SUPPLEMENTARY MATERIALS

Molecular Diagnostics

P. vivax Infection Detection

DNA was extracted from dried blood spots using Chelex-100 (Bio-Rad, Hercules, CA) and Saponin as previously described.[1,2] *P. vivax* infections were detected using a two-stage approach that combined a TaqMan quantitative PCR (qPCR) assay targeting the 18S rRNA gene and a confirmatory nested-PCR assay.[3,4] A two-step approach was utilized to increase specificity and limit potential false positives given the range of cycle-threshold (CT) values considered.

For the qPCR assay, individual reactions were performed at a final volume of 18 µL: 5 µL of template DNA, 9 µL of FastStart Universal Probe Master Mix (ROX) (Roche Diagnostics, Indianapolis, IN), 0.36 µL of each primer at 20 µM, 0.36 µL of probe at 10 µM, and 2.92 µL of molecular grade water (Supplementary Table 1). All qPCR reactions were performed on a QuantStudio 6 Flex Real-Time PCR System (ThermoFisher Scientific, Waltham, MA, USA) using the following thermal cycling conditions: hold at 50°C for 2 minutes, denaturation at 95°C for 10 minutes, and 45 cycles of 95°C for 15 seconds and annealing at 60°C for 1 minute. All bulk qPCR reactions included two replicates of positive controls: 10-fold standard dilutions of a *P. vivax* 18S PCR plasmid clone (MRA-178, BEI Resources, Manassas, VA, USA) from 4,550 parasites/µL (10^{-4} ng/µL) to 4.55 parasites/µL (10^{-7} ng/µL), assuming 6 copies of 18S per parasite.[5,6] In addition, four non-template controls were added to each qPCR plate. For all qPCR runs, a threshold of 0.04 log change in Rn was set and results were then exported for analysis. Any samples with uncharacteristic amplification profiles, or samples with amplification prior to 13 cycles, were marked as undetermined.

To confirm the presence of *P. vivax* infections, all samples that were positive by qPCR underwent reflex confirmatory screening using a nested PCR assay.[4] The first round of the nested PCR assay targeted a 1.6-1.7 kilobase region of the 18S gene specific to the *Plasmodium* genus using the Plu1 and Plu5 primers (Supplementary Table 1). For the first round PCR, reactions were performed in a final volume of 25 µL: 5uL of sample DNA, 12.5uL HotStarTaq Mastermix (Qiagen, Venlo, Netherlands), 0.5 µL of 20 µM each of the forward and reverse primers, and 6.5 µL molecular grade water. PCRs were performed on a BioRad T100 Thermal Cycler (Applied Biosystems, Foster City, CA, USA) using the following thermocycler conditions: 95°C for 15 minutes followed by 35 cycles of 94°C for 1 minute, annealing at 50°C for 1 minute, and extension at 72°C for 1 minute, with a final extension at 72°C for 10 minutes. Product from the first round of PCR was handled in a designated post-PCR section of the laboratory to inhibit contamination of pre-PCR work surfaces. The second round PCR reaction targeted a 121 base pair region specific to *P. vivax* using the rViv1 and rViv2 primers (Supplementary Table 1). The second round reaction recipe was identical to the first round reaction recipe with the exception that the PCR product from the first round was used as the DNA template for the second round. Thermal cycling conditions were the same with the exception of a raised annealing temperature of 62°C. Final PCR product was visualized using 5% ethidium bromide on a 1% agarose gel run in 1x Tris-borate-EDTA buffer at 100 volts for 1 hour. Positive confirmation of *P. vivax* was based on visualization of the 121 base pair PCR amplicon and was evaluated by two reviewers independently. A confirmed infection was only considered when reviewers were in agreement. Among the 579/17,972 qPCR-positive samples, the inter-observer agreement of the absence/presence of a PCR band was high (Agreement: 564/579, Cohen's $\kappa = 0.80$).

Assay	Primer	Sequence	Ref.
Diagnostic qPCR	PvForward	5'-ACGCTTCTAGATTAATCCACATAACT	[3]
	PvReverse	5'-ATTTACTCAAAGTAACAAGGACTTCCAAGC	
	Pv-probe (FAM-IowaBlack)	5'-TTCGTATCG/ZEN/ACTTTGTGCGCATTTTGC	
Confirmatory PCR	Plu1	5'-TCAAAGATTAAGCCATGCAAGTGA	[4]
	Plu5	5'- CCTGTTGTTGCCTTAAACTCC	
	rVivi1	5'-CGCTTCTAGCTTAATCCACATAACTGATAC	
	rVivi2	5'-ACTTCCAAGCCGAAGCAAAGAAAGTCCTTA	

Supplementary Table 1 - *P. vivax* Infection Detection Assays: The adapted protocols used for diagnostic qPCR and confirmatory PCR in the detection of *P. vivax* infections. For the qPCR assay, the probe differs from Srisutham *et al.* 2017 in the use of FAM as the fluorescent label, an additional ZEN quencher, and the use of Iowa Black as the 3' terminus quencher. All probes and primers were synthesized by Integrated Device Technology, Inc. (San Jose, CA, USA).

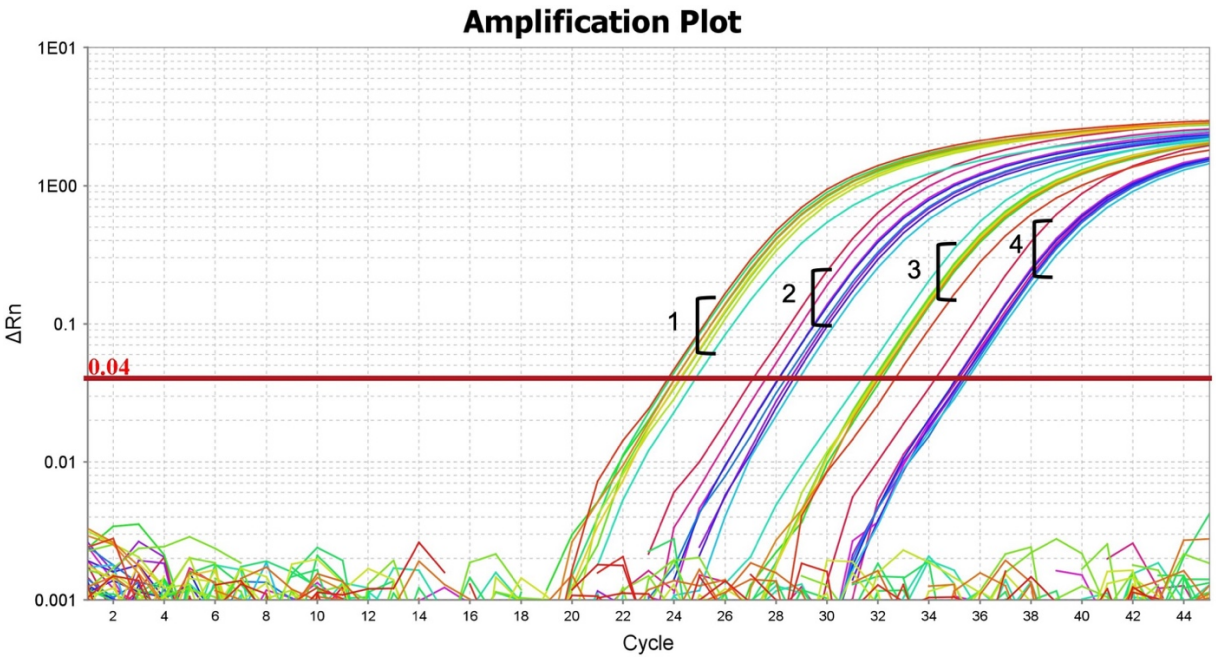
P. vivax Infection Assay Validation and Challenge

To assess the lower limit of detection of the *P. vivax* TaqMan assay, the *P. vivax* 18S qPCR plasmid clone (MRA-178, BEI Resources, Manassas, VA, USA) was diluted to 1.0 x 10^{-6} ng/µL and then serially diluted 2-fold to a lower limit of 0.03125 x 10^{-6} ng/µL. The assay was then ran with 12 replicates for the 1.0 x 10^{-6} and five 1.0 x 10^{-5} dilutions and with 22 replicates for the remaining dilutions (Supplementary Table 2). Cycle threshold values were read at a log change in Rn value of 0.04—identical to settings used for the detection of *P. vivax* infections. Overall, the 95% lower limit of detection for *P. vivax* was approximately 7 parasites/µL (assuming 6 copies of 18S per parasite), or a DNA concentration of 1.50 x 10^{-7} ng/µL (Supplementary Table 2).

In addition, the specificity of the assay was challenged using 20 replicates of 18S plasmid concentrated at 1.0 x 10^{-4} ng/µL for *P. falciparum* (MRA-177, BEI Resources), *Plasmodium malariae* (MRA-179, BEI Resources), and *Plasmodium ovale* (MRA-180, BEI Resources), respectively.[5,6] In all challenges, no off-target amplification was observed (Supplementary Figure 1).

Target concentration (x 10^{-6}) ng/µL	No. Tested	No. Detected	Percent Detected	Average parasites/µL	Mean Ct (of detected)
Pv 18S 1	12	12	100	45.50	32.68
Pv 18S 0.5	12	12	100	22.75	34.40
Pv 18S 0.25	22	22	100	11.38	35.98
Pv 18S 0.125	22	20	90.91	5.69	37.45
Pv 18S 0.0625	22	14	63.64	2.84	38.69
Pv 18S 0.03125	22	6	27.27	1.42	38.54
Pf 18S 100	20	0	0	--	--
Pm 18S 100	20	0	0	--	--
Po 18S 100	20	0	0	--	--

Supplementary Table 2 - *P. vivax* Lower Limit of Detection and Off-Target Amplification: The concentrations of the 18S plasmid for each of the *Plasmodium spp.: P. vivax* (Pv), *P. falciparum* (Pf), *P. malariae* (Pm), and *P. ovale* (Po), are provided alongside the number of replicates tested. The percent detected was calculated as the proportion of replicates identified with a cycle threshold value less than 45 among those tested. Among those samples that were detected, the mean cycle threshold value is provided. *Abbreviations*: No. – number, Ct – cycle threshold.



Supplementary Figure 1 - *P. vivax* qPCR Specificity Challenge: Amplification of *P. vivax* 18S plasmid dilutions containing 4,550 (1), 455 (2), 45.5 (3), and 4.55 (4) parasites/µL. Each *P. vivax* dilution was tested with eight replicates. Amplification failed for all 20 replicates of *P. falciparum*, *P. malariae,* and *P. ovale* 18S plasmids (concentration: $1x10^{-4}$ ng/µL).

Duffy-Genotyping

For each sample that was positive by qPCR, we used a previously validated high-resolution melt (HRM) assay to genotype the GATA-1 transcription factor (-33 T:C) point mutation that has been previously shown to silence Duffy Antigen/Chemokine Receptor (DARC) expression.[7,8] Each HRM reaction contained a final concentration of 1x MeltDoctor HRM Master Mix (Applied Biosystems, Foster City, CA, USA), 0.3 µM forward primer (DARCf), 0.3 µM reverse primer (DARCr), 100 pg of template DNA in a final volume of 20 µM (Supplementary Table 3). Reactions were performed using the following thermocycler conditions: denaturation at 95°C for 10 minutes, followed by 45 cycles of 95°C for 15 seconds, 60°C for 1 minute, 95°C for 10 seconds, 60°C for 1 minute, 95°C for 15 seconds, and 60°C for 15 seconds on a QuantStudio 6 Flex Real-Time PCR System (ThermoFisher Scientific, Waltham, MA, USA). Each HRM plate contained a DARC-positive (-33 C:C), DARC-negative (-33 T:C), and a non-template control which were used to call HRM results on each plate independently.

Samples that could not be definitively determined by HRM and a 10% random subset of *P. vivax* qPCR-positive samples underwent confirmatory Sanger sequencing genotyping at Eton Bioscience (Research Triangle, NC). PCR products were generated from a previously validated assay.[9] Final reactions contained 0.25 µL of FastStart High Fidelity Taq (Enzyme Blend; Roche©, Indianapolis, IN), 2.5 µL of 10x FastStart High Fidelity reaction buffer with 18 mM $MgCl_2$, 0.36 µM forward primer, 0.36 µM reverse primer, 250 µM dNTPs and 3 µL of template DNA in a volume of 25 µL. Reactions were amplified using the following thermocycler conditions: denaturation at 94°C for 15 minutes followed by 40 cycles of 94°C for 30 seconds, annealing at 58°C for 30 seconds, extension at 72°C for 90 seconds, and a final extension at

72°C for 10 minute on a BioRad T100 Thermal Cycler (Applied Biosystems, Foster City, CA, USA). PCR products and Sanger sequences were also generated for a DARC-positive control (-33 C:C) and DARC-negative control (-33 T:C).

For each sample, forward and reverse sequences were analyzed using Geneious 10.1.3 (Biomatters Limited, Auckland, New Zealand). First, the 5' and 3' ends of each sequence was trimmed using Geneious `Trim Ends` tool with a 0.05 error probability limit. For each sample, forward and reverse sequences were then *de novo* assembled using the Geneious `Assembler` tool with the sensitivity flag set to "Highest Sensitivity/Slow". Of the 51 randomly samples sequenced, one sample was unable to be assembled due to low sequencing quality. Of the 17 samples that underwent confirmatory sequencing, all samples were assembled. The mapped sequences were then visually assessed for the DARC (-33 T:C) point mutation. Duffy-Genotypes by Sanger sequencing were concordant with the HRM-qPCR results among the remaining 50/51 samples selected for validation.

Assay	Primer	Sequence	Ref.
HRM Genotyping	DARCf	5'-CGTGGGGTAAGGCTTCCTGA	[7]
	DARCr	5'-CTGTGCAGACAGTTCCCCAT	
Confirmatory PCR	ESf	5'-GTGGGGTAAGGCTTCCTGAT	[9]
	ESr	5'-CAAACAGCAGGGGAAATGAG	

Supplementary Table 3 - DARC-Genotyping Primers: All samples that were positive by qPCR underwent genotyping at the Duffy Antigen/Chemokine Receptor (DARC) promoter region using High Resolution Melt (HRM) Analysis. A subset of randomly selected samples and those samples that could not be absolutely confirmed by HRM underwent confirmatory Sanger sequencing of a GATA-1 transcription factor amplicon that contained the region of interest.

Epidemiological Analyses

Study Population and Data Sources

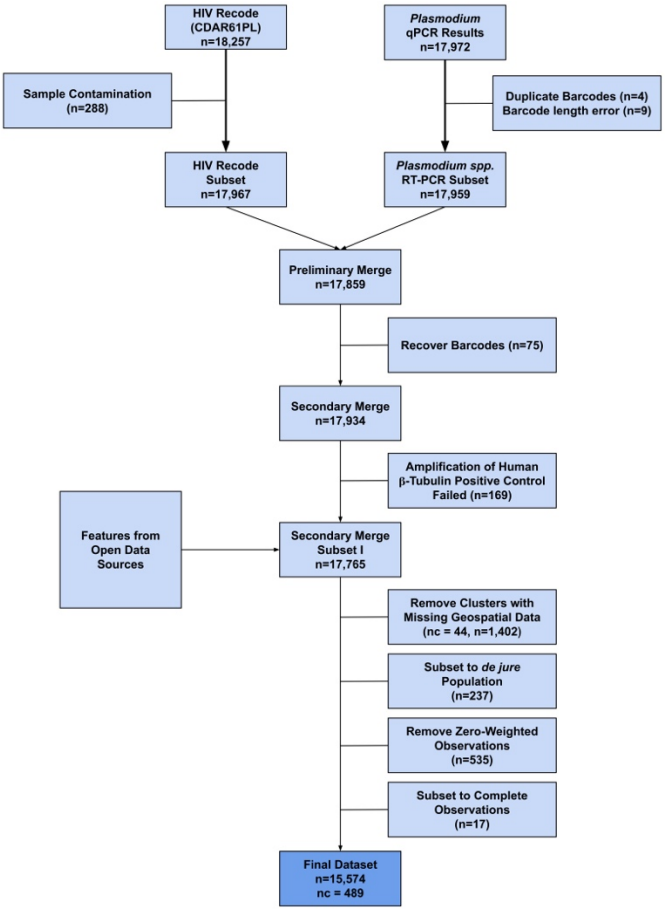
In the Democratic Republic of the Congo (DRC), the Demographic Health Survey (DHS) aims to create a nationally representative survey using a two-stage stratified cluster sampling design.[10] In the first stage, clusters, or enumeration area, are selected with a known and fixed probability. During the second stage, within each cluster, a subset of households are selected. Finally, among those adults residing in selected households, a subset are consented for HIV and other biomarker testing. To control for this sampling scheme, the DHS weights each individual with an inverse probability weights of selection, hereafter, sampling weights.[10]

The DRC 2013-2014 DHS survey was conducted from August 2013 - September 2013 and November 2013 - February 2014. Specifically, DHS surveyors screened Kinshasa and surrounding areas from August 2013 - September 2013 and then subsequently administered the survey across the rest of the country from November 2013 - February 2014.

For each household, DHS surveyors acquired informed consent and administered a substantial questionnaire to all individuals that had slept in the household the night prior to the interview.[10] Individuals that permanently reside in the household are classified as *de jure* while individuals that were coincidentally in the household the night preceding the interview were classified as *de facto*.[10] Given that household variables were considered as potential malaria risk factors, we limited observations to the *de jure* population, as *de facto* individuals' homes may differ substantially from the home that they were visiting.[10]

Among those adults that agreed to undergo HIV and other biomarker testing, a dried blood spot (DBS) was taken. DBS were then punched into 96-well plates and associated barcodes were manually recorded in a spreadsheet in the DRC. The 96-well plates were then sent to the University of North Carolina-Chapel Hill (UNC) for malaria testing.

In total, 17,959/17,972 samples with properly formatted barcodes were screened by qPCR at UNC. These samples were then linked to the DHS HIV (AR) recode excluding the 288 samples that were contaminated during shipment from the DRC to UNC. On the initial merge, 17,859/17,959 samples were successfully linked. In order to recover more samples, we allowed for a one-character mismatch between the manually recorded DBS barcode and the DHS barcode among those samples that did not have a match in the preliminary merge. Using this strategy, we successfully recovered an additional 75 samples accounting for our total of 17,934/17,959 samples that were screened by qPCR. Among these 17,934 samples, 169 samples failed to amplify human beta-tubulin, which was used as a within-sample positive control, and thus, were excluded from the study population.[1] Of these 17,765 samples, 1,402 were missing geospatial data (44 clusters), 237 individuals were not *de jure* household members, 535 have sampling-weights set to zero, and 17 had missing risk-factor covariate information and were excluded from the study. As a result, the total study population consisted of 15,574 individuals (Supplementary Figure 3). We assumed that all samples lost due to shipping contamination, failure to amplify human beta-tubulin, barcode typos, and missing geospatial data were due to "accidents" and were missing completely at random. Additional samples were excluded under the DHS sampling framework (i.e. missing sampling weights, *de jure*).[10] As a result, from this refined dataset, we had only 17/15,591 (0.11%) observations that were missing conditional on factors not considered in our study.



Supplementary Figure 2 - Flowchart of Study Participants that were Included in the Study: Of the 18,257 Demographic Health Survey (DHS) records that had a dried blood spot, 15,574 were included in the final study population. *Abbreviations*: Quantitative polymerase chain reaction - qPCR.

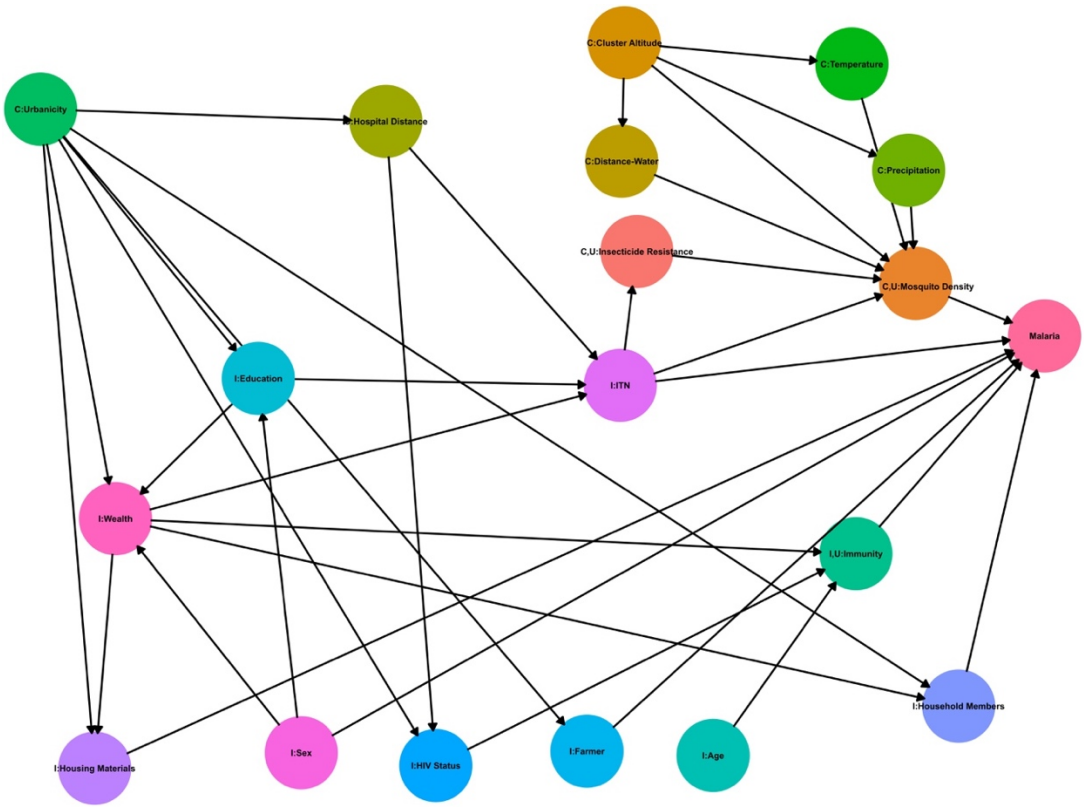
Covariate Feature Engineering

From the DHS questionnaires, we used data from the household members recode (PR), the HIV testing recode (AR), the geospatial covariate (GC) dataset, and the geographical dataset (GE).[10] Data from the CD2013 was downloaded using the `rDHS` package.[11]

In addition to the data provided by the DHS, we downloaded data from several open sources, including: (1) waterways lines and polygon shape-files for the DRC from the Humanitarian OpenStreetMap Team database (https://data.humdata.org/dataset/hotosm_cod_waterways; accessed October 30, 2019); (2) locations of public hospitals within sub-Saharan Africa (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JTL9VY; Accessed October 30, 2019) [12]; and (3) non-human ape (NHA) territories from the International Union for Conservation of Nature Red List database (https://www.iucnredlist.org/; accessed January 21, 2019). Temperature data for the 2013-2014 DRC DHS study period was downloaded from the Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (Goddard Space Flight Center, Greenbelt, MA). Specifically, we downloaded monthly

layers of land surface temperature and emissivity data from the MYD11C3 (v6) product with a 0.05° x 0.05° spatial resolution (accessed September 20, 2019).[13] Monthly precipitation data with a 0.05° x 0.05° spatial resolution was downloaded from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) server using an R-wrapper package (GitHub: `environmentalinformatics-marburg/heavyRain`) for the CD2013 study period.[14] OpenStreetMap extracts from Geofabrik (https://www.geofabrik.de/data/download.html) for Africa (accessed August 23, 2019) were downloaded and used as the road network input into the Open Source Routing Machine (`ORSM`) tool.[15] Finally, additional map features included: (1) ocean spatial polygons from Natural Earth (naturalearthdata.com); (2) geographical base-map layers from the Database of Global Administrative Areas (http://www.gadm.org/); and (3) country geographies from the R-package, `rnaturalearth`.[16]

Prior to analysis, we identified risk factors for *P. vivax* and *P. falciparum* from a comprehensive literature review.[1,17–19] The relationships among risk factors and our outcome of interest, malaria (i.e. either *P. vivax* or *P. falciparum*) was modeled using a directed acyclic graph (DAG) with the `daggity` graphical user interface and R-package (Supplementary Figure 3).[20] As a result, not all risk factors identified were measured and included in the analysis. Although anemia and anti-malarial use were considered to be *a priori* risk factors, both were determined to have cyclic relationships with our outcome of interest, malaria, and were excluded (i.e. anemia and anti-malarial use could not be resolved by the DAG).



Supplementary Figure 3 - Malaria Risk Factor Directed Acyclic Diagram: Risk factors were identified from an extensive literature search. Similarly, the causal relationships among the risk factors were based on the literature review and putative associations. Based on our directed acyclic diagram (DAG), we expected urbanicity, altitude, age, and biological sex to all be unconfounded in expectation (no ancestor nodes).

The majority of risk factors were abstracted from the DHS recodes and kept in their original form with the exception of standardizing continuous variables. Dichotomized variables were set to have an *a priori* protective referent level. Housing type was coded as either "traditional" or "modern" based on a composite score of floor, wall, and roof type as previously outlined by Tustings *et al.* 2017. We also considered any house that had a metal roof as "modern", given recent findings that metal roofs alone appear to be protective against malaria.[21]

Given that the DHS wealth variable accounts for housing type in its calculation, we recreated the wealth variable in order to avoid issues of collinearity and non-independence between the housing and wealth covariates.[10,19,22] The wealth covariate was recreated using the factor-score approach based on the instructions by Rustein 2015 and Tustings *et al.* 2017. Wealth factor scores were then considered as a continuous covariate in order to smooth over issues of positivity in wealth and residual confounding.

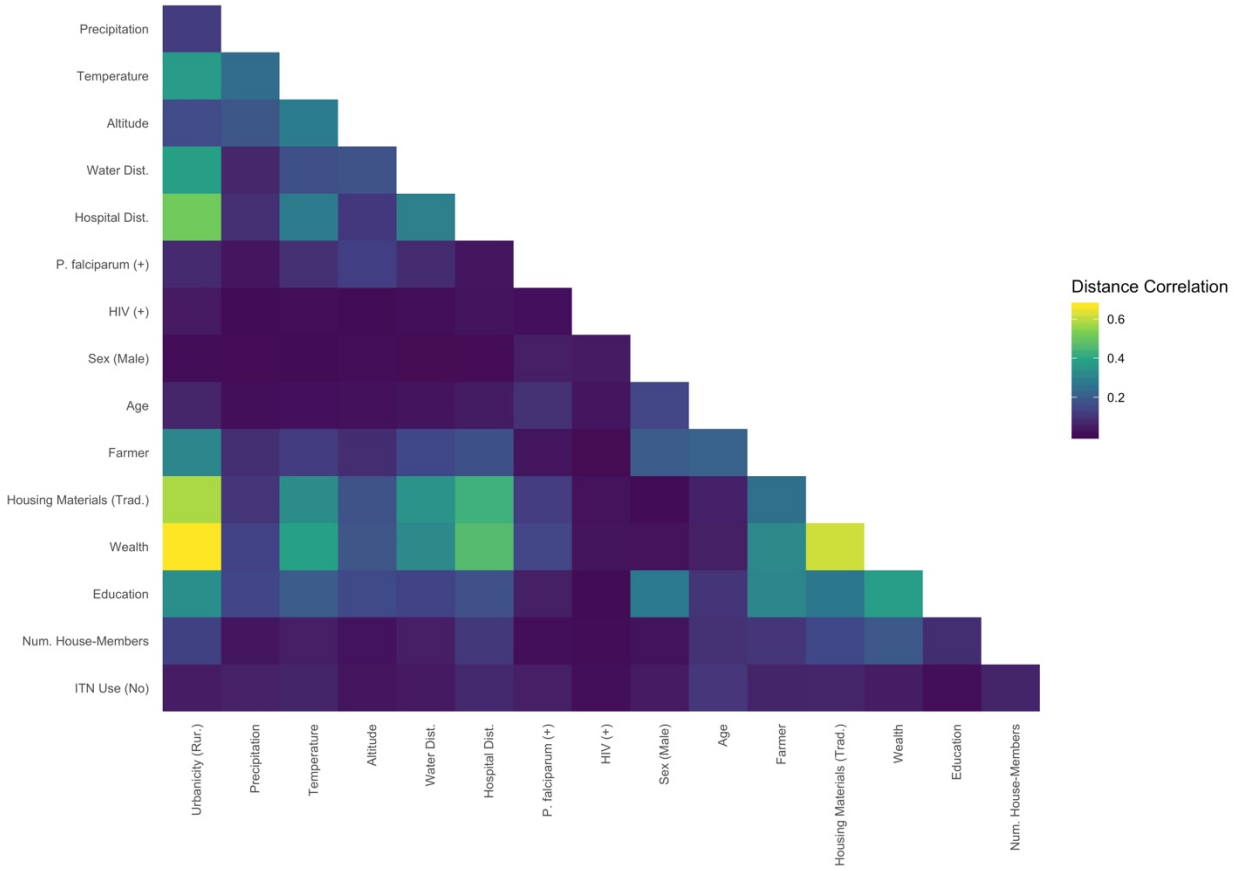
We defined insecticide treated net (ITN) usage based on the definition outlined in Tustings *et al.* 2017, which limits the ITN classification to long-lasting insecticidal nets less than or equal to three-years-old at the time of the survey, convention ITNs that were less than or equal to one-year-old at the time of the survey, or any net that was retreated within a year of the survey. All other net-usage was coded as "no net" alongside those individuals that reported not using a net the night prior to the survey.[10]

The distance from a hospital covariate was coded as the average duration of travel in minutes between a respective cluster and all public hospitals within the cluster's catchment area. A catchment area was defined as a circle with a 100 km radius with the cluster's location as the centroid. Catchment areas were considered in order to better approximate overall cluster accessibility to health-sites, which may otherwise be biased if a cluster is close to a single hospital but far from all others. If all hospitals were farther than 100 km from a given cluster, the minimum duration between the cluster and all hospitals was considered in place of the catchment area. Travel times were calculated using the OSRM tool.[15] Among the 489 clusters considered, one cluster (469) could not be resolved by OSRM. As a result, the hospital distance for cluster 469 was considered as the average duration among its five nearest-neighbor clusters. Clusters were then coded as "near" or "far" from public hospitals if they were within 120 minutes of average travel time or not, respectively.[12] Distance to water was measured as the minimum greater circle distance between a cluster and a body of water that was either labeled as a "river" or "lake" by the OpenStreetMap water-type (Humanitarian OpenStreetMap Team database). Greater circle distances were measured using the R `sf` package.[23,24]

Given that the 2013-2014 DRC DHS was conducted in two phases, with the first phase contained to Kinshasa and surrounding regions during months that coincided with the dry-season, while the remaining areas were surveyed during months mostly coinciding with the rainy-season, we elected to take the average monthly temperature and monthly precipitation across the six-months included in the study. Although previous studies have shown that lagging precipitation and temperature can improve predictions of malaria transmission in some cases, we felt that we were unable to lag our weather covariates without introducing spatial confounding.[25–29] As a result, for each cluster in a given study-period month, we first took the average amount of precipitation or daytime temperature among all raster squares within 2 km or 10 km radius of the cluster. The 2 versus 10 km boundary depended on the cluster's designation as urban or rural designation, respectively. This approximates the offsets of geographical coordinates applied by

the DHS for each cluster.[10,30] We then aggregated these catchment-area averages for each month into a final study period average. Among the 489 clusters considered, four urban clusters (200, 225, 271, 419) had missing values for temperature and/or precipitation. For these four clusters, the radius was extended to 6km and precipitation and temperature means were calculated as described above.

Correlations among risk factors were evaluated using the Szekely-Rizzo-Bakirov distance correlation with the `energy` R package.[31–33] Based on the covariate-pairwise correlations, we determined that covariate collinearity was manageable and no covariates needed to be excluded from the analysis (Supplementary Figure 4).



Supplementary Figure 4 - Covariate Collinearity: The correlation between each pair of covariates were explored for potential bias due to extreme collinearity. Although there were strong correlations that were consistent with *a priori* expectations (e.g. wealth and urbanicity, temperature and altitude), these correlations did not appear to be completely dependent. As a result, all covariates were kept in the analysis.

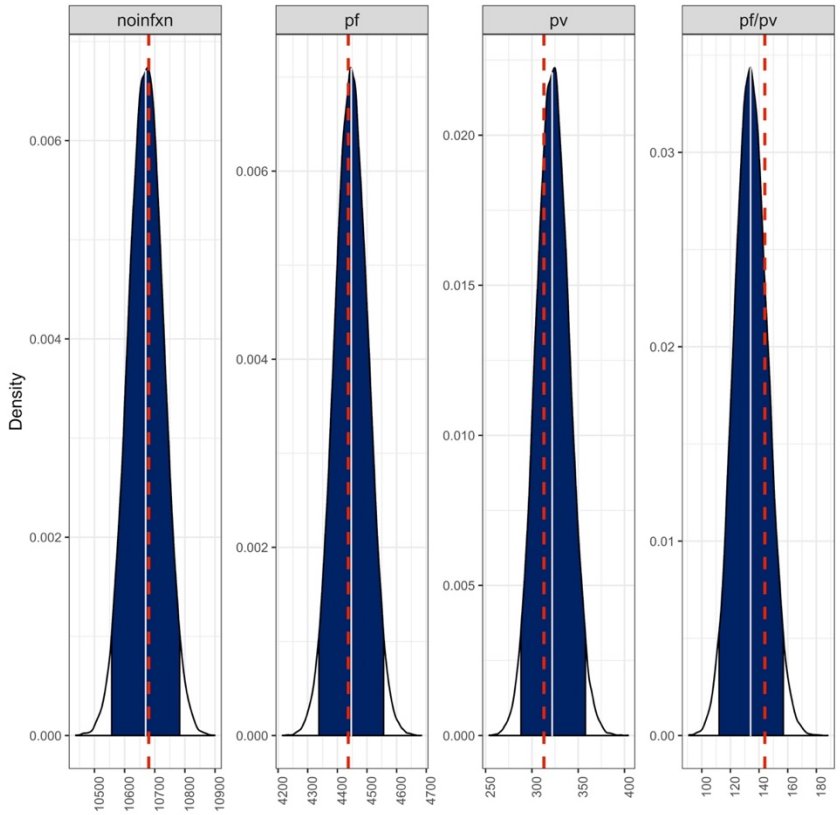
Species Interactions

Interactions between *P. vivax* and *P. falciparum* were examined using an extended version of the independent acquisition of infection model put forth by Akala & Watson *et al.* 2019 to account for individuals that were not infected but still considered in the study population. As in the

previous model, we used the observed frequency of each parasite species to fit the expected frequencies of mono-species and co-species infections using a multinomial likelihood. An additional category -- uninfected -- was added as a parameter to the multinomial model to account for the case when no successful infectious bites occurred. As a result, the unobserved sequence of species, Y that can be passed to a host is now modeled as:

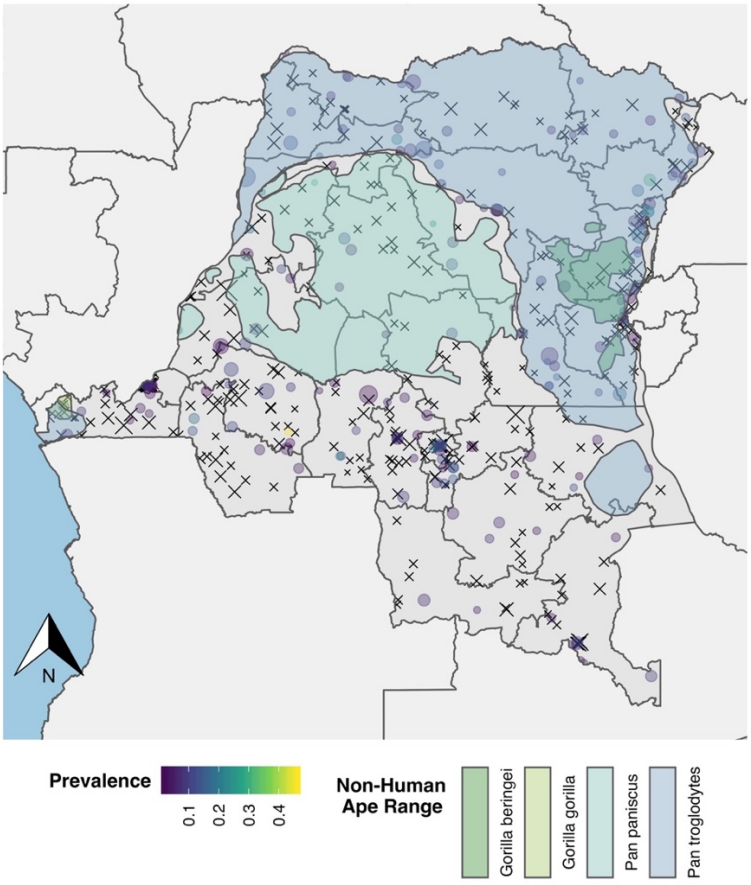
$$Y = \begin{cases} y_1, y_2, , y_k \in S, & \text{if } k > 0 \\ 0, & \text{if } k = 0 \end{cases}$$

Where S was previously defined as the set of *Plasmodium* species of interest and k as the number of infectious bites a host received.[34] Otherwise, the model was unchanged. For the *P. vivax*-*P. falciparum* model, we considered μ as a Poisson distribution and drew 50,000 bootstrap iterations to form the expected infection compositions. Expected infection compositions were then compared against the observed mono- and co-infection data. Overall, mono-infection and co-infection compositions were consistent with the expectation of independent acquisition of parasites, as the observed data fell within the simulated data (Supplementary Figure 5).



Supplementary Figure 5 - Composition of *P. vivax* and *P. falciparum* Co-infections: The expected versus observed composition of *P. vivax* and *P. falciparum* infections were explored using a multinomial likelihood model. The plot shows the expected distribution for individuals without infection ("noinfxn"), *P. falciparum* infections ("pf"), *P. vivax* infections ("pv"), and *P. falciparum*- *P. falciparum* coinfections ("pf/pv"). The blue shading indicates the 95% bootstrapped interval and the red-dotted line indicates the observed number of cases for each infection category. Overall, the observed data is very consistent with the simulated data.

Interactions between NHA territories and *P. vivax* prevalence were assessed using a permutation test with 10,000 iterations. Null distributions for the permutation test were calculated by drawing n_{ape} clusters at random, where n_{ape} was the number of 2013-2014 DRC DHS clusters that overlapped with NHA territories. We then calculated the prevalence of *P. vivax* infections among the selected clusters. We considered NHA territories for (1) *Pan troglodytes* and *Gorilla sp.* and (2) *Pan troglodytes*, *Pan paniscus*, and *Gorilla sp.*, separately, as *P. paniscus* (bonobos) have only recently been shown to harbor *P. vivax*-like parasites at a single field-site (TL2).[35] In contrast, *Pan troglodytes* (chimpanzees) and *Gorilla sp.* have previously been shown to harbor *P. vivax*-like parasites at various prevalences across the DRC.[36] From the permutation tests, NHA territories and *P. vivax* prevalence were not associated (p = 0.32 and p = 0.30, respectively). This lack of an association is also evident when visualizing a map of NHA territories and cluster level *P.* vivax prevalences (Supplementary Figure 6).



Supplementary Figure 6 - *P. vivax* and Non-Human Ape Distributions: Overall, *P. vivax* prevalence did not appear to be associated with non-human ape (NHA) habitat distribution. This lack of a *P. vivax* - NHA association was recapitulated with permutation testing. Clusters with *P. vivax* infections are shaded on a purple-yellow spectrum with respect to the cluster-level prevalence. Clusters without *P. vivax* infections are indicated by black X-ticks. Finally, the distribution of each non-human ape habitat is indicated in shades of green for the *Gorilla* genus and blue for the *Pan* genus.

Inverse Probability Weights and Prevalence Odds Ratios

The average effect of each risk-factor, A, on our binary outcome of interest Y (i.e. malaria infection), was estimated using marginal structural models (MSMs): $g(P(Y|A=a)) = \beta_0 + \beta_1 a$, where $g(\cdot)$ is a logit link for our prevalence odds ratio effect estimates.[37–40] For each MSM, we adjusted for confounders, L, using inverse probability weights (IPWs).[37–40] IPWs were modeled as $w_i = \frac{1}{f_{A|L}(A|L)}$ for each individual, i in the study population, N. Each weight was stabilized by the marginal mean of the risk factor, such that final weights were: $w_i = \frac{f(A)}{f_{A|L}(A|L)} \text{for} i \in N$. In the case of a binary risk factor, $f_{A|L}(A|L)$ was a probability mass function with each level of A representing the predictive probability of receiving a risk factor given a sequence of confounders. Similar, in the case of a continuous treatment, $f_{A|L}(A|L)$ was a probability density function with each level of A representing the predictive probability of receiving a dose of the risk factor given a sequence of confounders. In the continuous setting, we assumed that $f(A)$ and $f_{A|L}(A|L)$ followed normal distributions and could be estimated with a standard normal density.[37,39,41]

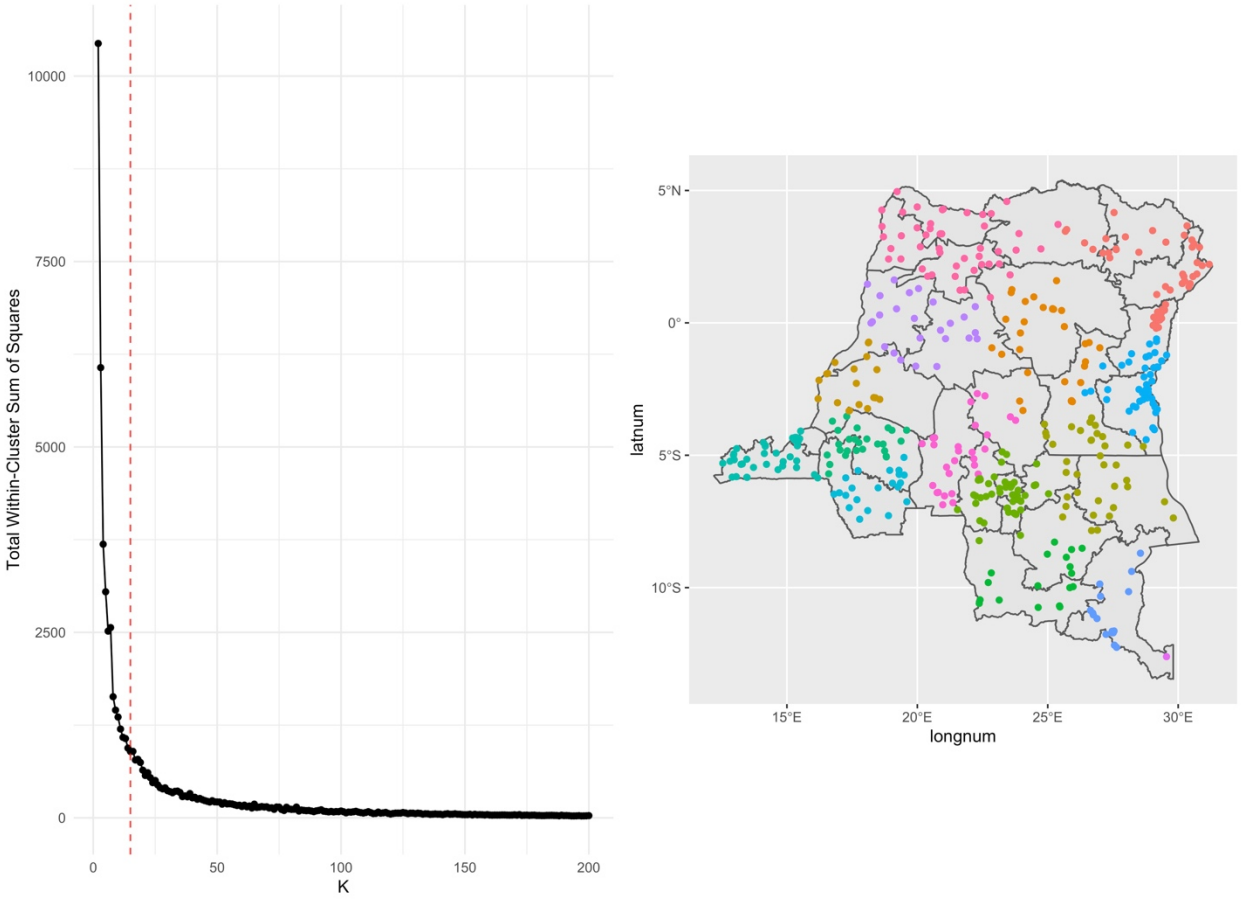
IPWs were calculated using the super learner algorithm with spatial cross-validation.[42–45] We used a diverse set of candidate algorithms, as the super learner is expected to asymptomatically outperform any individual candidate algorithm as the number of candidate algorithms becomes polynomial in sample size (Supplementary Table 4).[42–44] In some cases, if IPWs appeared to be unstable, we limited the candidate algorithm library to either logistic or linear regression, depending on the outcome type (Supplementary Table 5). We assumed that a single iteration of the super learner algorithm was adequate to predict the IPWs.

Base Learner	R-package, Function	Relevant Hyperparameters	Justification
Generalized Linear Regression*	stat, lm/glm [46]	-	-
Cross-Validated L1/L2 Regularized Regression (x3)	glmnet, cvglmnet [47]	α: 1 α: 0.5 α: 0	Shrinkage of covariates based on fit
Boosted Generalized Additive Modeling	mboost, gamboost [48]	-	Non-linearity in covariates
K-Nearest Neighbor	kknn, kknn [49]	k: 7 Kernel: optimal	Interactions, Non-linearity in Covariates
Single Vector Machines	e1071, svm [50]	Cost: 1 Kernel: radial	Interactions, Non-linearity in Covariates
Neural Net	nnet, nnet [51]	Hidden Layers: 1 Units in Hidden Layer: 3	Interactions, Non-linearity in Covariates
Random Forest	ranger, ranger [52]	Number of Trees: 500 Variables at Node split: √p	Interactions, Non-linearity in Covariates

Supplementary Table 4 - Base Learners used in the Super Learner Algorithm: Various base learners were inputted into the super learner algorithm. The super learner algorithm is an ensemble based method that optimizes

the predictions of base learners using a loss-based approach that minimizes the prediction error. A diverse suite of base learners was selected to account for various non-linear effects as well as interactions among covariates.

Folds for cross-validation were based on K-means clustering of geographical coordinates to account for potential spatial autocorrelation among observations.[45] We selected a K of 15, as it was the inflection point that appeared to minimize the within-cluster sum of squares while avoiding overfitting (Supplementary Figure 7).

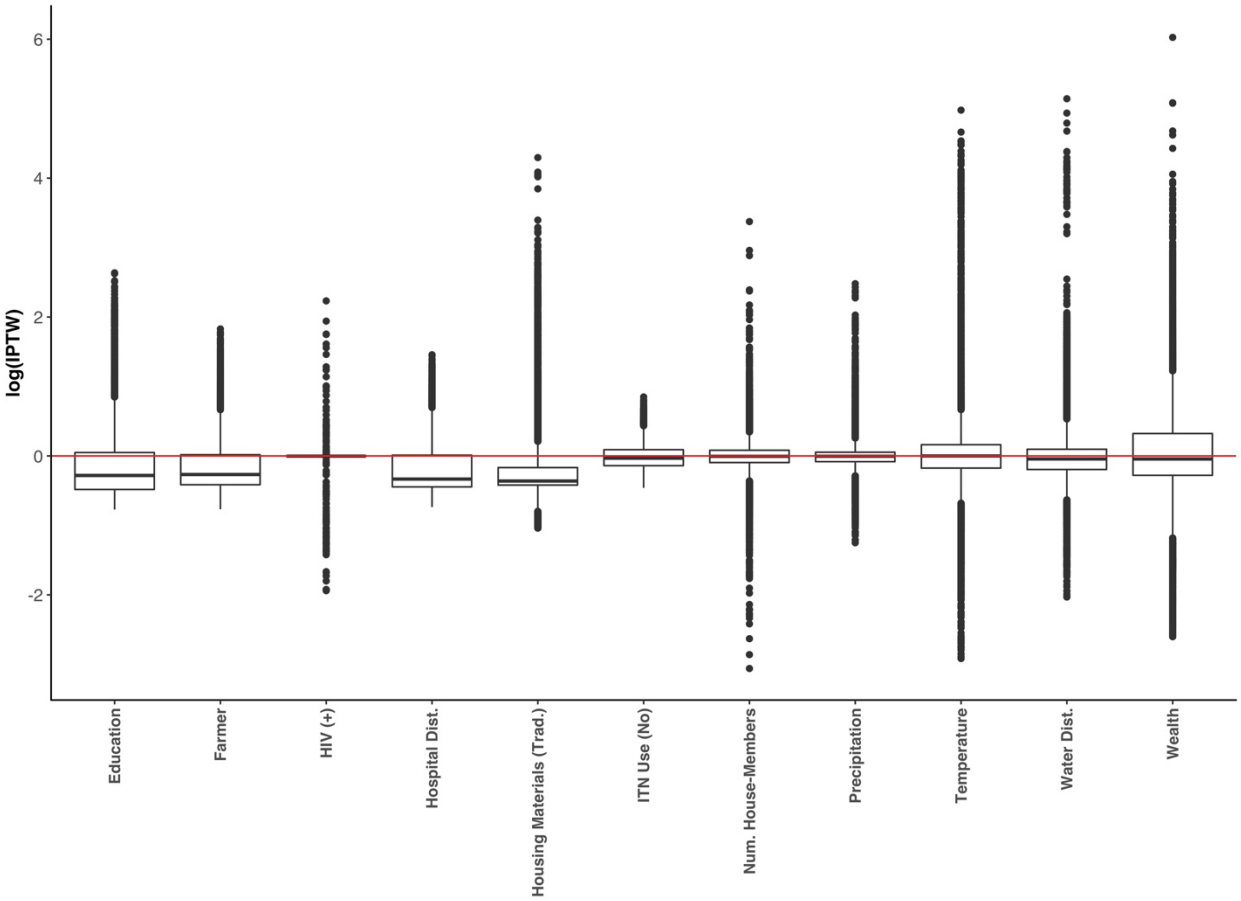


Supplementary Figure 7 - Spatial Cross-Validation K-Clusters: The DRC was partitioned into K-clusters for spatial cross-validation. Based on the geographical K-means total within-cluster sum-of-squares, fifteen clusters appeared to be a reasonable inflection point that did not overfit the data but still captured natural geographic partitions in the DRC (left). The fifteen partitions are mapped to show the geographical partition (right).

All machine-learning models were built and analyzed using the `mlr` package, which provides a machine-learning infrastructure within the R-environment.[53] The super learner algorithm was selected for IPW calculations to account for issues of functional form and non-linearity that can bias predictions.[54]. For each risk factor, we considered all descendants and ancestors of the risk factor and the outcome that were not on the causal pathway as predictors in the IPW-model to account for any "backdoor" paths not considered in our DAG, (i.e. the IPW adjustment set).[37] For risk-factors that were unconfounded in expectation (i.e. biological sex,

age, urbanicity, and altitude), no adjustment set was considered (Supplementary Figure 3). Weights were incorporated with the R `survey` package and base R functions.[55]

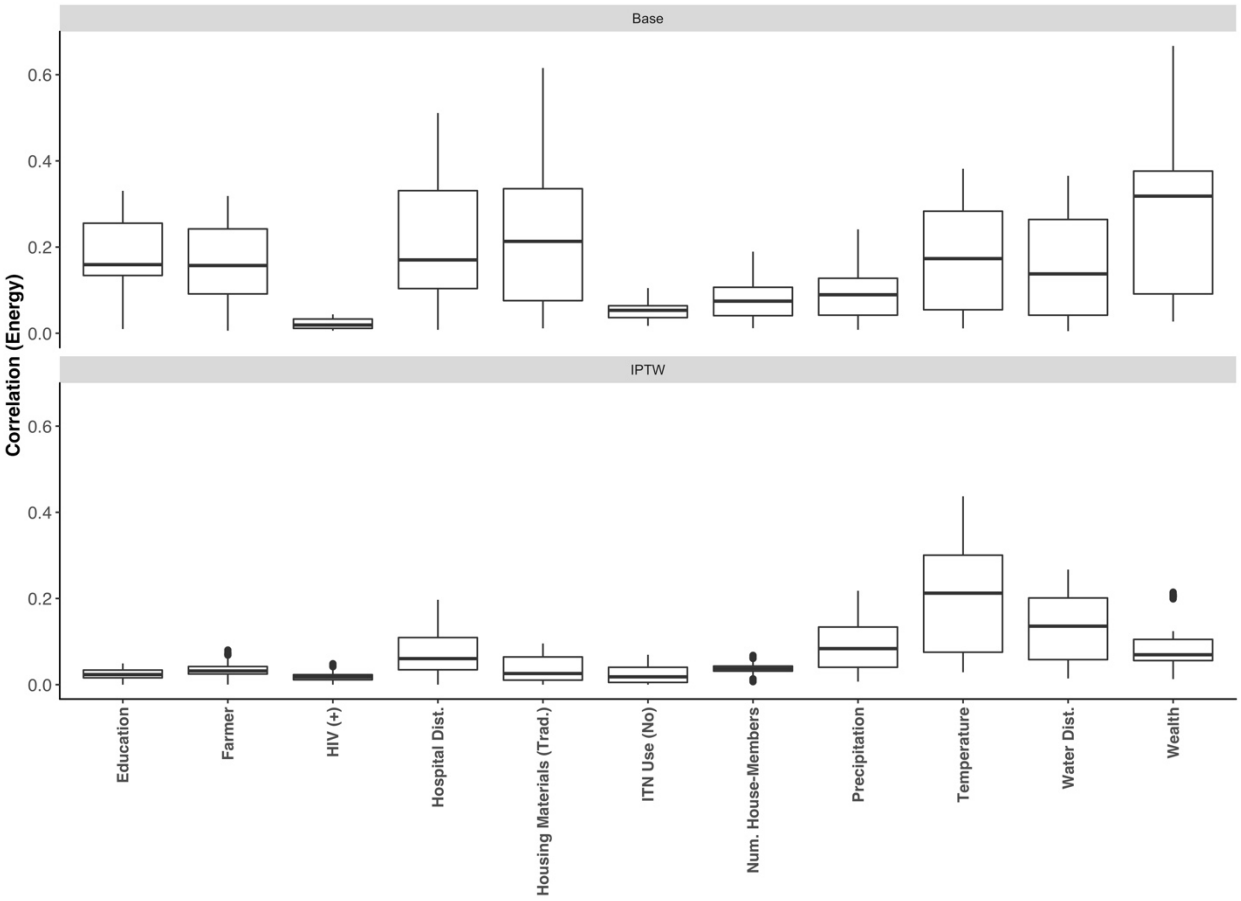
Overall stability of the IPWs were assessed visually and were determined to have log-transformed standard normal distribution (Supplementary Figure 8). IPW distributions that are not definitively centered may suffer from lingering issues of structural positivity or may be correctly identifying multimodal distributions in risk-factor distributions.



Supplementary Figure 8 - Distribution of Inverse Probability Weights: For each covariate, the distribution of weights for the 15,574 individuals included in the study are shown. Distributions have been log-transformed and appear to be approximately normally distributed. *Abbreviations*: Hospital Dist. – Distance to hospital, Trad. – traditional, ITN – insecticide treated net, Num. – number, Water Dist – Distance to water.

The effects of the IPW on baseline risk-factor associations (i.e. putative confounding) were assessed using Szekely-Rizzo-Bakirov distance correlations for each risk-factor pair.[31–33] Given that a weight option is not specified in the Szekely-Rizzo-Bakirov distance correlation calculation, we applied our IPWs by sampling observations according to their IPWs. To account for variability in sampling, we created 100 IPW-pseudopopulations for each risk-factor pair. The

distribution of pairwise distance correlations for the risk factors was then plotted and compared with no weights applied and with IPWs applied (Supplementary Figure 9).



Supplementary Figure 9 - Correlation among Covariates at Baseline and After Application of Inverse Probability Weights: A classic measure of confounding is baseline correlations among covariates, or the unequal distribution of covariates among different treatment classes. Shown for each covariate are the measures of pairwise covariate correlation at baseline (top) and after inverse probability weights (IPWs) have been considered (bottom). Baseline covariates show a large degree of correlation -- potentially indicating confounding -- while, for the most part, covariates with IPWs applied show a considerable reduction in pairwise covariate correlations (mean fold-reduction: 3.14, range: 0.85 - 7.63). Interestingly, temperature appeared to still have somewhat high pairwise correlations even after applying IPWs. *Abbreviations*: Hospital Dist. – Distance to hospital, Trad. – traditional, ITN – insecticide treated net, Num. – number, Water Dist – Distance to water.

Covariate	Cross-Validated Risk Coefficient	Base Learner
Precipitation	1	Simple Linear Regression
Temperature	0.16	Simple Linear Regression
Temperature	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Temperature	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Temperature	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Temperature	0.19	Support Vector Machines (libsvm)
Temperature	0.12	K-Nearest-Neighbor regression
Temperature	0.51	Gradient Boosting with Smooth Components
Temperature	0.03	Neural Network
Temperature	0	Random Forests
Water Dist.	0.10	Simple Linear Regression
Water Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Water Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Water Dist.	0.31	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Water Dist.	0	Support Vector Machines (libsvm)
Water Dist.	0	K-Nearest-Neighbor regression
Water Dist.	0	Gradient Boosting with Smooth Components
Water Dist.	0.20	Neural Network
Water Dist.	0.39	Random Forests
HIV (+)	1	Logistic Regression
Farmer	0	Logistic Regression

Farmer	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Farmer	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Farmer	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Farmer	0.56	Gradient boosting with smooth components
Farmer	0.23	Support Vector Machines (libsvm)
Farmer	0.01	k-Nearest Neighbor
Farmer	0.18	Neural Network
Farmer	0.02	Random Forests
Wealth	0	Simple Linear Regression
Wealth	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Wealth	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Wealth	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Wealth	0.02	Support Vector Machines (libsvm)
Wealth	0.41	K-Nearest-Neighbor regression
Wealth	0.57	Gradient Boosting with Smooth Components
Wealth	0	Neural Network
Wealth	0	Random Forests
Education	1	Logistic Regression
Housing Materials (Trad.)	1	Logistic Regression
ITN Use (No)	1	Logistic Regression
Hospital Dist.	0	Logistic Regression
Hospital Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)

Hospital Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Hospital Dist.	0.98	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Hospital Dist.	0	Gradient boosting with smooth components
Hospital Dist.	0	Support Vector Machines (libsvm)
Hospital Dist.	0.02	k-Nearest Neighbor
Hospital Dist.	0	Neural Network
Hospital Dist.	0	Random Forests

Supplementary Table 5 - Cross-Validated Risk and Contribution of Base Learners for each Covariate: Given that the super learner algorithm optimizes the contribution of individual base learners, not all base learners are included in the final predictions for each covariate. In some instances, super learner predictions resulted in unstable weights. As a result, we culled the base learner library to either a linear or logistic regression algorithm for continuous and dichotomous covariates, respectively (indicated by a 1 in the Cross-Validated Risk Coefficient column). *Abbreviations*: Hospital Dist. – Distance to hospital, Trad. – traditional, ITN – insecticide treated net, Num. – number, Water Dist – Distance to water.

Risk Factor	Species	IPTW-pOR	IPTW-pOR, L95	IPTW-pOR, U95	pOR	pOR, L95	pOR, U95
Age	Pv	0.97	0.87	1.07	0.97	0.87	1.07
Altitude	Pv	1.13	0.88	1.45	1.13	0.89	1.44
Education (Lower)	Pv	0.91	0.64	1.3	0.99	0.74	1.34
Farmer	Pv	1.42	1.08	1.88	1.32	1	1.75
HIV (+)	Pv	0.93	0.33	2.67	1.86	0.76	4.54
Hospital Dist.	Pv	0.86	0.53	1.4	0.86	0.53	1.38
Housing Materials (Trad.)	Pv	1.12	0.62	2.04	1	0.64	1.57
ITN Use (No)	Pv	0.76	0.55	1.04	0.8	0.58	1.09

Precipitation	Pv	0.79	0.63	0.99	0.78	0.63	0.97
Sex (Male)	Pv	1.17	0.89	1.53	1.17	0.89	1.54
Temperature	Pv	0.83	0.62	1.11	0.78	0.62	0.97
Urbanicity (Rur.)	Pv	1.13	0.7	1.83	1.13	0.7	1.82
Water Dist.	Pv	1.19	0.93	1.52	0.97	0.79	1.19
Wealth	Pv	1.12	0.78	1.59	0.93	0.81	1.07
Age	Pf	0.81	0.77	0.86	0.81	0.77	0.86
Altitude	Pf	0.73	0.65	0.82	0.73	0.66	0.8
Education (Lower)	Pf	1.44	1.25	1.67	1.18	1.02	1.35
Farmer	Pf	1.03	0.9	1.18	1.08	0.94	1.24
HIV (+)	Pf	0.54	0.18	1.58	0.5	0.26	0.93
Hospital Dist.	Pf	1.15	0.89	1.48	1.37	1.1	1.7
Housing Materials (Trad.)	Pf	1.25	0.98	1.61	1.84	1.54	2.19
ITN Use (No)	Pf	1.23	1.07	1.42	1.27	1.11	1.45
Precipitation	Pf	0.96	0.83	1.12	0.99	0.87	1.12
Sex (Male)	Pf	1.31	1.2	1.43	1.31	1.2	1.43
Temperature	Pf	1.41	1.05	1.9	1.07	0.97	1.19
Urbanicity (Rur.)	Pf	0.7	0.54	0.89	0.7	0.56	0.86
Water Dist.	Pf	0.87	0.77	0.99	1.12	0.99	1.28
Wealth	Pf	0.82	0.73	0.92	0.75	0.69	0.81

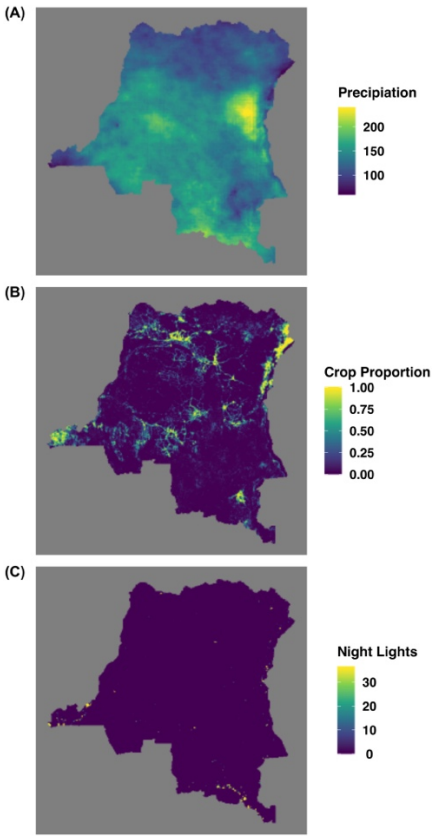
Supplementary Table 6 - Inverse Probability Weight (IPW) Adjusted and Unadjusted Prevalence Odds Ratios for the Malaria Risk Factors: Inverse probability weight (IPW) adjusted and unadjusted prevalences odd ratios (pOR) risk factor effect estimates for *P. vivax* (Pv) and *P. falciparum* (Pf) are provided with corresponding 95% confidence intervals. IPW adjustments were performed using the super learner algorithm. Unadjusted estimates are modeled using generalized estimating equations with a logit-link and binomial variance accounting for the DHS

sample-weights. These bivariate association models are essentially two-by-two tables weighted for the 2013-2014 Demographic Health Survey in the Democratic Republic of the Congo sampling scheme. In instances where the adjusted and unadjusted estimates are the same (age, biological sex, urbanicity, and altitude), the risk factor was expected to be unconfounded at baseline and IPWs were not considered (**Supplementary Figure 3**). *Abbreviations*: Hospital Dist. – Distance to hospital, Trad. – traditional, ITN – insecticide treated net, Rur. - rural, Num. – number, Water Dist – Distance to water.

Spatial and Raster Feature Engineering

In order to incorporate the risk factor covariate information into our spatial models, we downloaded spatial raster data for significant risk factors identified by the MSMs. The precipitation raster was used from above, with the surface consisting of mean values over the study period. To account for the risk factor associated with farming, we downloaded a raster of light intensity and land coverage for the DRC. Specifically, we used the 2015 annual night light composite vcm-orm-ntl version raster (https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html, accessed Nov 8, 2019), which provides an average night-light intensity for each point in the DRC at a 15 arcsecond resolution.[56,57] In addition, the vcm-orm-ntl version has been pre-processed to exclude outliers and spurious measurements due to fires or cloud coverage.[56,57] The 2015 annual night light composite raster was selected as rasters for 2013 and 2014 were not available. Land coverage in the DRC was accessed through the Land Cover Climate Change Initiative (CCI) Climate Research Data Package from the European Space Agency Climate Change Initiative, which provides yearly land coverage maps at 300 x 300 meter resolution for 1992-2015 (https://maps.elie.ucl.ac.be/CCI/, accessed Nov 8, 2019). Specifically, we used the 2013 land coverage raster and reclassified raster points as a binary of cropland or not-cropland based on the CCI classifications (values 10, 20, 30, 40, Yes; all others, No; Supplementary Figure 10). From these raster surfaces, we then aggregated raster points to fit within the DHS cluster design and DHS province boundaries. Specifically, for each cluster and covariate of interest, we took the mean value from all raster squares within a 2 km or 10 km radius with respect to the cluster urban/rural designation.[10,30] For each province, all raster cells within the province boundary were aggregated and summarized as a mean value.

As described above, precipitation values were standardized. Similarly, cropland proportion was transformed onto the real-line using a logit-transformation and was then standardized. Given that most points in the DRC had no measured light-intensity throughout the year, night-light standardization was performed under a zero-truncated framework (i.e. standardization did not include zeros). As above, standardization was performed in favor of model stability.



Supplementary Figure 10 - Spatial Raster Covariates: Spatial covariates that were associated with *P. vivax* infection by the risk factor analysis were included in the spatial prediction prevalence models and included: precipitation (A) and farming. Farming was captured through the proportion of crops (B) at each raster cell as well as raster cell night light intensity (C) across the DRC.

Bayesian Mixed Spatial Models and Predictions

Prevalence maps were fit as mixed generalized linear models with spatially correlated random effects in a Bayesian framework. We modeled prevalence at two different levels: (1) Province-level using the `CARBayes` R-package and (2) Cluster-level using the `PrevMap` R-package.[58,59] DHS sampling weights were accounted for by rounding the number of cases, Y_i, to the nearest whole individual in order to conform with the binomial error distribution of our model. For the province-level models, there are i total survey regions, such that $i \in \geq \mathbb{Z}_1$, and survey regions are defined as non-overlapping areal units with defined boundaries: $A = a_1, \ldots, a_i$. Risk factors that were identified as significant were included as linear predictors, β. As a result, the model was specified as:

$$Y_i|n_i, p_i \sim \text{Binomial}(n_i, p_i)$$
$$logit(p_i) \sim \beta X_i^\intercal + S(a_i) + Z_{a_i}$$

Following Lee 2017, for the province-level model, the spatial ($S(a_i)$) and non-spatial (Z_{a_i}) random effects were modeled using a random effect, ϕ with the conditional-autoregressive prior proposed in Leroux *et al.* 2000 (hereafter referred to as the Leroux CAR model). Specifically,

$$\beta \sim \text{MultivariateNormal}(\mu_\beta,\ \Sigma_\beta)$$
$$\phi_k|\phi_{-k}, W, \tau^2, \rho \sim \text{Normal}(\frac{\rho\sum_{i=1}^{K} w_{ki}\phi_i}{\rho\sum_{i=1}^{K} w_{ki}+1-\rho},\ \frac{\tau^2}{\rho\sum_{i=1}^{K} w_{ki}+1-\rho})$$
$$\tau^2 \sim \text{InverseGamma}(1,\ 0.01)$$
$$\rho \sim Uniform(0,\ 1)$$

The adjacency matrix, W, was a simple neighborhood matrix, where border sharing was indicated as a binary.[60] Models with the ρ parameter fixed at one assume complete spatial autocorrelation among the random effects (i.e. the Intrinsic CAR or Besag model), while models the ρ parameter fixed at zero assume independence.[61–63] By allowing ρ to vary under the model, as specified above, we can fit this spatial autocorrelation process.[58,61] Finally, we set the multivariate Gaussian mean prior as a vector of zeros and the diagonal elements of the covariance matrix, σ_β, to 50,000.[58]

For the cluster model, the survey region is the DHS second-level enumeration area, which is a collection of households aggregated at a single set of GPS coordinates (i.e. clusters).[10] In total, there are there are j total clusters, where $j \in \geq \mathbb{Z}_1$ and clusters (i.e. sampling locations) are indexed as: $C = c_1, \ldots, c_j$. As a result, the model was specified as:

$$logit(p_i) \sim \beta X_i^\intercal + S(c_i) + Z_{c_i}$$

Following the model presented in Giorgi and Diggle 2017, the spatial random effect, S_i, was modeled as a stationary isotropic Gaussian process with variance σ^2 and a Matérn covariance function, $\rho(d; \phi, \kappa)$. Here, d is the distance between any two clusters, $c_i,\ c_{\bar{i}}$. Based on an exploratory analysis of the κ that maximized the log-likelihood of our logit-transformed prevalence data, we fixed κ at 1. The remainder of the model was specified using diffuse priors:

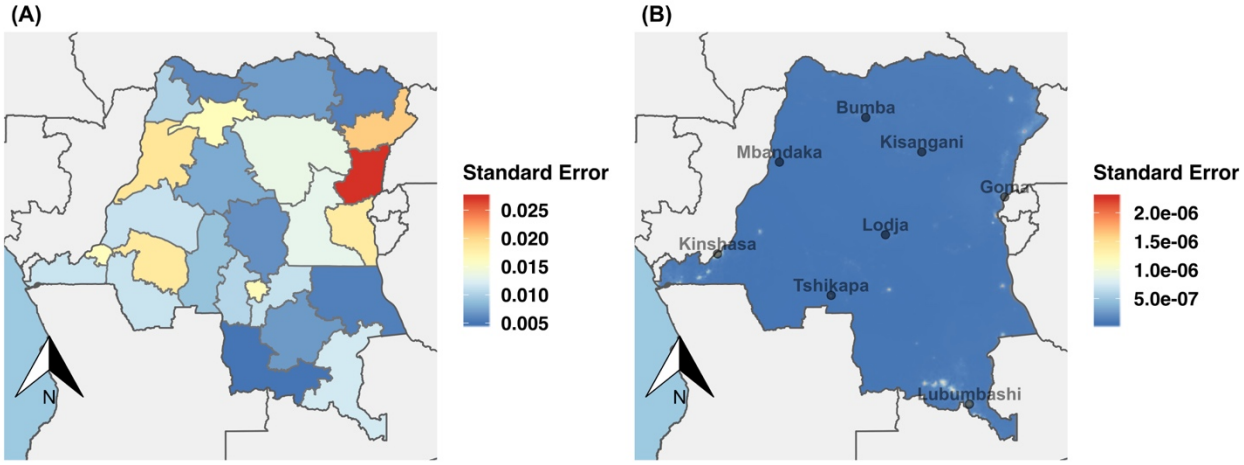
$$\beta|\sigma^2 \sim \text{Normal}(0, \sigma^2)$$
$$log(\tau^2) \sim \text{Normal}(0,\ 25)$$
$$\phi \sim \text{Uniform}(0,\ 50)$$
$$log(\sigma^2) \sim \text{Normal}(0,\ 50)$$

Each model was first evaluated with four diagnostic chains using 1,000 burn-in iterations and 10,000 sample iterations. Chains were then visually assessed for convergence and appropriate mixing patterns. A final long chain with 10,000 burn-in iterations and 100,000 sampling iterations was then considered for each model. Chains were again visually assessed and all parameters were required to have an effective sample size of at least 500. The effective sampling size was calculated with the `coda` R package while the highest posterior density interval (HPDI) was calculated with the `HDInterval` package for the `CarBayes` models (the HPDI is provided by `PrevMap`) [64,65].

For the province level, predictions were calculated from the fitted province responses. These posterior responses were aggregated as means and standard errors for each province.

For the cluster level, predictions were made out-of-sample using the fitted covariates under the assumption of a multivariate Gaussian distribution as previously described in Giorgi and Diggle 2017. Covariate observations for predictions were taken from the precipitation, crop-proportion, and night light intensity rasters described above. For the crop-proportion and night light intensity raster, we aggregated the rasters to a 0.05° x 0.05° resolution by taking the mean and sum of raster cells, respectively (a 0.05° x 0.05° resolution was selected as this was the least

precise spatial resolution among the covariates). For each of the prediction sampling locations, the covariate matrix was calculated by taking the mean value for each raster cell within a six km radius (mean of DHS maximum offset).[10,30] Any value in the covariate prediction matrix that exceeded the observed maximum in our fitted covariate matrix was truncated (i.e. the observed maximum for each covariate served as an upper bound among the predictions to avoid extrapolation). Predictions were then calculated for each of the 100,000 sampling iterations and aggregated as means and standard errors. For the sake of computational burden, we subsetted the approximately 160,000 potential prediction sampling locations in the DRC that would need to be estimated at 100,000 sampling iterations (16 billion estimates) to 20,000 randomly selected sampling locations. To map the unsampled sampling locations, we performed local interpolation using inverse distance weighting and an inverse power parameter of two with the R `gstat` package.[66,67]



Supplementary Figure 11 - Spatial Model Standard Errors: The standard errors of the posterior prevalence distribution for the final province-level (left) and cluster-level (right) and model. For the final province-level model, the standard error range was small (range: 4.44×10^{-3}, 2.76×10^{-2}). Standard errors at the province-level appeared to be greatest along the Eastern and Western borders. Similarly, the final cluster-level model exhibited a small standard error range (range: 1.62×10^{-8}, 2.30×10^{-6}). Standard errors were highest where the prevalence estimates were greatest, indicating a degree of uncertainty that coincides with higher covariates values (Supplementary Figure 10).

Level	Model	Parameter	Mean	Median	2.5% HPDI	97.5% HPDI	Effective N	DICg
Province	Intercept	Intercept	-3.58	-3.58	-3.71	-3.45	27,002	-55.65
		τ	0.52	0.46	0.16	1.01	15,797	
		ρ	0.35	0.30	0.00	0.80	11,425	
	Covariate	Intercept	-3.59	-3.59	-3.71	-3.46	11,162	-61.96
		Precip.	0.03	0.03	-0.28	0.37	636	
		Crop Prop.	0.22	0.21	-0.08	0.53	1,014	
		Nightlight	-0.22	-0.22	-0.47	0.04	1,869	
		τ^2	0.46	0.40	0.11	0.96	2,762	
		ρ	0.30	0.24	0.00	0.76	7,585	
Cluster	Intercept	Intercept	-2.66	-2.64	-6.49	1.20	92,453	1,331,975.70
		σ^2	7.57	5.30	0.78	20.60	4,543	
		ϕ	37.21	38.91	19.55	50.00	18,353	
		τ^2	3.16	3.11	2.24	4.21	1,844	
	Covariate	Intercept	-2.66	-2.64	-5.37	-0.11	71,797	157,849.23
		Precip.	-0.04	-0.04	-0.32	0.24	19,276	
		Crop Prop.	0.06	0.06	-0.22	0.34	18,305	
		Nightlight	-0.08	-0.08	-0.55	0.40	19,862	
		σ^2	3.59	2.78	0.64	8.80	6,757	
		ϕ	35.44	36.98	16.82	50.00	13,206	
		τ^2	3.22	3.18	2.23	4.25	1,949	

Supplementary Table 7 - Spatial Model Parameter Estimates and Fits: The posterior mean, posterior median, and 95% highest posterior density interval (HPDI) summary statistics are provided for each parameter with respect to the models evaluated. The fit of each model was calculated using Gelman's deviance information criteria and compared at the province-level and cluster-level, respectively. Overall, the best fitting province-level and cluster-level models included a precipitation, crop, and night light intensity covariate. For reference, the posterior ϕ values for each province are also provided (Supplementary Table 8).

Model	Province	Mean	Median	2.5% HPDI	97.5% HPDI	Effective N
Intercept	Bas-Uele	-0.27	-0.26	-0.88	0.33	49662
	Equateur	0.92	0.92	0.50	1.36	4439
	Haut-Katanga	-0.32	-0.32	-0.79	0.14	22758
	Haut-Lomami	-0.24	-0.23	-0.82	0.33	63915
	Haut-Uele	-0.21	-0.19	-0.93	0.47	52444
	Ituri	1.03	1.03	0.62	1.47	2967
	Kasai	-0.06	-0.05	-0.55	0.42	58547
	Kasai-Central	-0.36	-0.35	-0.83	0.13	46811
	Kasai-Oriental	0.31	0.31	-0.15	0.77	5144
	Kinshasa	-0.80	-0.80	-1.24	-0.37	10242
	Kongo-Central	-0.69	-0.68	-1.24	-0.16	25972
	Kwango	-0.38	-0.37	-0.85	0.07	30992
	Kwilu	0.15	0.15	-0.27	0.57	5581
	Lomami	-0.01	0.00	-0.43	0.41	41269
	Lualaba	-0.50	-0.48	-1.19	0.16	57783
	Mai-Ndombe	-0.26	-0.26	-0.72	0.18	37433
	Maniema	0.34	0.34	-0.08	0.76	30291
	Mongala	0.84	0.83	0.42	1.25	12839
	Nord-Kivu	0.44	0.46	-0.05	0.86	931
	Nord-Ubangi	-0.19	-0.17	-0.87	0.46	61733
	Sankuru	-0.13	-0.12	-0.70	0.43	58822
	Sud-Kivu	-0.06	-0.06	-0.54	0.39	2478
	Sud-Ubangi	0.09	0.09	-0.43	0.59	35362
	Tanganyika	-0.31	-0.30	-0.95	0.28	62905
	Tshopo	0.80	0.80	0.40	1.21	26008
	Tshuapa	-0.13	-0.12	-0.67	0.39	50960
Covariate	Bas-Uele	-0.22	-0.22	-0.98	0.55	1585
	Equateur	1.09	1.08	0.57	1.63	1555
	Haut-Katanga	-0.29	-0.28	-0.74	0.16	18098
	Haut-Lomami	-0.01	-0.01	-0.68	0.72	3428
	Haut-Uele	-0.21	-0.20	-1.01	0.53	3980
	Ituri	0.72	0.72	0.08	1.37	982
	Kasai	-0.16	-0.15	-0.69	0.37	3791
	Kasai-Central	-0.44	-0.43	-0.96	0.06	4439
	Kasai-Oriental	0.21	0.21	-0.21	0.61	4953
	Kinshasa	-0.13	-0.12	-1.23	0.98	1880
	Kongo-Central	-0.85	-0.83	-1.54	-0.20	3048
	Kwango	-0.36	-0.35	-0.85	0.12	5457
	Kwilu	-0.02	-0.02	-0.48	0.44	1822
	Lomami	0.22	0.21	-0.36	0.81	2029
	Lualaba	-0.30	-0.29	-1.04	0.43	7227
	Mai-Ndombe	-0.13	-0.13	-0.63	0.37	5164
	Maniema	0.25	0.25	-0.23	0.73	2520
	Mongala	0.57	0.56	0.05	1.10	2171
	Nord-Kivu	0.26	0.27	-0.28	0.80	801
	Nord-Ubangi	-0.25	-0.24	-1.02	0.50	3034
	Sankuru	-0.20	-0.19	-0.78	0.38	7946
	Sud-Kivu	-0.19	-0.18	-0.84	0.49	786
	Sud-Ubangi	-0.04	-0.04	-0.63	0.56	2770
	Tanganyika	-0.18	-0.17	-0.86	0.47	9500
	Tshopo	0.73	0.72	0.28	1.19	2789
	Tshuapa	-0.08	-0.07	-0.62	0.47	19404

Supplementary Table 8 - Summary of the posterior for each province: The posterior mean, posterior median, and 95% highest posterior density interval (HPDI) was calculated with respect to the province. These are provided for reference.

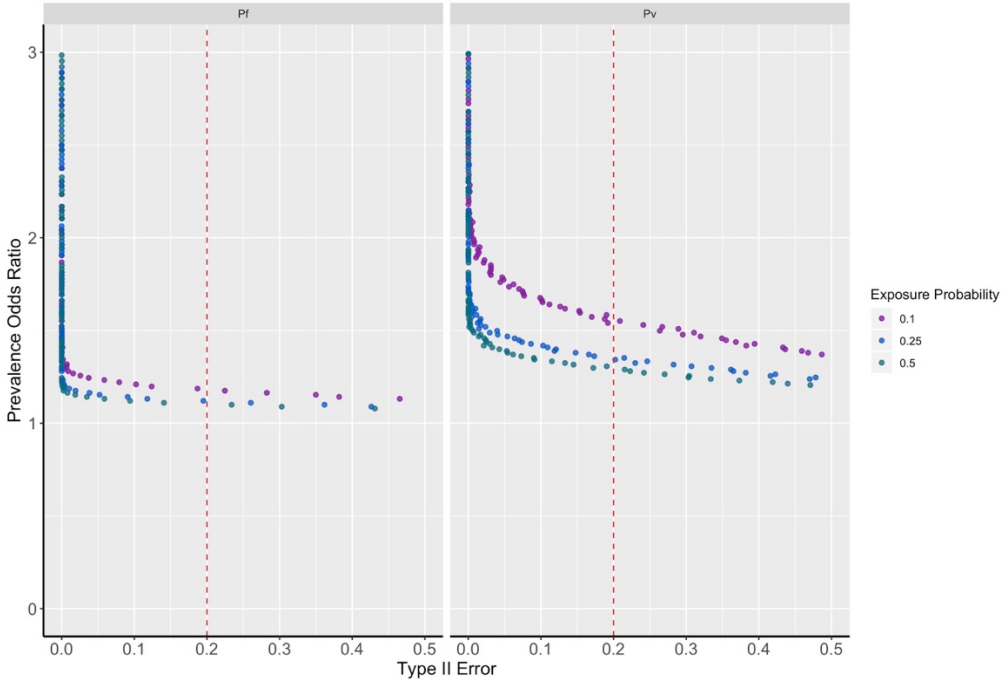
post-hoc Power Calculations

Power calculations were simulated from a population of 15,490 individuals (the weighted N from our study population), where the probability of exposure, $\Pr(E)$ was varied at 10%, 25%, and 50% within the population. For the *P. vivax* models, the overall prevalence of the outcome, O, was set at 3% but was varied in the unexposed group from 0.01 - 3.0% (O_U). In contrast, for *P. falciparum* models, the overall prevalence of the outcome was set at 30% and was varied in the unexposed group from 1.0 - 30.0%. ORs were simulated under the following framework:

$$A_i \sim Bernoulli(\Pr(E)) for i \in 1:N$$
$$O_E = 2 * O - O_U$$

$$D_i = \begin{cases} Bernoulli(\Pr(O_E)), & \text{if } A_1 = 1\ , \\ Bernoulli(\Pr(O_U)), & \text{if}_i = 0\ . \end{cases}$$

As a result, from the exposure status, A_i, and disease status, D_i, we can calculate the simulated OR using the standard generalized linear model function with a logit-link in R. Power was calculated as the number of iterations that the parameter estimate α was less than 0.05 with respect to each OR.



Supplementary Figure 12 - Power Calculations for *P. vivax* and *P. falciparum*: We performed *a posteriori* power calculations to determine the minimum detectable risk factor at varying levels of exposure with 80% power given the prevalence of *P. vivax* and *P. falciparum* in our study. At the lowest exposure probability (lowest expected power), we could detect a harmful prevalence odds ratio of approximately 1.54 and 1.18 for *P. vivax* ("Pv") and *P. falciparum* ("Pf"), respectively.

Population Genetics

Hybrid Selection and Next Generation Sequencing

Samples from the DRC were amplified using the Illustra Genomic Phi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Pittsburgh, PA) and prepared for sequencing using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England BioLabs Inc., Ipswich, MA). Amplified libraries were then enriched using custom MYbaits targeting the *P. vivax* genome (version 3.0; MYcroarray: The Oligo Library Company, Ann Arbor, MI). Enriched genomes were sequenced on MiSeq 150 base-pair paired-end and HiSeq2500 125 base-pair paired-end chemistry (Illumina, San Diego, CA).

Publicly Available Whole Genome Sequences

We downloaded 684 publicly available Illumina paired-end *P. vivax* or *P. vivax*-like whole genome sequences from across the globe from the European Nucleotide Archive (Additional File).[68–81] In addition, we downloaded Illumina single-end sequences for a single isolate that was recovered from a microscopy slide dating to Spain, 1944.[82] *P. cynomologi* Illumina paired-end sequences were downloaded for both the M- and B-strains (Accessions: DRS000258, ERS001838, ERS023609).[83,84]

Alignment, Quality Control, and Variant Discovery

Reads were aligned to the *P. vivax* P01 reference genome (ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/CURRENT/PvivaxP01.genome.fasta.gz) with `bwa mem` (v0.7) after undergoing adaptor-trimming with `cutadapt` (v1.16).[85] Alignments were then deduplicated and mate-tags were added using `samblaster` (v0.1.24). The quality of the alignments were assessed using the Genome Analysis Toolkit (GATK) `CallableLoci` tool (v3.8-0). We defined a "callable" loci as sites with greater than or equal to five high-quality reads (MQ >= 10, BQ >= 20). Upon inspection of the DRC isolates, we found that genomic coverage was sparse and only the mitochondria passed quality-thresholds. As a result, all further analyses were limited to the mitochondria. We then performed short variant discovery using GATK `HaplotypeCaller` followed by joint genotyping across all *P. vivax* samples with GATK `GenotypeGVCFs` (v4.0.3).[86]

Variant Filtering and Consensus Haplotypes

Samples were excluded from downstream processing if less than 95% of their mitochondrial genome was callable (24/685 samples). Loci were then filtered using the GATK "hard filtering" approach, following previously established guidelines for both single nucleotide variants (SNVs) and insertion-deletions (INDELs).[87] Specifically, we filtered loci with a quality-depth of less than two (QD < 2), position bias (ReadPosRankSum < -8.0 for SNV, ReadPosRankSum < -20.0 for INDELs), strand bias (FS > 60 for SNV, FS > 200 for INDEL, SOR > 3 for SNV, SOR > 10 for INDELs), and low mapping-quality (MQ < 35, MQSR < -12.5) using the GATK `VariantFiltration` and `SelectVariants` tools (v4.0.3).

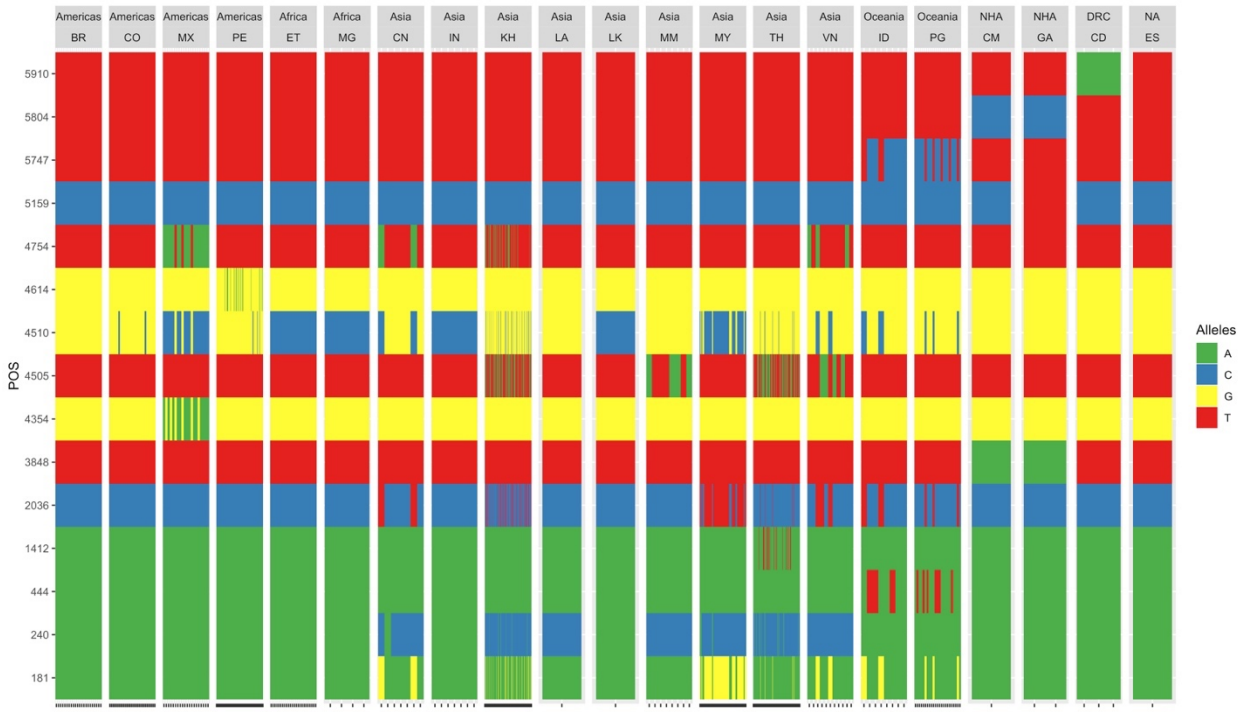
Following hard-filtering, we performed post-processing of loci and samples using the `vcfR` package and other custom scripts (GitHub: IDEELResearch/vcfRmanip).[88] Passed loci were first limited to SNVs and sites that encoded a deletion as an alternative allele were excluded (i.e. `*` in the ALT category). Samples with more than 20% of SNV genotyped as heterozygous were excluded under an assumption of heteroplasmy. We then imputed the genotype of missing loci based on the sample's within-country allele frequency. Two samples that were the only isolate from their country of origin, ERS347479 (Laos) and ERS040109 (Sri Lanka), were combined into Thailand and India for imputation, respectively. Following imputation, heterozygous sites were recoded as the major allele. Finally, we removed alleles within a country if the within-country allele frequency was less than $\frac{n}{n-1} * 0.1$. In a large population, this expression simplifies to removing alleles that are at less than 10% frequency within a country.

From the resulting stringently filtered genotype calls, we created a consensus haplotype for each sample using the P01 mitochondrial sequence as a backbone using the `SeqinR` and `Biostrings` packages.[89,90]

. Two samples -- both a part of the *P. vivax*-like Clade 2 from Gilabert *et al.* 2018 -- were found to have a higher-order of diversity than expected (ERS333076, ERS352725) and were subsequently excluded from further analysis.

In total, 636/685 samples passed quality thresholds and were included in analyses. The Ebro-1944 sample was originally excluded at the callable loci stage (3,148/5,989 bases callable) but was later recovered for visual comparison (Supplementary Figure 13). To identify variants for the Ebro-1944 sample, we used the `mpileup` and `call` (consensus-caller) tools within the `bcftools` suite (*N.B.* joint genotyping was not performed).

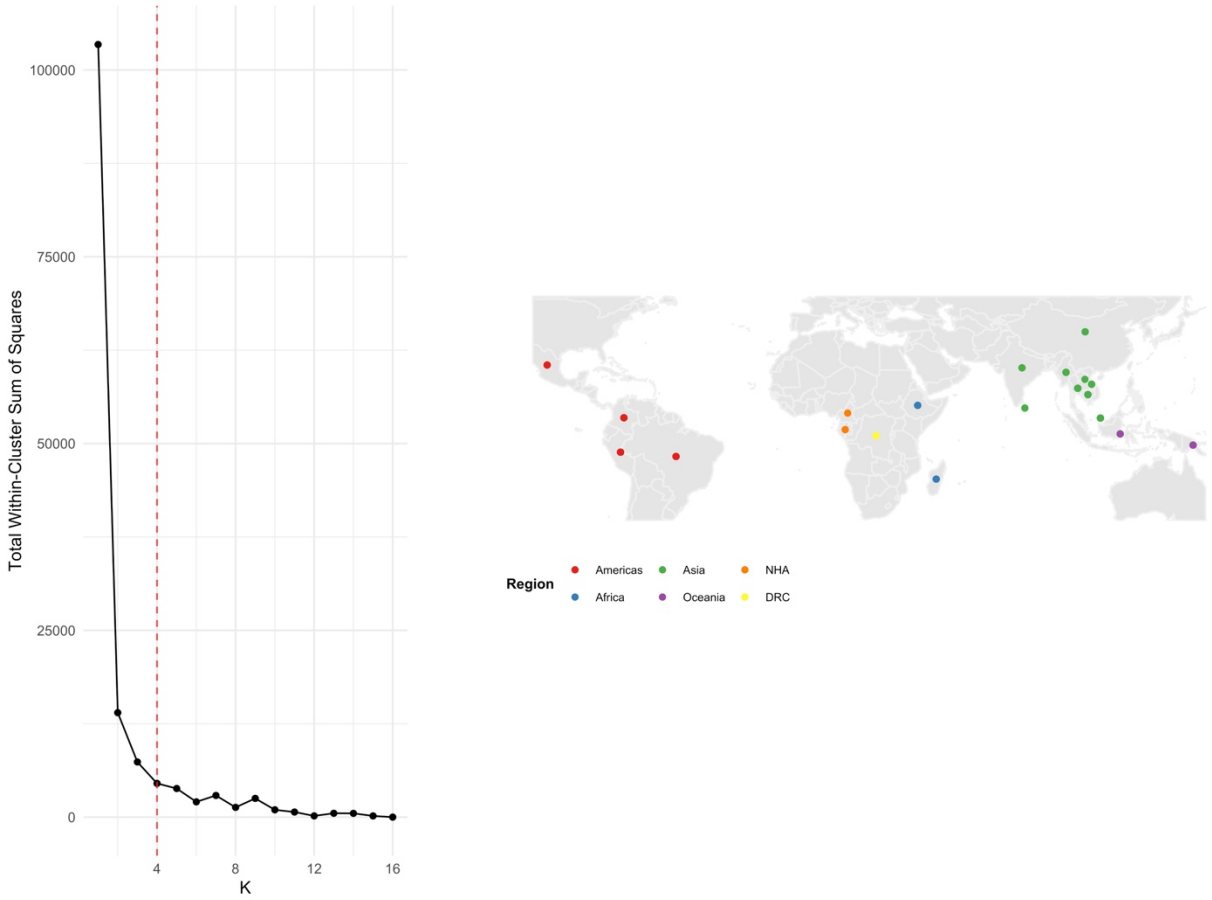
Separately, the *P. cynomolgi* samples also underwent variant discovery, joint genotyping, and hard-filtering as described above. The resulting hard-filtered variants among the three *P. cynomolgi* isolates were then processed by recoding heterozygous alleles as homozygous based on the major allele. Variants were then limited to SNVs and for each variant site, the most common allele among the three isolates was selected. As above, using these consensus SNVs, we created a *P. cynomolgi* consensus haplotype with a *P. vivax* P01 backbone.



Supplementary Figure 13 - Consensus Haplotypes: Haplotypes are shown for each isolate that passed quality-control (QC) threshold with the exception of the sample from Spain (ES) dating to 1944 in the Ebro region (Ebro-1944). As described above, the Ebro-1944 sample did not initially pass QC thresholds but was later recovered for visual comparison. *Abbreviations*: DRC – Democratic Republic of the Congo, NHA – non-human apes, Brazil (BR), Colombia (CO), Mexico (MX), Peru (PE), Spain (ES), China (CN), Indonesia (ID), Cambodia (KH), Laos (LA), Myanmar (MM), Malaysia (MY), Papua New Guinea (PG), Thailand (TH), and Vietnam (VN). India (IN), Sri Lanka (LK), Ethiopia (ET), Madagascar (MG), Democratic Republic of the Congo (CD), Cameroon (CM), and Gabon (GA).

Population Genetic Statistics and Phylogenetics

Isolates were first divided into global regions using geographic K-means clustering. We selected K to be four based on minimizing the within-cluster sum of squares while avoiding overfitting. Samples from the DRC and NHA samples were also designated separate clusters (Supplementary Figure 14).



Supplementary Figure 14 - Spatial Cross-Validation K-Clusters: Countries with *P. vivax* isolates included in the study were partitioned into K-groups for diversity and population differentiation measures. Based on the geographical K-means total within-cluster sum-of-squares, four sub-populations appeared to be a reasonable balance between minimizing the total within-cluster sum of squares while avoiding overfitting the data (left). The DRC samples and non-human ape samples were included as separate populations based on the overall study question and prior assumptions (right).

To explore patterns of diversity among our global isolates, we first measured within-region nucleotide and haplotype diversity using the R-package, `PopGenome` (Supplementary Table 9).[91–94] We then evaluated the degree of population differentiation among parasite using measures of between-region nucleotide and haplotype diversity as well as pairwise measures of both between- and within-regions using Hudson's F_{st} (Supplementary Table 10).[91,92,95,96] Population differentiation was also calculated using a Hamming's distance between consensus haplotypes with the `ape` R-package.[97] Haplotype differences were then mapped and visualized directly for the DRC (Figure 5).

Population	Nucleotide Diversity	Haplotype Diversity
Americas	0.70	0.38
Africa	0	0
Asia	1.80	0.77
Oceania	1.65	0.68
NHA	0.67	0.67
DRC	0	0

Supplementary Table 9 - Within Population Measures of Diversity: For each population, the within-population nucleotide diversity and haplotype diversity was evaluated. Overall, there was little within population diversity among samples from Africa as a whole. This lack of diversity may be an effect of the sample size.

Pop1	Pop2	Between Haplotype Diversity	Between Nucleotide Diversity	Hudson's Fst
Africa	Americas	0.97	1.14	0.81
Asia	Americas	0.70	1.50	0.18
Asia	Africa	0.97	1.74	0.61
Oceania	Americas	0.95	1.93	0.44
Oceania	Africa	1	2.26	0.66
Oceania	Asia	0.95	2.56	0.23
NHA	Americas	1	3.05	0.48
NHA	Africa	1	3.67	0.67
NHA	Asia	1	3.86	0.28
NHA	Oceania	1	4.25	0.32
DRC	Americas	1	1.39	0.81
DRC	Africa	1	2	1
DRC	Asia	1	2.19	0.62
DRC	Oceania	1	2.58	0.66
DRC	NHA	1	3.67	0.67

Global F_{st}	-	-	0.81

Supplementary Table 10 - Between Population Measures of Diversity and Population Structure: Pairwise comparisons were made for each population with respect to genetic diversity and population differentiation. Overall, the DRC differed from samples from the Americas the least. However, based on Hudson's F_{st} this similarity was ancestral and did not represent recent mixing. Instead, the DRC samples appeared to be relatively isolated based on Hudson's F_{st}. Overall lack of haplotype sharing is likely -- in part -- due to small sample sizes.

Evolutionary relationships among the isolates were explored using phylogenetic analysis. We first identified the mutational model that best fit the observed data by comparing the Jukes-Cantor versus the General Time Reversible substitution model (GTR + $\gamma(4)$) using maximum likelihood estimation with the `ape` and `phangorn` R-packages.[97–100] For both substitution models, the tree topology, base frequencies, rate matrix, and gamma rate parameters were simultaneously optimized while finding the maximum likelihood. Model fit was compared using AIC, with the GTR model demonstrating a lower AIC and a better model fit. We then performed 1,000 bootstrap iterations of our phylogenetic tree under the GTR model. The phylogenetic tree with the bootstrapped node support was then plotted using the R-package `ggtree`. Finally, we set *P. cynomologi* as the outgroup to root the tree.

References

1 Deutsch-Feldman M, Brazeau N, Parr J, *et al.* Spatial and epidemiological drivers of P. falciparum malaria among adults in the Democratic Republic of the Congo. *medRxiv* 2020; published online Jan 28. DOI:10.1101/2020.01.28.20018978.

2 Plowe CV, Djimde A, Bouare M, Doumbo O, Wellems TE. Pyrimethamine and proguanil resistance-conferring mutations in Plasmodium falciparum dihydrofolate reductase: polymerase chain reaction methods for surveillance in Africa. *Am J Trop Med Hyg* 1995; **52**: 565–8.

3 Srisutham S, Saralamba N, Malleret B, Rénia L, Dondorp AM, Imwong M. Four human Plasmodium species quantification using droplet digital PCR. *PLoS One* 2017; **12**: e0175771.

4 Snounou G, Singh B. Nested PCR analysis of Plasmodium parasites. *Methods Mol Med* 2002; **72**: 189–203.

5 Mercereau-Puijalon O, Barale J-C, Bischoff E. Three multigene families in Plasmodium parasites: facts and questions. *Int J Parasitol* 2002; **32**: 1323–44.

6 Gruenberg M, Moniz CA, Hofmann NE, *et al.* Plasmodium vivax molecular diagnostics in community surveys: pitfalls and solutions. *Malar J* 2018; **17**: 55.

7 Tanaka M, Takahahi J, Hirayama F, Tani Y. High-resolution melting analysis for genotyping Duffy, Kidd and Diego blood group antigens. *Leg Med* 2011; **13**: 1–6.

8 Tournamille C, Colin Y, Cartron JP, Van Kim CL. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. *Nat Genet* 1995; **10**: 224–8.

9 Ménard D, Barnadas C, Bouchier C, *et al.* Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc Natl Acad Sci U S A* 2010; **107**: 5967–71.

10 Croft TN, Marshall AMJ, Allen CK, Others. Guide to DHS statistics. *Rockville, Maryland, USA: ICF* 2018.

11 API Client and Dataset Management for the Demographic and Health Survey (DHS) Data [R package rdhs version 0.6.3]. https://cran.r-project.org/web/packages/rdhs/index.html (accessed July 31, 2019).

12 Ouma P, Okiro EA, Snow RW. Sub-Saharan Public Hospitals Geo-coded database. 2017. DOI:10.7910/DVN/JTL9VY.

13 Wan, Z., Hook, S., Hulley, G. MYD11C3 MODIS/Aqua Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006 [Data set]. NASA EOSDIS Land Processes DAAC. 2015. https://doi.org/10.5067/MODIS/MYD11C3.006 (accessed Sept 20, 2019).

14 Funk C, Peterson P, Landsfeld M, *et al.* The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data* 2015; **2**: 150066.

15 Luxen D, Vetter C. Real-time routing with OpenStreetMap data. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2011: 513–6.

16 South A. rnaturalearth: World Map Data from Natural Earth. 2017. https://CRAN.R-project.org/package=rnaturalearth.

17 Taylor SM, Messina JP, Hand CC, *et al.* Molecular Malaria Epidemiology: Mapping and Burden Estimates for the Democratic Republic of the Congo, 2007. *PLoS One* 2011; **6**: e16420.

18 Millar J, Psychas P, Abuaku B, *et al.* Detecting local risk factors for residual malaria in northern Ghana using Bayesian model averaging. *Malar J* 2018; **17**: 343.

19 Tusting LS, Bottomley C, Gibson H, *et al.* Housing Improvements and Malaria Risk in Sub-Saharan Africa: A Multi-Country Analysis of Survey Data. *PLoS Med* 2017; **14**: e1002234.

20 Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *Int J Epidemiol* 2016; **45**: 1887–94.

21 Lindsay SW, Jawara M, Mwesigwa J, *et al.* Reduced mosquito survival in metal-roof houses may contribute to a decline in malaria transmission in sub-Saharan Africa. *Sci Rep* 2019; **9**: 7770.

22 Rutstein SO. Steps to constructing the new DHS Wealth Index. *Rockville, MD: ICF International* 2015. https://www.dhsprogram.com/programming/wealth%20index/Steps_to_constructing_the_new_DHS_Wealth_Index.pdf.

23 Karney CFF. Algorithms for geodesics. *J Geodesy* 2013; **87**: 43–55.

24 Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal. 2018; **10**: 439–46.

25 Darkoh EL, Larbi JA, Lawer EA. A Weather-Based Prediction Model of Malaria Prevalence in Amenfi West District, Ghana. *Malar Res Treat* 2017; **2017**: 7820454.

26 Ferrão JL, Mendes JM, Painho M. Modelling the influence of climate on malaria occurrence in Chimoio Municipality, Mozambique. *Parasit Vectors* 2017; **10**: 260.

27 Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White NJ, Kaewkungwal J. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan. *Malar J* 2010; **9**: 251.

28 Nkurunziza H, Gebhardt A, Pilz J. Bayesian modelling of the effect of climate on malaria in Burundi. *Malar J* 2010; **9**: 114.

29 Janko MM, Irish SR, Reich BJ, *et al.* The links between agriculture, Anopheles mosquitoes, and malaria risk in children younger than 5 years in the Democratic Republic of the Congo: a population-based, cross-sectional, spatial study. *The Lancet Planetary Health* 2018; **2**: e74–82.

30 Mayala B, Fish TD, Eitelberg D, Dontamsetti T. The DHS Program Geospatial Covariate Datasets Manual. 2018.

31 Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Stat* 2007; **35**: 2769–94.

32 Székely GJ, Rizzo ML. Brownian distance covariance. *Ann Appl Stat* 2009; **3**: 1236–65.

33 Rizzo M, Szekely G. energy: E-Statistics: Multivariate Inference via the Energy of Data. 2019. https://CRAN.R-project.org/package=energy.

34 Akala HM, Watson O, Mitei KK, *et al.* Longitudinal characterization of Plasmodium inter-species interactions during a period of increasing prevalence of Plasmodium ovale. *medRxiv* 2020; published online Jan 2. DOI:10.1101/2019.12.28.19015941.

35 Liu W, Sherrill-Mix S, Learn GH, *et al.* Wild bonobos host geographically restricted malaria parasites including a putative new Laverania species. *Nat Commun* 2017; **8**: 1635.

36 Liu W, Li Y, Shaw KS, *et al.* African origin of the malaria parasite Plasmodium vivax. *Nat Commun* 2014; **5**: 3346.

37 Hernán MA RJM. Causal Inference. Boca Raton: Chapman & Hall/CRC.

38 Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006; **60**: 578–86.

39 Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–60.

40 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; **168**: 656–64.

41 Zhu Y, Coffman DL, Ghosh D. A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. Journal of Causal Inference. 2015; **3**. DOI:10.1515/jci-2014-0022.

42 van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007; **6**: Article25.

43 Polley EC, van der Laan MJ. Super Learner In Prediction. 2010. https://biostats.bepress.com/ucbbiostat/paper266/ (accessed Oct 6, 2019).

44 Gruber S, Logan RW, Jarrín I, Monge S, Hernán MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat Med* 2015; **34**: 106–17.

45 Brenning A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. 2012: 5372–5.

46 R Core Team. R: A Language and Environment for Statistical Computing. 2019. https://www.R-project.org/.

47 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**: 1.

48 Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. mboost: Model-Based Boosting. 2018. https://CRAN.R-project.org/package=mboost.

49 Schliep K, Hechenbichler K. kknn: Weighted k-Nearest Neighbors. 2016. https://CRAN.R-project.org/package=kknn.

50 Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. 2019. https://CRAN.R-project.org/package=e1071.

51 Venables WN, Ripley BD. Modern Applied Statistics with S. 2002. http://www.stats.ox.ac.uk/pub/MASS4.

52 Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software. 2017; **77**: 1–17.

53 Bischl B, Lang M, Kotthoff L, *et al.* mlr: Machine Learning in R. *J Mach Learn Res* 2016; **17**: 5938–42.

54 Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol* 2015; **181**: 108–19.

55 Lumley T, Others. Analysis of complex survey samples. *J Stat Softw* 2004; **9**: 1–19.

56 Mills S, Weiss S, Liang C. VIIRS day/night band (DNB) stray light characterization and correction. In: Earth Observing Systems XVIII. International Society for Optics and Photonics, 2013: 88661P.

57 Elvidge CD, Baugh K, Zhizhin M, Hsu FC, Ghosh T. VIIRS night-time lights. *Int J Remote Sens* 2017; **38**: 5860–79.

58 Lee D. CARBayes version 4.6: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors. *University of Glasgow, Glasgow* 2017. https://mran.microsoft.com/snapshot/2017-02-

20/web/packages/CARBayes/vignettes/CARBayes.pdf.

59 Giorgi E, Diggle PJ, Others. PrevMap: an R package for prevalence mapping. *J Off Stat* 2017. https://jstatsoft.tr1k.de/article/download/v078i08/v78i08.pdf.

60 Bivand RS, Pebesma E, Gómez-Rubio V. Applied Spatial Data Analysis with R. Springer, New York, NY, 2013.

61 Leroux BG, Lei X, Breslow N. Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. In: Statistical Models in Epidemiology, the Environment, and Clinical Trials. Springer New York, 2000: 179–91.

62 Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 1991; **43**: 1–20.

63 Lee D. CARBayes: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors. https://pdfs.semanticscholar.org/95b1/8642faa6d1e2222e6cf52d99b2d858e6d982.pdf.

64 Meredith M, Kruschke J. HDInterval: Highest (Posterior) Density Intervals. 2018. https://CRAN.R-project.org/package=HDInterval.

65 Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006; **6**: 7–11.

66 Pebesma EJ. Multivariable geostatistics in S: the gstat package. *Comput Geosci* 2004; **30**: 683–91.

67 Gräler B, Pebesma E, Heuvelink G. Spatio-Temporal Interpolation using gstat. The R Journal, 8 (1), 204--218. 2016.

68 Parobek CM, Lin JT, Saunders DL, *et al.* Selective sweep suggests transcriptional regulation may underlie Plasmodium vivax resilience to malaria control measures in Cambodia. *Proc Natl Acad Sci U S A* 2016; **113**: E8096–105.

69 Hupalo DN, Luo Z, Melnikov A, *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat Genet* 2016; **48**: 953–8.

70 Loy DE, Plenderleith LJ, Sundararaman SA, *et al.* Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. *Proc Natl Acad Sci U S A* 2018; **115**: E8450–9.

71 Shen H-M, Chen S-B, Wang Y, Chen J-H. Whole-genome sequencing of a Plasmodium vivax isolate from the China-Myanmar border area. *Mem Inst Oswaldo Cruz* 2015; **110**: 814–6.

72 Shen H-M, Chen S-B, Wang Y, Xu B, Abe EM, Chen J-H. Genome-wide scans for the identification of Plasmodium vivax genes under positive selection. *Malar J* 2017; **16**: 238.

73 Gilabert A, Otto TD, Rutledge GG, *et al.* Plasmodium vivax-like genome sequences shed

new insights into Plasmodium vivax biology and evolution. *PLoS Biol* 2018; **16**: e2006035.

74 Popovici J, Friedrich LR, Kim S, *et al.* Genomic Analyses Reveal the Common Occurrence and Complexity of Plasmodium vivax Relapses in Cambodia. *MBio* 2018; **9**. DOI:10.1128/mBio.01888-17.

75 Pearson RD, Amato R, Auburn S, *et al.* Genomic analysis of local variation and recent evolution in Plasmodium vivax. *Nat Genet* 2016; **48**: 959–64.

76 Auburn S, Benavente ED, Miotto O, *et al.* Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics. *Nat Commun* 2018; **9**: 2585.

77 Cowell AN, Valdivia HO, Bishop DK, Winzeler EA. Exploration of Plasmodium vivax transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing. *Genome Med* 2018; **10**: 52.

78 Menard D, Chan ER, Benedet C, *et al.* Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy Plasmodium vivax strains. *PLoS Negl Trop Dis* 2013; **7**: e2489.

79 Chan ER, Menard D, David PH, *et al.* Whole genome sequencing of field isolates provides robust characterization of genetic diversity in Plasmodium vivax. *PLoS Negl Trop Dis* 2012; **6**: e1811.

80 Sanguinetti L, Toti S, Reguzzi V, Bagnoli F, Donati C. A novel computational method identifies intra- and inter-species recombination events in Staphylococcus aureus and Streptococcus pneumoniae. *PLoS Comput Biol* 2012; **8**: e1002668.

81 Auburn S, Getachew S, Pearson RD, *et al.* Genomic Analysis of Plasmodium vivax in Southern Ethiopia Reveals Selective Pressures in Multiple Parasite Mechanisms. *J Infect Dis* 2019; **220**: 1738–49.

82 Gelabert P, Sandoval-Velasco M, Olalde I, *et al.* Mitochondrial DNA from the eradicated European Plasmodium vivax and P. falciparum from 70-year-old slides from the Ebro Delta in Spain. *Proc Natl Acad Sci U S A* 2016; **113**: 11495–500.

83 Pasini EM, Böhme U, Rutledge GG, *et al.* An improvedPlasmodium cynomolgigenome assembly reveals an unexpected methyltransferase gene expansion. *Wellcome Open Res* 2017; **2**: 42.

84 Tachibana S-I, Sullivan SA, Kawai S, *et al.* Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. *Nat Genet* 2012; **44**: 1051–5.

85 Auburn S, Böhme U, Steinbiss S, *et al.* A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res* 2016; **1**: 4.

86 Poplin R, Ruano-Rubio V, DePristo MA, *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017; : 201178.

87 Shultz A. Whole-genome resquencing for population genomics (Fastq to VCF). Harvard FAS Informatics. 2018; published online Jan 10. https://informatics.fas.harvard.edu/whole-genome-resquencing-for-population-genomics-fastq-to-vcf.html#variantcalling (accessed Sept 15, 2019).

88 Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour* 2017; **17**: 44–53.

89 Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. 2019.

90 Charif D, Lobry JR. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, eds. Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 207–32.

91 Pfeifer B, Wittelsbuerger U, Ramos-Onsins SE, Lercher MJ. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. Molecular Biology and Evolution. 2014; **31**: 1929–36.

92 Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. *Genetics* 1992; **132**: 583–9.

93 Nei M. Molecular Evolutionary Genetics. 1987. DOI:10.7312/nei-92038.

94 Wakeley J. The variance of pairwise nucleotide differences in two populations with migration. *Theor Popul Biol* 1996; **49**: 39–57.

95 Hudson RR. A new statistic for detecting genetic differentiation. *Genetics* 2000; **155**: 2011–4.

96 Verity R, Nichols RA. What is genetic differentiation, and how should we measure it—GST, D, neither or both? *Mol Ecol* 2014. https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12856.

97 Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018; **35**: 526–8.

98 Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics* 2011; **27**: 592–3.

99 Jukes TH, Cantor CR. Evolution of Protein Molecules. Mammalian Protein Metabolism. 1969; : 21–132.

100 Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* 1986; **17**: 57–86.