Supplementary Materials

Materials and Methods

Genetic measures

Multiple measures of genetic similarity and genetic differentiation between unions were used (table S2). Genetic diversity was measured using the complexity of infection (COI), proportion of polygenomic infections, pairwise barcode difference, pairwise drug-resistance marker difference, principal component analysis (PCA) distance, the proportion of identical by descent (IBD), and the proportion of identical barcodes. Pairwise difference measures the proportion of SNP differences and was calculated using the number of SNP differences divided by the total number of sites v1) excluding missing data and assuming that SNP difference between heterozygous call and homozygous call is 0.5 and v2) excluding all missing data and heterozygous calls. THE REAL McCOIL (*1*) was used to estimate COI and allele frequencies. The proportion of IBD was estimated using a Hidden Markov Model described in (*2*). We further estimated the proportion of IBD for each pair of samples by assuming that all samples were from a single randomly mixed population (*pooled*), or by analyzing samples from each union separately (*separated*). PCA distance was defined as the Euclidean distance in the PC1/PC2 plane. F_{ST} was calculated for both barcodes and drug markers using Weir and Cockerham's method (*3*) between all union pairs with sample sizes > 20. Normalized pairwise difference was calculated by subtracting the average within-union pairwise difference between two unions from the between-union pairwise difference. *Infomap (4)* was used for clustering unions together based on genetic similarity and travel among unions.

The odds ratio of observing nearly identical barcodes with respect to the residence location or travel pattern was calculated as follows:

$$Odds\ ratio = \frac{\frac{Prob(nearly\ identical\ barcoes \mid C)}{Prob(not\ nearly\ identical\ barcodes \mid C)}}{\frac{Prob(nearly\ identical\ barcodes \mid not\ C)}{Prob(not\ nearly\ identical\ barcodes \mid not\ C)}}$$

where *C* is the residence location or travel pattern (e.g., the condition that two individuals live in the same union, or work in the same union, or travel to the same union, etc.). Nearly identical barcodes were defined as barcodes with less than a 10 % SNP difference, and “not nearly identical barcodes” was defined as barcodes with SNP differences between the 25th and 75th percentiles for all SNP differences.

Geographic distance and travel measures

Geographic distances were calculated as both the Euclidean distance and road distance between union centroids. For simplicity, the results in the main text only include those based on road distance. Two unions were considered to have “indirect” travel if they were both connected by travel to another union that had non-zero incidence. Travel survey was collected down to the village level wherever patients could provide it. However, complete and

accurate data on the locations of these villages was not available, limiting the analyses to union level. Efforts are underway by members of our team to map these villages, so they can be used in future analyses.

The association between genetic data, travel survey, and mobile phone data

We compared genetic data with population level travel survey data. We examined how SNP differences (denoted by *x*) related to the travel survey data, and compared the empirical results with 100 permutations (Fig. 2). We considered three scenarios: 1) individuals living in the same union (denoted by T_1), 2) individuals coming from places with direct travel (denoted by T_2), and 3) individuals coming from places with indirect travel (denoted by T_3). Specifically, we calculated the proportion of parasite pairs under these three scenarios, given different SNP thresholds (denoted by *S*), as follows:

$$Prop\ (T_1\ |x < S) = \frac{Prop\ (T_1, x < S,)}{Prop\ (x < S)}$$

$$Prop\ (T_2\ |x < S, not\ T_1) = \frac{Prop\ (T_2, x < S \mid not\ T_1)}{Prop\ (x < S \mid not\ T_1)}$$

$$Prop\ (T_3\ |x < S, not\ T_1, not\ T_2) = \frac{Prop\ (T_3, x < S \mid not\ T_1, not\ T_2)}{Prop\ (x < S\ |not\ T_1, not\ T_2)}$$

We excluded T_1 when calculating T_2, and excluded T_1 and T_2 when calculating T_3, in order to identify the signal associated with each scenario separately. Our results show that parasite pairs with smaller SNP differences were more likely to come from the same unions, unions with direct, or unions with indirect travel, than random permutations (Fig. 2A, B, C), indicating that genetic data was consistent with travel survey.

We performed a similar analysis using mobile phone data. Because almost all pairs of locations have direct travel inferred from mobile phone data, instead of calculating the proportion of locations with direct travel, we calculated the proportion of parasite pairs from locations with *higher* direct travel (>0.1%). The results show parasite pairs with smaller SNP differences, not living in the same unions, were more likely to come from unions with higher direct travel, indicating the association between genetic similarity and mobile phone data (Fig. 2D).

Quantifying the probability of a geographic distance given a SNP difference

To investigate the relationship between geographic and SNP distance, we considered six geographic windows (0, 0–10, 10–15, 15–20, 20–40, 40–100km) and six SNP difference windows (0, 0–10, 10–17.5, 17.5–25, 25–30, 30–35%). For all pairs of unions within a specified geographic distance window d, the proportion of sample pairs with SNP differences within each SNP difference window s was calculated $Prob(s|d)$. We then calculated $Prob(d < X|s)$ for a given threshold distance X representative of local transmission. Here, $Prob(d < X|s)$ is the probability that for all pairs of samples from unions less than distance X, they have a given SNP difference s:

$$Prob(d < X|s) = \frac{Prob(d < X, s)}{Prob(s)} = \frac{\sum_{d<X} Prob(s|d)\ Prob(d)}{\sum_{\forall d} Prob(s|d) Prob(d)}$$

assuming a uniform prior for d, this simplifies to: $Prob(d < X|s) = \frac{\sum_{d<X} Prob(s|d)}{\sum_{\forall d} Prob(s|d)}$

Comparison between genetic measures and results from epidemiological models

The proportion of imported cases inferred from the epidemiological model and genetic mixing index were positively correlated (Spearman's correlation test, ρ>0). The proportion of imported cases was higher for unions with a high genetic mixing index than those for unions with neutral genetic mixing index (Fig. 4C), although not statistically significant (*p*-values > 0.05). We also identified the pairs of unions with an unusually high proportion of nearly identical barcodes given the geographic distance between them (fig. S11). These unions were significantly more likely to have inferred parasite flow from the epidemiological model parameterized by the travel survey data (permutation test with matched geographic distance, 10,000 replicates; *p*-value = 3×10^{-4}). Finally, parasite flow inferred from the epidemiological model was higher among the unions within the same genetic clusters identified using the proportion of identical barcodes (fig. S2B). Parasite flow among unions within the same genetic cluster was higher than that among unions in different genetic clusters (permutation test, 1000 replicates; *p*-value <0.001 [travel survey] and = 0.07 [mobile phone]). Finally, after controlling for geographic distance, we found that genetic similarity and genetic differentiation were positively and negatively correlated with parasite flow inferred from the epidemiological model, respectively, using the Mantel test (table S3).

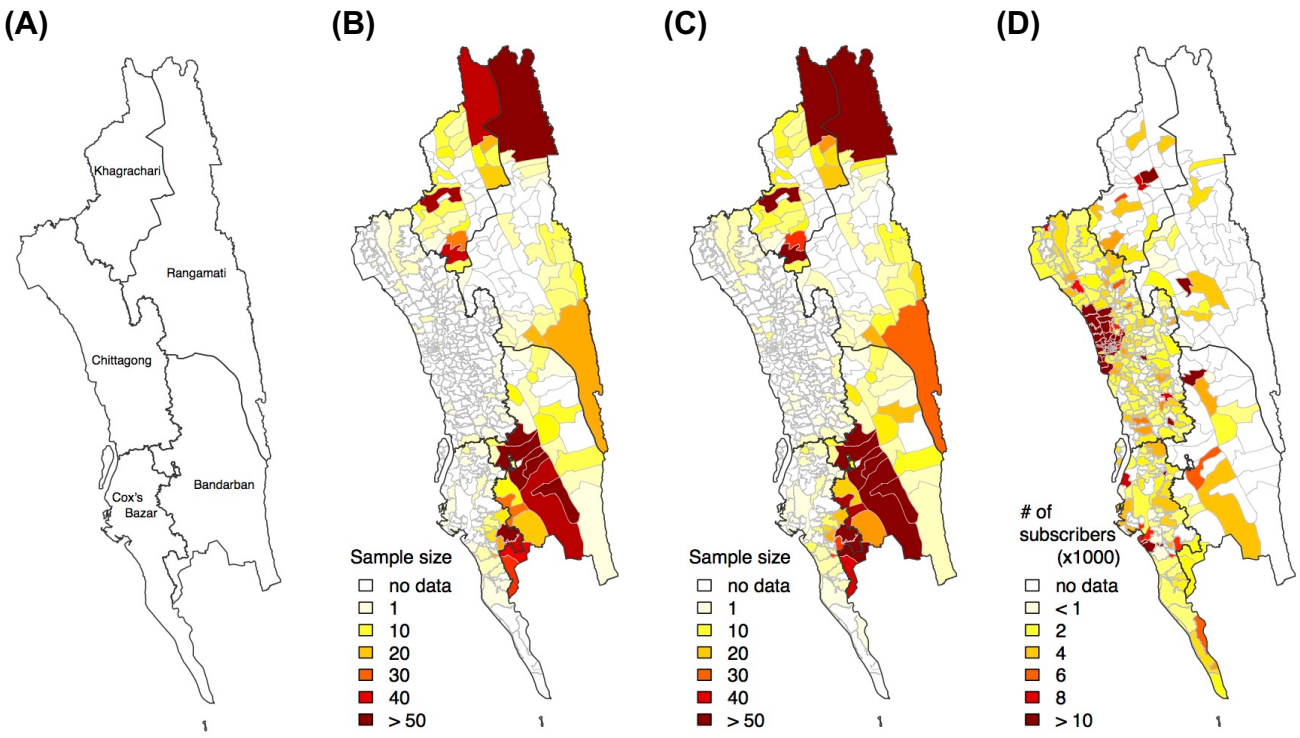

Fig. S1. Sample distribution. (A) District map in the CHT. Sample distribution of genetic (B), travel survey (C) and mobile phone (D) data.

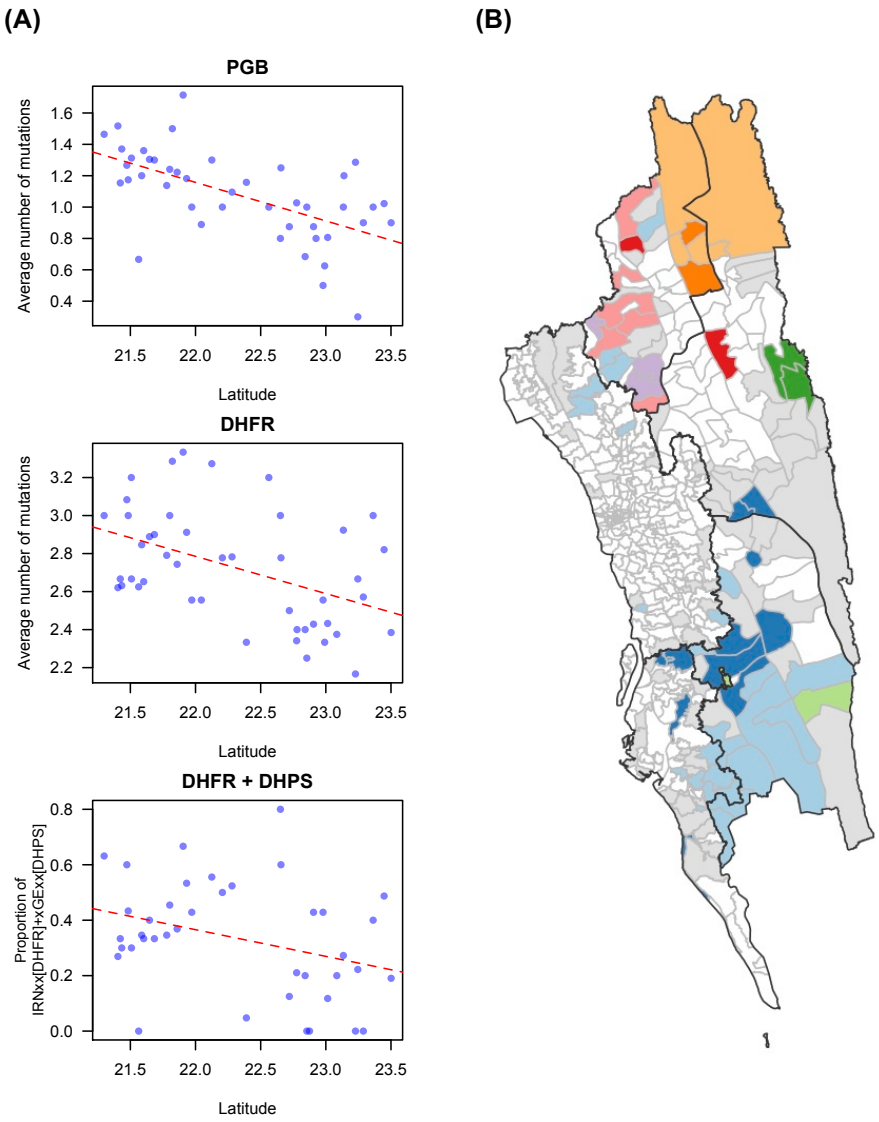

Fig. S2. Drug resistant markers and the proportion of identical parasites showing spatial signal. (A) The drug resistance-related markers were significantly associated with latitude, including PGB mutations that were found in genetic background of K13 mutations that lead to artemisinin resistance (Pearson's correlation test, *p*-value= 1.58×10^{-5}, *r*=–0.601), DHFR mutations that mediated pyrimethamine resistance (Pearson's correlation test, *p*-value= 0.0018, *r*=–0.453), and the proportion of the haplotype of IRNxx [DHFR] and xGExx [DHPS], which was shown to be associated with treatment failure for the combination of pyrimethamine and sulfadoxine (Pearson's correlation test, *p*-value= 0.035, *r*=–0.335). Red dotted line is the fitted linear regression line. **(B)** The unions were clustered using genetic information, the proportion of identical parasites between locations (see table S2 for clustering using other genetic measures). White color represents the unions without genetic data; grey color represents the unions that had genetic data but were not clustered with any other union; other colors represent identified clusters (i.e. unions in the same cluster were colored using the same color).

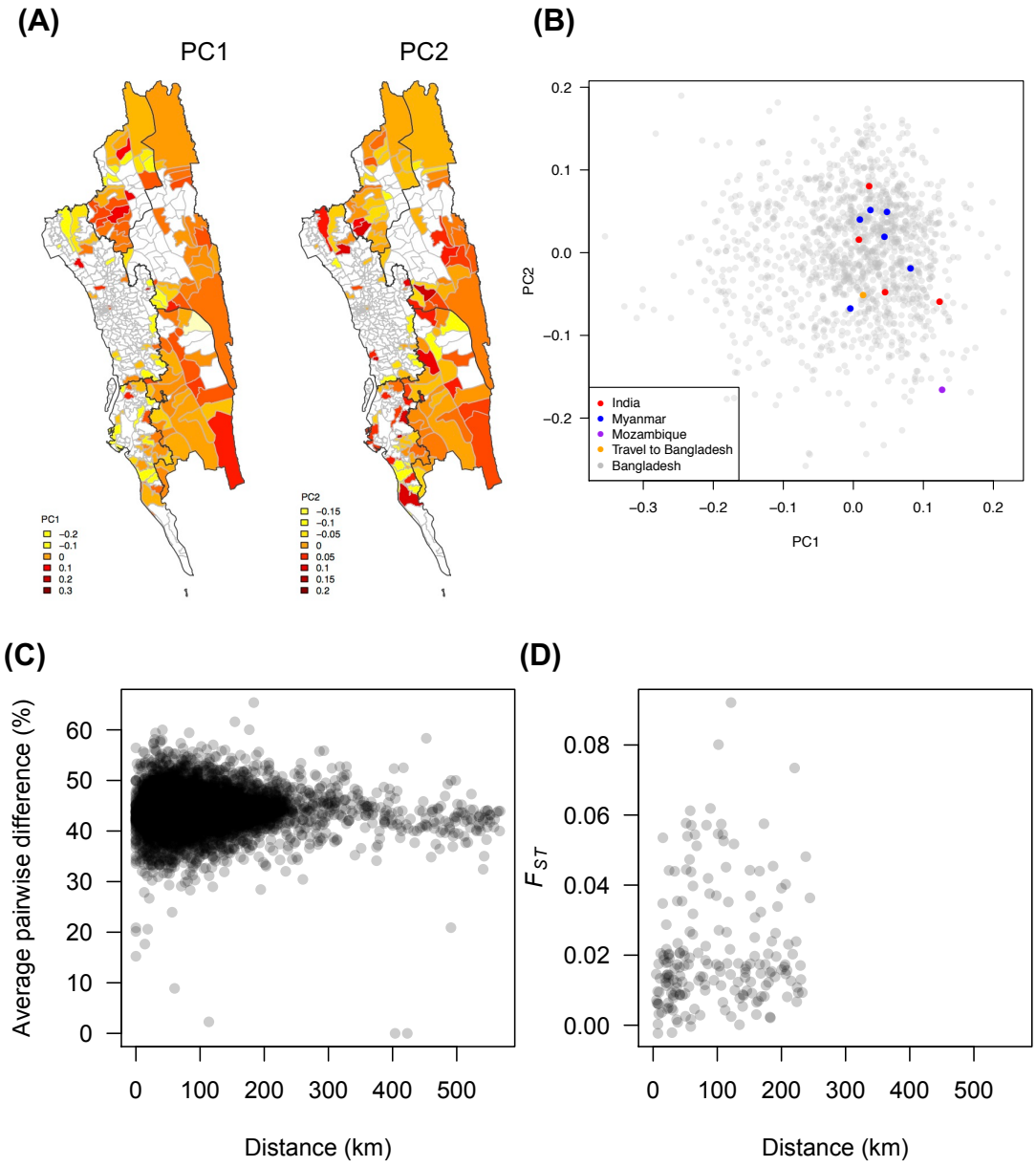

Fig. S3. Commonly used genetic measures showing little spatial signal.
(A) Pattern of genetic variation was presented by the first two principal components from PCA analysis. The color shows the average PC1 or PC2 values for each union (white means no data). There was no clear spatial trend in PC1 or PC2 values. **(B)** Genetic barcodes of parasites in international travelers were not distinguishable from people who did not travel or only traveled within Bangladesh, from PCA analysis. The case from Mozambique was an immigrant and was an outlier in the plot. **(C)(D)** Average pairwise difference (%) (C) and F_{ST} (D) were not associated with geographic distance.

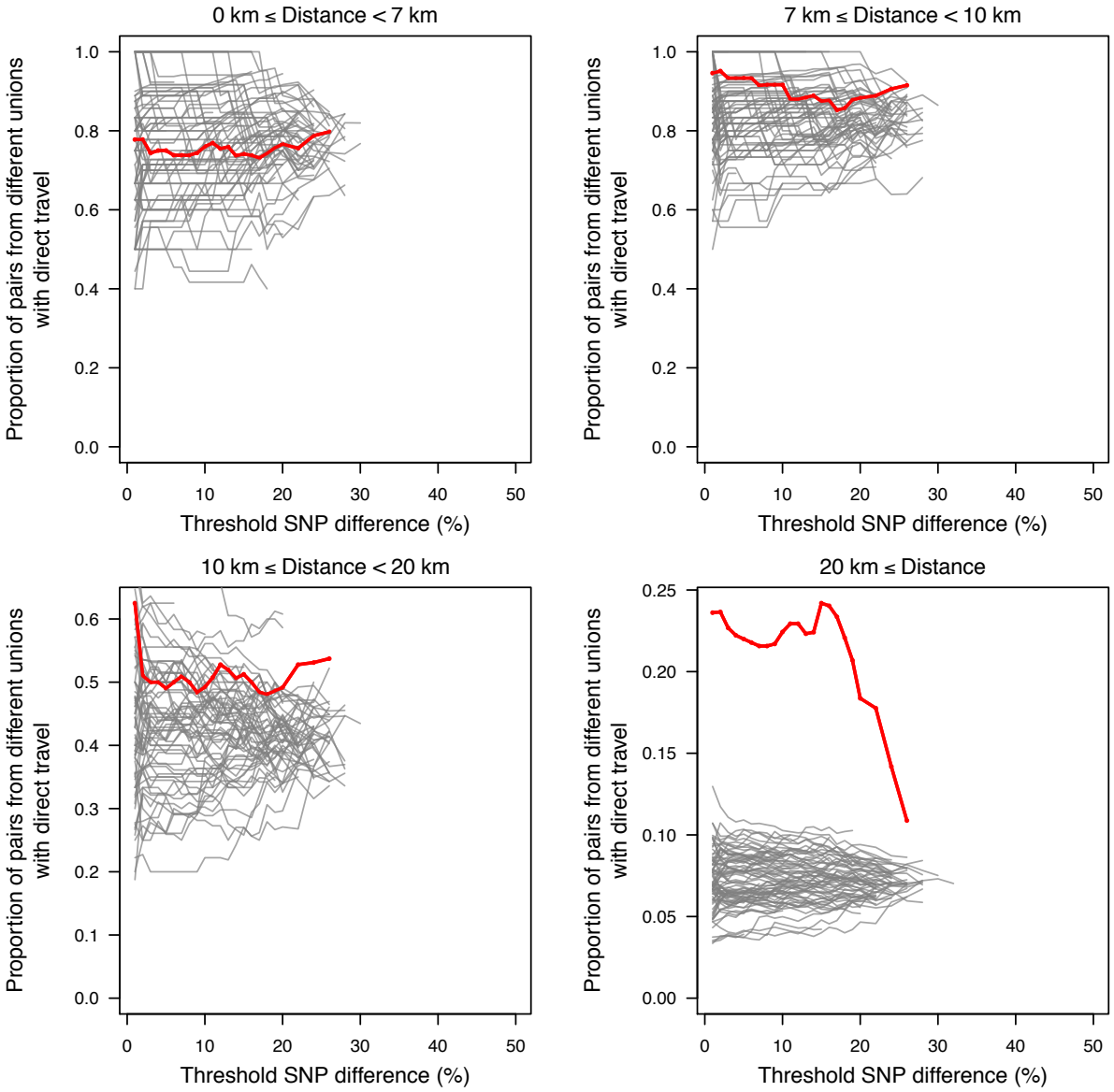

Fig. S4. The association between SNP difference and travel at varying distances. Sample pairs with smaller SNP differences were more likely to come from unions with direct travel; this is mainly driven by samples from unions that were ≥20km apart.

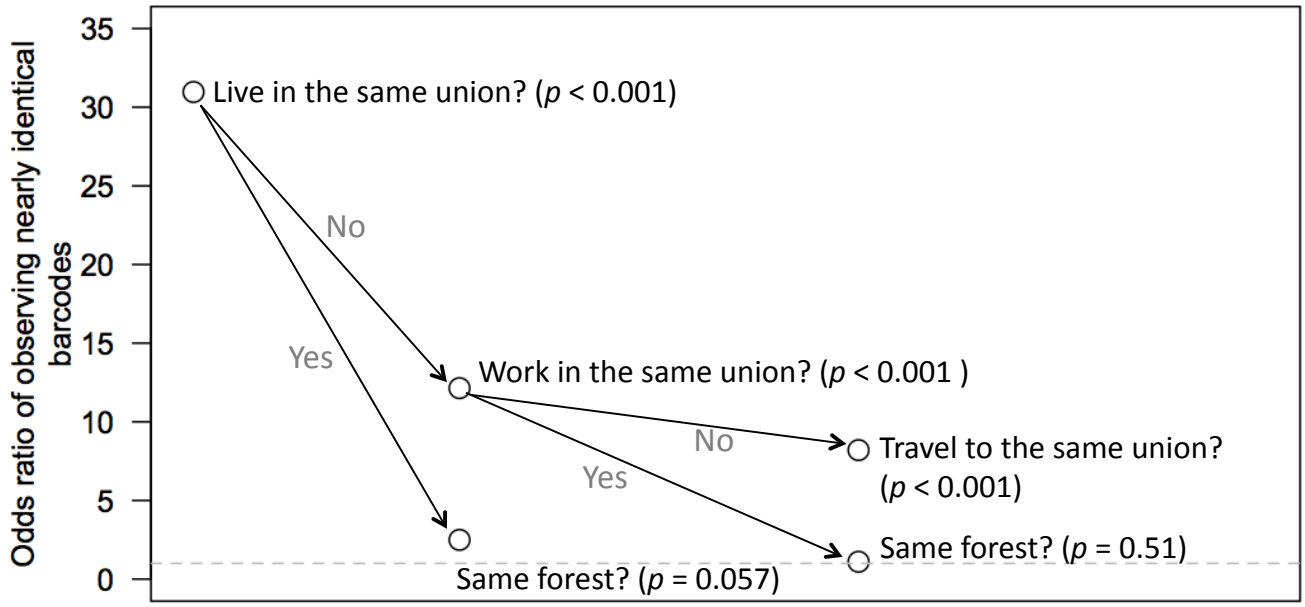

Fig. S5. Odds ratio of observing nearly identical barcodes with respect to resident locations and travel patterns. The odds ratio of observing nearly identical barcodes with respect to living in the same location was 30.97 (*p*-value <0.001). Limiting to individuals living in the same locations, the odds ratio with respect to living in the same forest was 2.50 (*p*-value = 0.057). Given that individuals did not live in the same location, the odds ratio with respect to working in the same forest was 12.14 (*p*-value <0.001); Given that individuals did not live or work in the same locations, the odds ratio with respect to traveling to the same location was 8.2 (*p*-value <0.001).

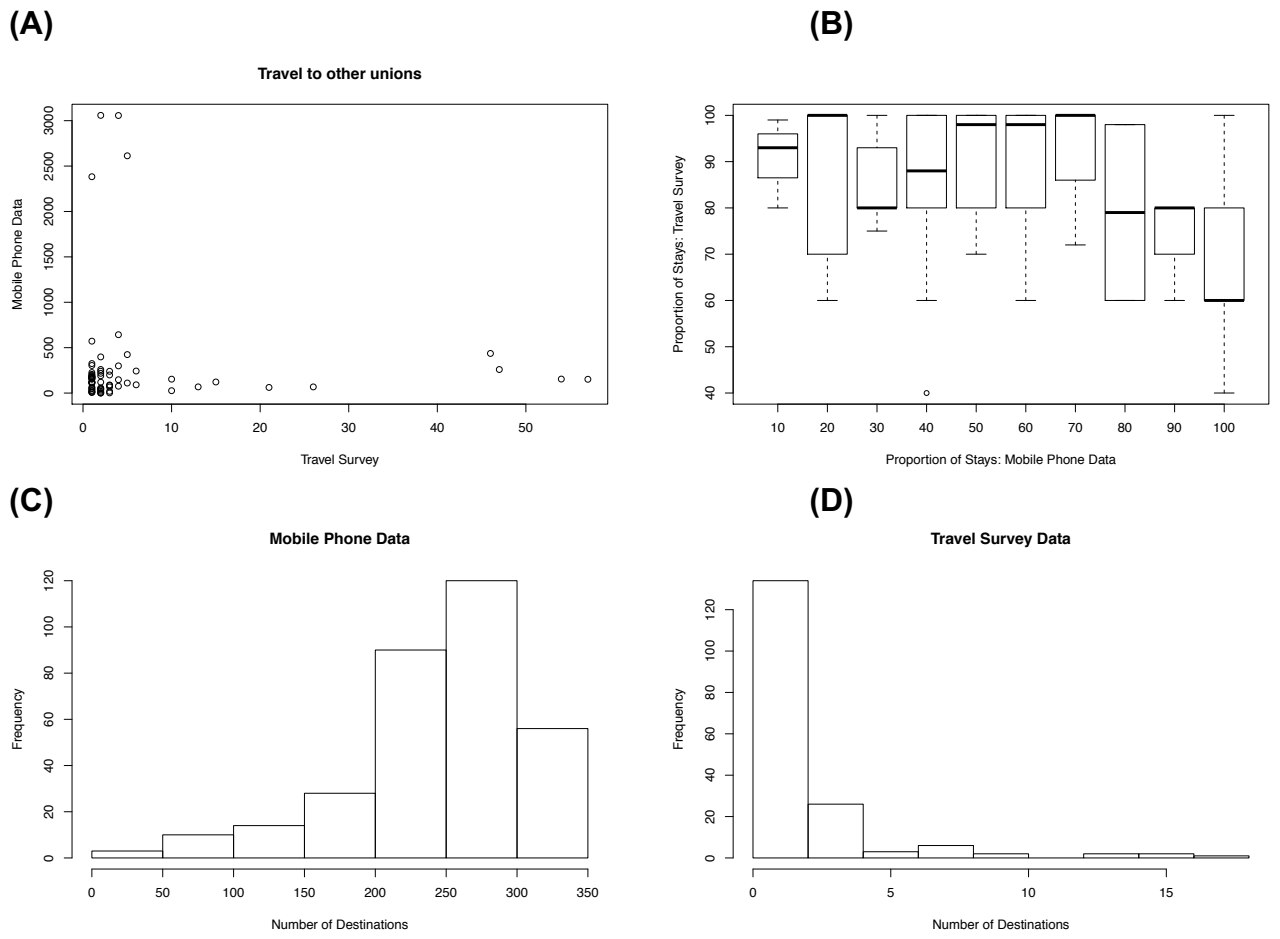

Fig. S6. The differences in mobile phone versus travel survey data. (A) The number of trips to other unions from either the travel survey or mobile phone data. In general, the mobile phone data trip counts are 1-2 orders of magnitude greater than the travel survey data. **(B)** From both data sets, we estimated the proportion of time subscribers or individuals who do not travel (stays) by either remaining in their residence location or their previous day's tower location. The travel survey data estimates a higher proportion of time staying. The **(C)** mobile phone data and **(D)** travel survey data also vary in the number of destinations from each union. The mobile phone data estimates a much higher number of destinations than the travel survey data.

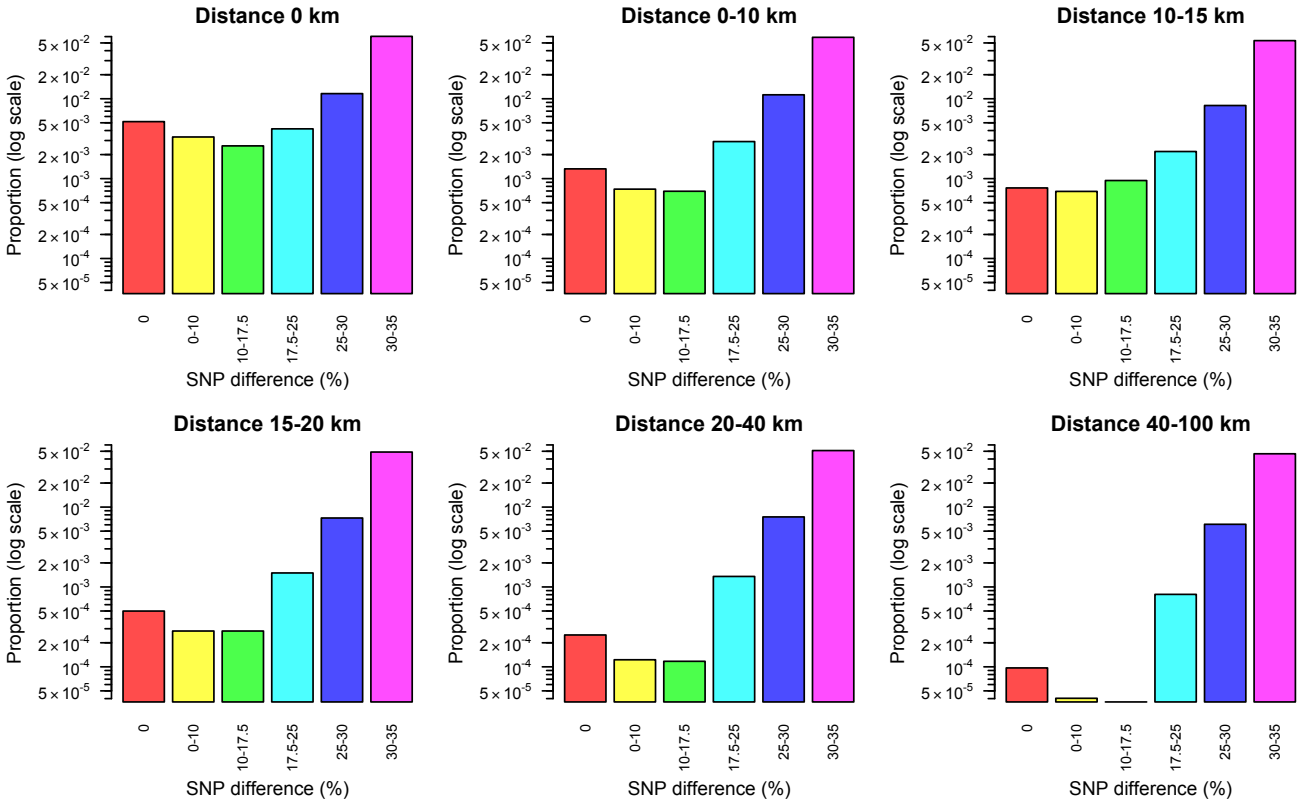

Fig. S7. The association between genetic data and geographic distance was only obvious for small SNP differences. Pairs of parasites sampled from unions that are geographically closer were more likely to be genetically similar. The proportion of intermediate SNP differences did not vary much with geographic distance.

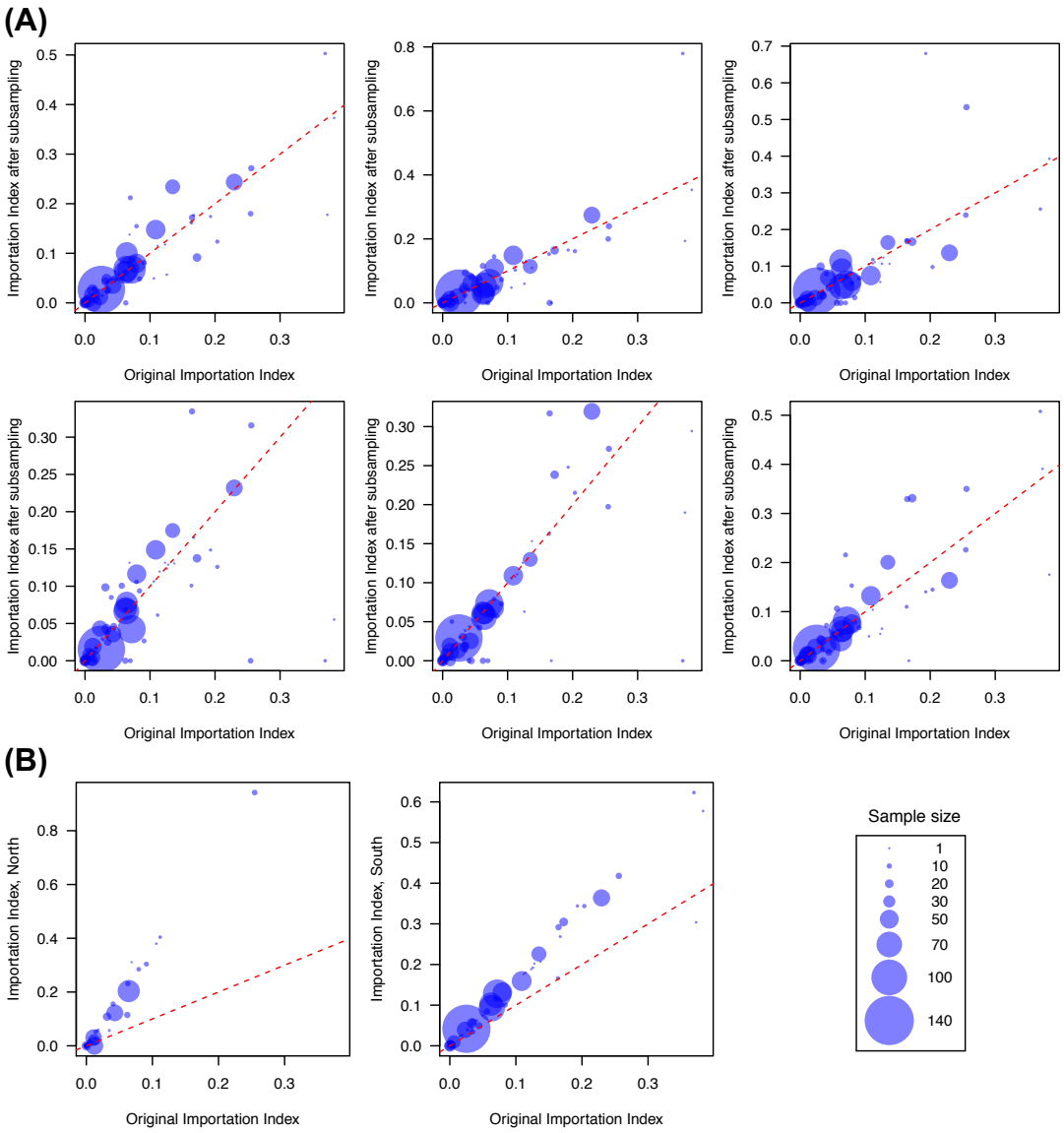

Fig. S8. Genetic mixing index was robust to subsampling randomly and geographically. We performed subsampling (80%) **(A)** and separated the northern and southern samples **(B)** to test the sensitivity of genetic mixing index to sampling, and the results remained qualitatively similar**.** The latitude of 22.6 was used as the cutoff for separating the northern and southern samples. These results suggest that the importation index is a robust measure.

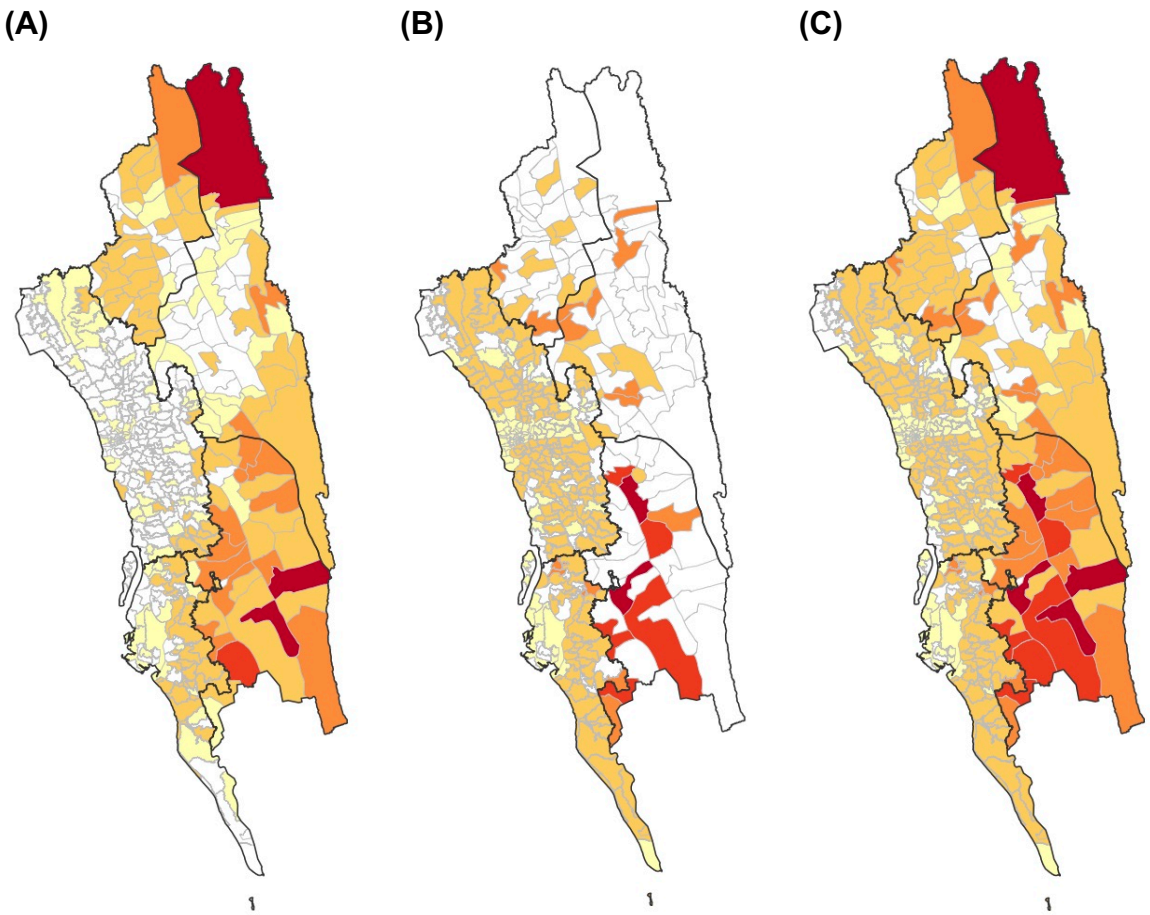

Fig. S9. Sources of parasite importations based on the epidemiological models parameterized by the mobility data. Travel survey data **(A)**, mobile phone data **(B)** or a combination of both data sets **(C)** were used to calculate the source value for each union. Source ranks were calculated using the total contribution of each location to all other locations in each data set. Source ranks (the highest source values are colored red and the lowest values are colored light yellow) are shown using each data set individually (A)(B). To combine source ranks from both data sets, we used either the higher source value based on each data set or the source value if they were equal from each type of data (C). White color means no data.

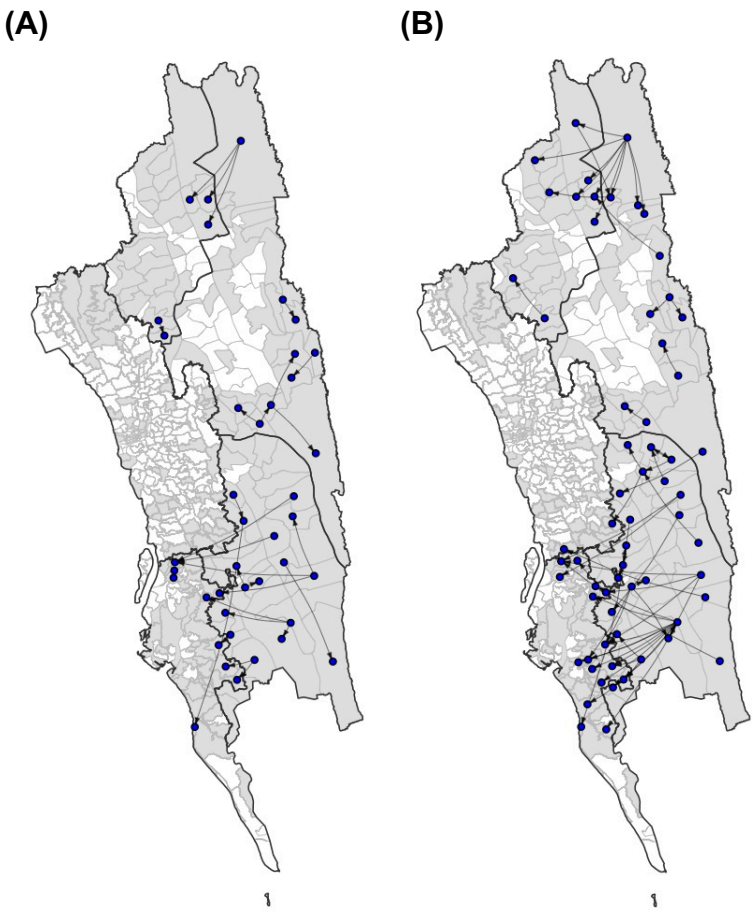

Fig. S10. The top routes of importation based on the travel survey data. We further analyzed the travel survey to **(A)** include only work travel**,** or **(B)** exclude work travel since this type of travel was quantified per week, as opposed to every 2 months. Similar to all travel, we calculated importations using the incidence values per union. In both instances, the top 25% of routes are shown.

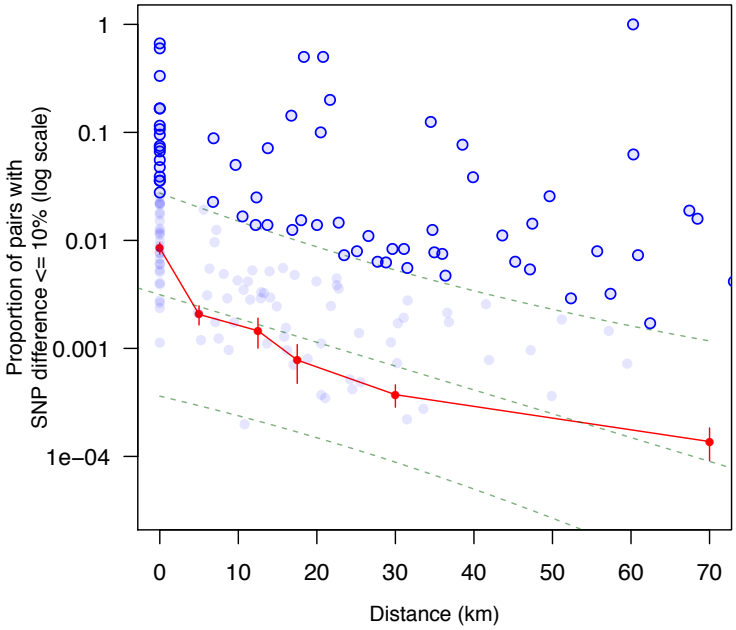

Fig. S11. Genetic signals were associated with parasite flow inferred from epidemiological models. The pairs of unions sharing an unusually high proportion of nearly identical barcodes given the geographic distance between them are shown in circled blue. Green dashed lines show the fitted line and the 95% prediction interval of the proportion of pairs with SNP difference ≤10% given distance.

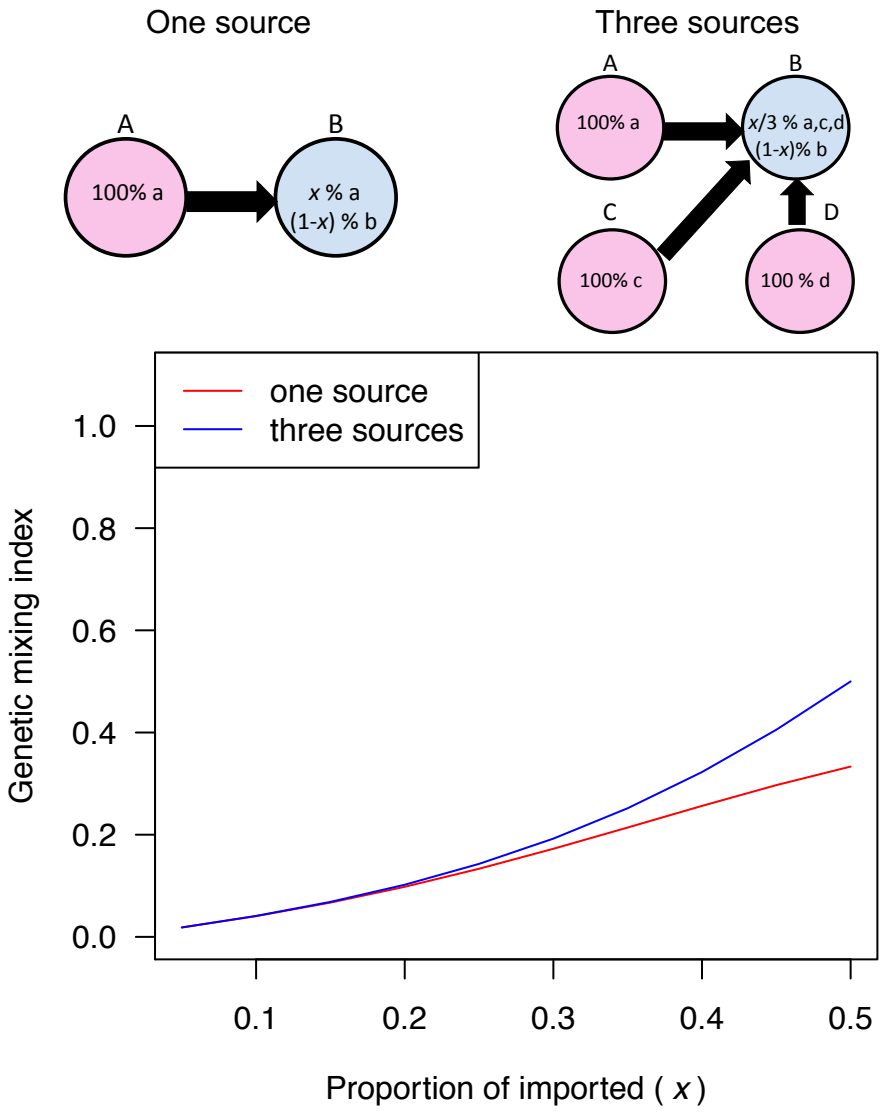

Fig. S12. Examples of genetic mixing index. We constructed simplified genetic models in order to provide intuitive interpretation for the genetic mixing index. In the simplified genetic model, we assumed that each location had its own genetic lineage (*a*, *b*, *c*, or *d*) and migration introduced genetic lineage from one location to the other (*x*). The model results indicate that the genetic mixing index increased with the proportion of imported cases and the number of source populations.

Table S1. The correlation between genetic diversity measures, incidence, and forest coverage*.

Genetic diversity measure	Spearman's ρ with incidence	*P*-value	Spearman's ρ with forest coverage	*P*-value
Average COI	0.222	0.014*	0.227	0.012*
Proportion of polygenomic infections	0.222	0.014*	0.227	0.012*
Average pairwise difference (v1)	0.007	0.953	0.090	0.419
Average drug marker difference (v1)	-0.196	0.079	-0.066	0.560
Average PCA distance	-0.203	0.065	-0.072	0.520
Average IBD proportion (pooled)	-0.031	0.800	-0.090	0.465
Proportion of identical barcodes (v1)	0.171	0.121	0.168	0.129
Genetic mixing index	-0.029	0.792	-0.081	0.462

* Only average COI and proportion of polygenomic infections are associated with incidence and forest coverage.

Table S2. Clustering by genetic data.

Genetic measure	Number of groups	Proportion (number) of unions that are not included in any group	Median geographic distance within and between groups (km)	*P*-value
Genetic similarity				
Proportion of identical barcodes (v1)	9	57% (76)	57 vs. 104	<0.001
Proportion of identical barcodes (v2)	10	48% (64)	56 vs. 109	<0.001
Proportion of nearly identical barcodes (v1) (≤0.1)	10	43% (58)	61 vs. 99	<0.001
Proportion of nearly identical barcodes (v2) (≤0.1)	22	31% (42)	30 vs. 91	<0.001
Proportion of IBD >0.9 (pooled)	9	54% (71)	39 vs. 124	<0.001
Proportion of IBD >0.9 (separated)	3	89% (117)	21 vs. 88	<0.001
Proportion of IBD >0.5 (pooled)	7	42% (55)	49 vs. 119	<0.001
Proportion of IBD >0.5 (separated)	2	88% (115)	58 vs. 70	0.15
Average IBD proportion (pooled)	1	17% (23)	NA	NA
Average IBD proportion (separated)	1	87% (114)	NA	NA
Genetic differentiation*				
Average pairwise difference (v1)	1	0% (0)	NA	NA
Average pairwise difference (v2)	1	0% (0)	NA	NA
Average pairwise difference in drug markers (v1)	1	0% (0)	NA	NA
Average pairwise difference in drug markers (v2)	1	0% (0)	NA	NA
Average normalized pairwise difference (v1)	5	36% (48)	79 vs. 86	0.045
Average normalized pairwise difference (v2)	3	39% (51)	83 vs. 85	0.240
Average normalized pairwise difference in drug markers (v1)	11	43% (57)	66 vs. 85	<0.001
Average normalized pairwise difference in drug markers (v2)	5	41% (54)	82 vs. 84	0.230
Average PCA distance	1	0% (0)	NA	NA
F_{ST} (barcode)	2	85% (111)	91 vs. 57	0.033
F_{ST} (drug markers)	2	86% (113)	85 vs. 89	0.320

*We inferred clusters of unions where parasites were genetically more similar based on genetic differentiation/similarity between unions using *Infomap*. Genetic measures similar to proportion of identical barcodes were able to identify meaningful genetic clusters (the number of groups was greater than 1 and within-group geographic distance was smaller than between-group distance), while other common genetic measures, including average pairwise difference, F_{ST}, and PCA were not.

Table S3. *P*-values[#] of the Mantel test between genetic measures and results from epidemiological modeling controlling for geographic distance.

Genetic	Proportion of parasite flow (travel survey)	Proportion of parasite flow (mobile phone)	Amount of parasite flow (travel survey)	Amount of parasite flow (mobile phone)
Genetic similarity				
Prop. of identical barcodes (v1)	**0.002**	0.058	**0.002**	**0.007**
Prop. of identical barcodes (v2)	**0.002**	0.126	**0.002**	**0.021**
Prop. of nearly identical barcodes (v1) (≤0.1)	**0.002**	0.102	**0.002**	**0.016**
Prop. of nearly identical barcodes (v2) (≤0.1)	**0.002**	0.109	**0.002**	**0.015**
Prop. of IBD >0.9 (pooled)	**0.002**	**0.016**	**0.002**	**0.005**
Prop. of IBD >0.9 (separated)	–	–	–	–
Prop. of IBD >0.5 (pooled)	**0.002**	**0.003**	**0.002**	**0.002**
Prop. of IBD >0.5 (separated)	–	–	–	–
Ave. IBD Prop. (pooled)	**0.002**	0.135	**0.002**	0.082
Ave. IBD Prop. (separated)	–	–	–	–
Genetic differentiation				
Ave. pairwise diff. (v1)	0.074	0.491	0.055	0.112
Ave. pairwise diff. (v2)	**0.030**	0.414	**0.031**	0.075
Ave. pairwise diff. in drug markers (v1)	0.090	0.207	0.125	0.157
Ave. pairwise diff. in drug markers (v2)	0.140	0.178	0.192	0.161
Ave. normalized pairwise diff. (v1)	**0.020**	0.456	**0.011**	0.206
Ave. normalized pairwise diff. (v2)	**0.004**	0.480	**0.008**	0.131
Ave. normalized pairwise diff. in drug markers (v1)	**0.005**	0.083	**0.006**	0.178
Ave. normalized pairwise diff. in drug markers (v2)	**0.005**	0.139	**0.004**	0.198
Ave. PCA distance	0.231	0.379	0.191	0.405
F_{ST} (barcode)	–	**0.024**	–	0.055
F_{ST} (drug marker)	–	**0.011**	–	–

[#]Spearman's correlation test was used. "–" means that *p*-values were not available.

Table S4. Questions in the travel survey

- Residence
- Place of work
- Have you visited the forest in the previous 2 months? If yes, where did you go?
- Have you been to another country in the previous 2 months? If yes, where did you go?
- Did you frequently travel to another village/town/city for a purpose other than work? If yes, where did you go?
- Other than this regular travel, have you been to another village/town/city in this country for a purpose other than work in the past 2 months? If yes, where did you go?

Table S5. Drug resistance markers

	Antimalarial	Gene	Amino Acid Positions	Wild Type Haplotype
K13	artemisinin	*pfkelch13*	any mutation seen in BTB/POZ and propeller domains	
DHFR	pyrimethamine	*pfdhfr*	51, 59, 108, 164	NCSI
DHPS	sulfadoxine	*pfdhps*	436, 437, 540, 581, 613	SAKAA
EXO	piperaquine	*exonuclease*	415	E
MDR-1	chloroquine, amodiaquine, lumefantrine, mefloquine	*pfmdr1*	86, 184, 1246	NYD
CRT	chloroquine	*pfcrt*	72, 73, 74, 75, 76	CVMNK
PGB (ART-R genetic background)	artemisinin	*pfarps10* *ferredoxin* *pfcrt* *pfmdr2*	127 193 326, 356 484	VDNIT

References

1. H. H. Chang *et al.*, THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. *PLoS computational biology* **13**, e1005348 (Jan, 2017).
2. S. F. Schaffner, A. R. Taylor, W. Wong, D. F. Wirth, D. E. Neafsey, hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *bioRxiv*, (2017).
3. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358 (1984).
4. M. Rosvall, C. T. Bergstrom, Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one* **6**, e18209 (Apr 08, 2011).