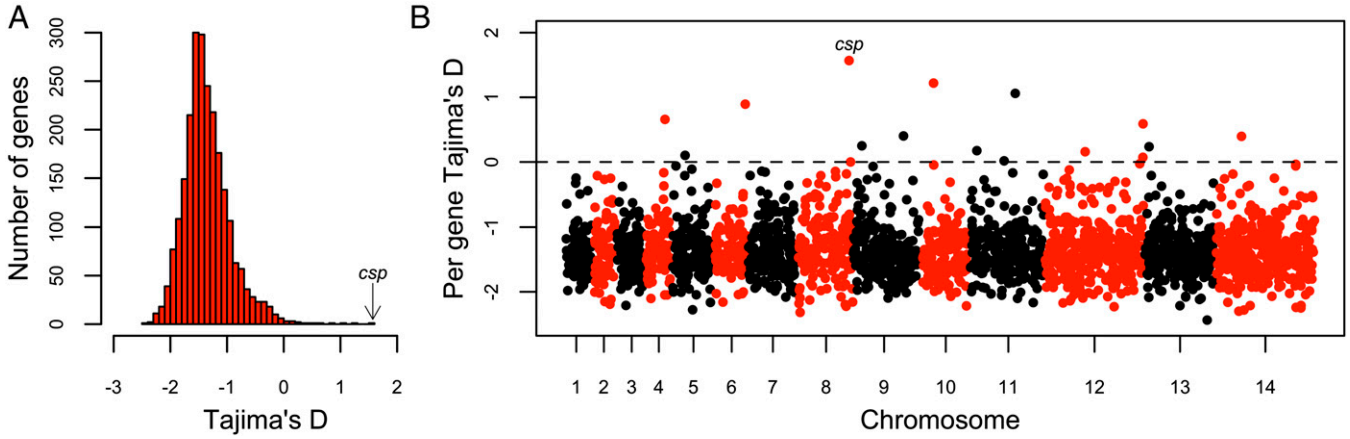

Fig. 4. Scan of Tajima's D values for 2,381 genes with a minimum of three SNPs within the major *P. knowlesi* subpopulation cluster 1 (n = 38 isolates). (*A*) Frequency distribution of Tajima's D values shows a highly negative skew genomewide and only a minority of genes with positive values (the gene with the highest value was *csp*, encoding the circumsporozoite protein). (*B*) Tajima's D values for 2,381 genes with a minimum of three SNPs plotted according to their chromosomal positions (black and red colors indicate consecutive chromosomes numbered from the smallest upwards). Tajima's D values for each of the individual genes are listed in Dataset S1.

endemic human malaria parasite species, *P. falciparum* or *P. vivax* (22–25). Consistent with previous analyses of diversity of mitochondrial genome sequences (11) and microsatellite genotyping (18), the estimate of genomic complexity here confirms that most human infections are dominated by single genotypes. A very deep population substructure was uncovered within this zoonotic species. The mean genomewide differentiation between the two main clusters of clinical isolates (F_{ST} = 0.21) was similar to that indicated by a recent study using microsatellite markers, which indicated that these clusters were associated with long-tailed and pig-tailed macaque reservoir hosts, respectively (18).

A third cluster comprised the small number of lines that have been maintained in laboratory rhesus macaques for many years. It is possible that the divergence of this latter cluster relates to geographical origins, because the isolates originally were obtained in the 1960s from peninsular Malaysia and the Philippines, although sequence analysis here indicates uncertainty in the true identity of some of these isolates. A recent study has indicated dimorphism in whole-genome sequences of six clinical isolates (three belonging to each of two divergent types) from Sarikei in Sarawak (26), where our previous microsatellite analysis identified both cluster 1 and cluster 2 as occurring at approximately equal frequencies (18).

Although it is likely that strong reproductive isolation is maintained between these sympatric subpopulations, mosaicism in the level of differentiation across the genome is apparent. Further sampling and population genetic analyses may address whether there has been occasional hybridization leading to introgression of parts of the genome between the different subpopulations, potentially enabling parasites to establish infections in different host species, including humans. Introgression between subspecies has been associated with gain of virulence in another human eukaryotic pathogen, *Trypanosoma brucei* (27), and was reported potentially to have enhanced growth rates and infectivity in hybrids between schistosome populations from cattle and human hosts (28). Such adaptive roles of introgression have been implicated in other cases, including adaptation of pigs to hot and cold climates (29). Although differences in F_{ST} have been used to scan for genomic islands related to speciation in other taxa, such measures can be misleading if values are inflated by a lack of diversity in particular populations (30). Because in the present study there were fewer isolates in cluster 2 than in cluster 1, it will be valuable to analyze a larger number of cluster 2 isolates in future to investigate further the genomewide variation in divergence.

Because of the pronounced population substructure, the scan for signatures of selection in this study was focused on the majority of the clinical isolates that appeared to belong to a single, undivided population, the 38 isolates in cluster 1. The overall negative skew of allele frequency distributions as shown by the Tajima's D values suggests previous long-term population growth of this parasite. It is interesting that the *csp* gene appears to be under stronger balancing selection in *P. knowlesi* (having the highest Tajima's D value genomewide) than in *P. falciparum* (in which its ortholog has a weaker positive value), whereas the *ama1* gene did not have an elevated value in *P. knowlesi* although it is consistently identified as under balancing selection in other species (21). To investigate whether these or other loci show evidence of selection within clusters 2 and 3, further sampling is required to obtain sufficient numbers of isolate sequences of these rarer clusters.

The high integrated haplotype scores observed in multiple chromosomal regions suggests current or recent occurrence of strong positive selection on a number of different genes. In *P. falciparum* population studies, such strong selection signals were reported around drug-resistance loci where selective sweeps recently have increased the frequency of resistance-associated haplotypes (31). The absence of any signature of positive selection on the *P. knowlesi* orthologs of drug-resistance genes is consistent with the hypothesis that most selection on these parasites occurs in macaque rather than human hosts (11), specifically in long-tailed macaques for the cluster 1 subpopulation (18). Interestingly, the *csp* gene was located within an elevated window of haplotype homozygosity on chromosome 8, suggesting that it could be a target of both balancing and directional selection or that it might have hitchhiked to intermediate allele frequencies by a linked locus under selection within the cluster 1 subpopulation. The relatively small number of isolate sequences of cluster 2 precluded comparable tests of selection for that subpopulation, although, given the different population structure and association with pig-tailed macaque reservoir hosts (18), it will be important to investigate selection in cluster 2 by further sampling.

To understand further the signatures of selection and to test for evidence of host-switching and possible adaptation to humans, it is also a priority to obtain whole-genome sequence data from *P. knowlesi* parasites in natural infections of macaque reservoir hosts. Although these data are difficult to obtain because of the generally low densities of parasites within the blood of these animals and because most macaques are infected with multiple

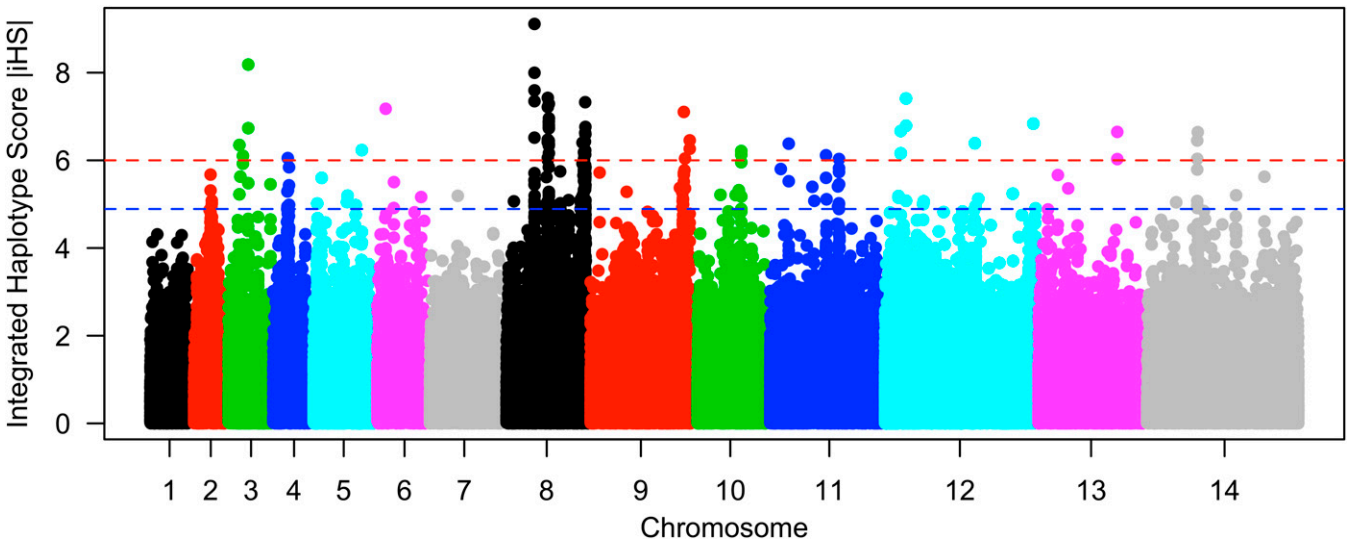

Fig. 5. Scan for evidence of recent positive selection in the main *P. knowlesi* subpopulation cluster 1. Plot of genomewide |iHS| scores shows regions of the genome that have windows of elevated values, consistent with the operation of recent positive directional selection. The dashed lines represent values of 4.89 (blue) and 6 (red), used to define nine windows containing SNPs with overlapping regions of extended haplotype homozygosity, as described in *Materials and Methods*. The coordinates of these windows and the genes within them are listed in Tables S4 and S5.

species of *Plasmodium* (11, 32), a promising approach may be flow cytometric sorting of individual parasites for single-cell sequencing, as recently applied to human malaria parasites (33). In addition, future culture adaptation of *P. knowlesi* clones from the divergent genotypic clusters and the application of high-efficiency transfection technologies to study gene functions systematically could enable the investigation of phenotypic differences. These studies could include attempts at genetic crosses by feeding mixed cultures to vector mosquitoes to identify competent vectors for each of the clusters and test for the possible existence of reproductive isolation mechanisms maintaining the differences between the clusters.

Materials and Methods

Peripheral blood samples were collected from patients with *P. knowlesi* malaria at Betong and Kapit Hospitals in Sarawak, Malaysian Borneo, following informed consent and with the approval of the Medical Research and Ethics Committee of the Ministry of Health, Malaysia, the Ethics Committee of the London School of Hygiene and Tropical Medicine, and the Institutional Biosafety and BioEthics Committee of King Abdullah University of Science and Technology, Saudi Arabia. *P. knowlesi* genomic DNA was prepared after depletion of human leukocytes (34) and processed for Illumina paired-end short-read sequencing. DNA samples from five laboratory strains of *P. knowlesi*, isolated in the 1960s and subsequently maintained in laboratory rhesus macaques, were similarly sequenced. Three of the laboratory strains (labeled "Hackeri," "Malayan," and "H") were originally isolated from peninsular Malaysia (35, 36), and one (labeled as "Philippine") was from the Philippines (37). They were obtained through the American Type Culture Collection (ATCC) and MR4 repository, and, along with a previously in vitro adapted line of the H strain (38, 39), here termed "H(AW)," were culture-adapted to rhesus macaque erythrocytes in vitro, cloned by limiting dilution and grown for 1–2 wk before DNA extraction.

For additional information on patient sampling and laboratory strains, DNA library preparation and genome sequencing, quality control and processing of sequence data (40, 41), and population genetic and comparative sequence analyses (24, 42–46), see *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank all patients, nurses, doctors, and laboratory technicians at Kapit and Betong Hospitals and core staff in our research institutes for their assistance; Hifzur Rahman Ansari (King Abdullah University of Science and Technology) for submission of the sequences to European Nucleotide Archive; the Malaria Research and Reference Reagent Resource Center for providing the *P. knowlesi* strains contributed by William E. Collins; and the Director General of Health in Malaysia for permission to publish this paper. This study was supported by Universiti Malaysia Sarawak Grants E14054/F05/54PK1/09/2012(01) and 01(TD03)/1003/2013(01); UK Medical Research Council Grants MRC G1100123, MR/K000551/1, and MR/M01360X/1; European Research Council Advanced Award AdG-2011-294428; NIH Grant 5R01AI091787; Bill and Melinda Gates Foundation Grant OPP1023594; a postgraduate scholarship from the Ministry of Education in Malaysia; and faculty baseline funding from King Abdullah University of Science and Technology.

1. William T, et al. (2011) Severe *Plasmodium knowlesi* malaria in a tertiary care hospital, Sabah, Malaysia. *Emerg Infect Dis* 17(7):1248–1255.
2. Singh B, Daneshvar C (2013) Human infections and detection of *Plasmodium knowlesi*. *Clin Microbiol Rev* 26(2):165–184.
3. Daneshvar C, et al. (2009) Clinical and laboratory features of human *Plasmodium knowlesi* infection. *Clin Infect Dis* 49(6):852–860.
4. Cox-Singh J, et al. (2008) *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. *Clin Infect Dis* 46(2):165–171.
5. Garnham P (1966) *Malaria Parasites and Other Haemosporidia* (Blackwell Scientific Publications Ltd., Oxford, UK).
6. Singh B, et al. (2004) A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet* 363(9414):1017–1024.
7. Tyagi RK, Das MK, Singh SS, Sharma YD (2013) Discordance in drug resistance-associated mutation patterns in marker genes of *Plasmodium falciparum* and *Plasmodium knowlesi* during coinfections. *J Antimicrob Chemother* 68(5):1081–1088.
8. Müller M, Schlagenhauf P (2014) *Plasmodium knowlesi* in travellers, update 2014. *Int J Infect Dis* 22:55–64.
9. Lee KS, Cox-Singh J, Singh B (2009) Morphological features and differential counts of *Plasmodium knowlesi* parasites in naturally acquired human infections. *Malar J* 8:73.
10. Lee KS, Cox-Singh J, Brooke G, Matusop A, Singh B (2009) *Plasmodium knowlesi* from archival blood films: Further evidence that human infections are widely distributed and not newly emergent in Malaysian Borneo. *Int J Parasitol* 39(10):1125–1128.
11. Lee KS, et al. (2011) *Plasmodium knowlesi*: Reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathog* 7(4):e1002015.
12. Tan CH, Vythilingam I, Matusop A, Chan ST, Singh B (2008) Bionomics of *Anopheles latens* in Kapit, Sarawak, Malaysian Borneo in relation to the transmission of zoonotic simian malaria parasite *Plasmodium knowlesi*. *Malar J* 7:52.
13. Vythilingam I, et al. (2008) *Plasmodium knowlesi* in humans, macaques and mosquitoes in peninsular Malaysia. *Parasit Vectors* 1(1):26.
14. Collins WE (2012) *Plasmodium knowlesi*: A malaria parasite of monkeys and humans. *Annu Rev Entomol* 57:107–121.
15. William T, et al. (2013) Increasing incidence of *Plasmodium knowlesi* malaria following control of *P. falciparum* and *P. vivax* Malaria in Sabah, Malaysia. *PLoS Negl Trop Dis* 7(1):e2026.
16. Jongwutiwes S, et al. (2011) *Plasmodium knowlesi* Malaria in humans and macaques, Thailand. *Emerg Infect Dis* 17(10):1799–1806.
17. Jeslyn WP, et al. (2011) Molecular epidemiological investigation of *Plasmodium knowlesi* in humans and macaques in Singapore. *Vector Borne Zoonotic Dis* 11(2):131–135.
18. Divis PCS, et al. (2015) Admixture in humans of two divergent *Plasmodium knowlesi* populations associated with different macaque host species. *PLoS Pathog* 11(5): e1004888.
19. Pain A, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455(7214):799–803.
20. Gardner MJ, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906):498–511.
21. Conway DJ (2015) Paths to a malaria vaccine illuminated by parasite genomics. *Trends Genet* 31(2):97–107.
22. Volkman SK, et al. (2007) A genome-wide map of diversity in *Plasmodium falciparum*. *Nat Genet* 39(1):113–119.
23. Neafsey DE, et al. (2012) The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nat Genet* 44(9):1046–1050.
24. Manske M, et al. (2012) Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* 487(7407):375–379.
25. Mu J, et al. (2007) Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat Genet* 39(1):126–130.
26. Pinheiro MM, et al. (2015) *Plasmodium knowlesi* genome sequences from clinical isolates reveal extensive genomic dimorphism. *PLoS One* 10(4):e0121303.
27. Goodhead I, et al. (2013) Whole-genome sequencing of *Trypanosoma brucei* reveals introgression between subspecies that is associated with virulence. *MBio* 4(4): e00197-13.
28. Huyse T, et al. (2009) Bidirectional introgressive hybridization between a cattle and human schistosome species. *PLoS Pathog* 5:e1000571.
29. Ai H, et al. (2015) Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat Genet* 47:217–225.
30. Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* 23: 3133–3157.
31. Nwakanma DC, et al. (2014) Changes in malaria parasite drug resistance in an endemic population over a 25-year period with resulting genomic evidence of selection. *J Infect Dis* 209(7):1126–1135.
32. Muehlenbein MP, et al. (2015) Accelerated diversification of nonhuman primate malarias in Southeast Asia: Adaptive radiation or geographic speciation? *Mol Biol Evol* 32(2):422–439.
33. Nair S, et al. (2014) Single-cell genomics for dissection of complex malaria infections. *Genome Res* 24(6):1028–1038.
34. Venkatesan M, et al. (2012) Using CF11 cellulose columns to inexpensively and effectively remove human DNA from Plasmodium falciparum-infected whole blood samples. *Malar J* 11:41.
35. Chin W, Contacos PG, Coatney GR, Kimball HR (1965) A naturally acquired quotidian-type malaria in man transferable to monkeys. *Science* 149(3686):865.
36. Wharton RH, Eyles DE (1961) *Anopheles hackeri,* a vector of *Plasmodium knowlesi* in Malaya. *Science* 134(3474):279–280.
37. Collins WE, Contacos PG, Chin W (1978) Infection of the squirrel monkey Saimiri sciureus, with Plasmodium knowlesi. *Trans R Soc Trop Med Hyg* 72(6):662–663.
38. Kocken CH, et al. (2002) *Plasmodium knowlesi* provides a rapid in vitro and in vivo transfection system that enables double-crossover gene knockout studies. *Infect Immun* 70(2):655–660.
39. Lim C, et al. (2013) Expansion of host cellular niche can drive adaptation of a zoonotic malaria parasite to humans. *Nat Commun* 4:1638.
40. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
41. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
42. Auburn S, et al. (2012) Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS One* 7(2):e32891.
43. Mobegi VA, et al. (2014) Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Mol Biol Evol* 31(6):1490–1499.
44. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.
45. Gautier M, Vitalis R (2012) rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 28(8):1176–1177.
46. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4(3):e72.