Tracing the origins of antimalarial resistance in *Plasmodium vivax*

INTRODUCTION

Plasmodium vivax is a malaria parasite with an immense global burden. Although infections are commonly treated with antimalarial drugs, resistance alleles have arisen throughout Southeast Asia and South America that threaten global efforts toward malaria eradication. The mechanisms and attributes of these alleles have been well-documented in its sister species *P. falciparum*, but little is known about the characteristics of positively selected alleles in *P. vivax*. In particular, few studies have addressed the global distribution of resistance alleles and the scenarios under which they may have arisen. To investigate the origins of selected regions in *P. vivax*, we assembled whole genome sequences from 148 Southeast Asian samples and identified genetically distinct populations within them. We applied a series of tests for positive selection both across all populations and within individual populations and pinpointed several previously unidentified selected regions with putative functional significance. To better understand the dispersal of positively selected alleles, we produced haplotype networks and found that selected alleles appear to have propagated broadly across Southeast Asia. Collectively, we determined that selected traits are generally not associated with extended haplotype blocks, which suggests that these alleles originated from standing variation rather than recent mutations. We expect our insights to inform strategies toward *P. vivax* elimination, since a more robust understanding of the origins of antimalarial drug resistance could shed light on their equilibrium in the face of new selective pressures.

INTRODUCION

Plasmodium vivax, like its sister species *P. falciparum*, is a parasitic protozoan that causes malaria in humans. Although the biological features, transmission patterns, and genetics of antimalarial drug resistance in *P. falciparum* have become much clearer in recent years, far fewer studies have explored these characteristics within *P. vivax*. This is due to its low mortality and a lack of—until very recently—a robust *P. vivax* blood stage culturing system [1, 2]. Nonetheless, over 2 billion people across the globe are at risk for *P. vivax* infection, which can result in serious conditions such as severe anemia, thrombocytopenia, and low birth weight [3-8].

Geographically, *P. vivax* is endemic primarily to Southeast Asia and South America. Its transmission in Africa is limited to the Horn of Africa and Madagascar, unlike *P. falciparum* [8]. The differing geographic environments, climates, and topographic features within each of these subcontinents have consequently given rise to significant diversification within the *P. vivax* genome. Across Southeast Asia in particular, migration patterns (due to refugee movements, migrant work patterns, and/or deforestation) have led to the highest intraspecific variation across all global *P. vivax* populations.

The proliferation of *P. vivax* whole genome sequencing in recent years has expanded upon these underlying characterizations of geographic populations, refining our understanding of how local populations of *P. vivax* may be established and evolve independently. For instance, a recent survey of selection within South American samples demonstrated that the current genetic diversity observed across the continent could be a product of multiple migration waves from the Old World that have outcrossed since their arrival [11]. In Southeast Asia, an analysis of whole genome sequences across the region found 40 single nucleotide polymorphisms (SNPs) that were

highly differentiated between Western Thailand and Western Cambodia, suggestive of regional adaptation [12].

The case for population genetics as a tool for strategic malaria containment is more compelling in the context of antimalarial drug resistance. First line treatments for uncomplicated *P. vivax* generally involve the antimalarials chloroquine and primaquine in combination—both chosen for their synergistic properties as well as concerns over growing chloroquine resistance [13, 14]. The expansion of antimalarial resistance in Asia has also contributed to the adoption of arteminism combination therapies (ACTs), which are aimed at minimizing drug resistance and maintaining clinical efficacy during treatment [15]. However, the adoption and deployment of these protocols varies widely by region. For instance, a combination therapy of sulfadoxine and pyrimethamine was first introduced in Thailand in 1973 for *P. falciparum* infections and is believed to be the source of several unique selective signatures in both *P. vivax and P. falciparum* [12, 16]. On the other hand, the country of Papua New Guinea uses primarily the antimalarial primaquine toward treatment of infection [17]. Since mutations that confer antimalarial resistance are key drivers of malaria evolution, a more comprehensive overview of the differences between these alleles within subpopulations will undoubtedly be immensely beneficial toward efforts to eliminate the disease.

Although many SNPs associated with drug resistance in *P. vivax* have been previously identified through genome-wide scans of selection, the majority of these analyses have been cursory, with minimal emphasis on mapping the haplotypes associated with such alleles. Along these lines, the geographic trajectories and population-level frequencies of these alleles remain poorly understood, unlike characterizations of antimalarial resistance in *P. falciparum* [18, 19]. Furthermore, although some facets of multidrug resistance in *P. vivax* have been linked to transcriptional regulation in the past, there has never been an attempt to map signatures of selection that fall outside of protein-coding regions to epigenetic marks, a technique which could shed light on how regulatory regions might confer a fitness advantage [20].

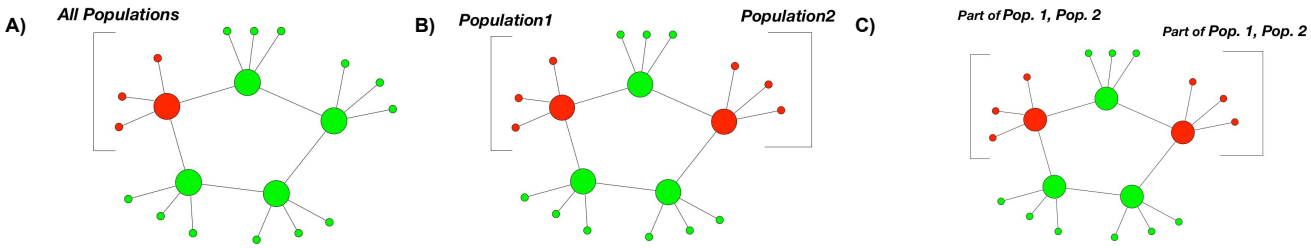

Figure 1: Scenarios depicting the propagation of antimalarial resistance alleles. Red dots represent samples with a mutation which provides drug resistance while green dots are wild-type samples. Panel A, a mutation associated with drug resistance originates once and spreads across multiple populations. In Panel B, resistance alleles originate independently in populations and propagate almost entirely within their originating population. Our third scenario, outlined in Panel C, includes multiple origins of resistance mutations that are transmitted across populations.

In our analysis, we explore three possibilities for the rise and transmission of antimalarial resistance genotypes or other selected alleles (**Figure 1A**). In the first, selected variants originate once, independent of population, yet rapidly propagate toward fixation across all populations. In the second, variants which provide antimalarial resistance originate from independent populations and are transmitted exclusively within their host population (**Figure 1B)**. In the final scenario, mutations associated with antimalarial resistance occur recurrently at the same locus and are transmitted across all populations (**Figure 1C**). An additional layer of complexity

underlying these scenarios comes from the question of whether selected drug resistance mutations have risen in frequency from standing genetic variation or are of more recent origin.

A growing body of research suggests that reduced malaria transmission can lead to higher inbreeding and increased linkage disequilibrium [21, 22]. These factors have been shown to maintain loci linked to antimalarial resistance in equilibrium, which could disrupt elimination efforts and exacerbate current resistance trends. Consequently, a more robust understanding of which scenarios of resistance allele transmission operate within populations of *P. vivax* would be beneficial for predicting how specific elimination strategies might impact or aggravate the balance of resistance alleles distributed throughout independent populations. Our inferences could also have applications in genome-wide association studies (GWAS) linking reduced rates of parasite clearance to putatively selected genetic variants. If there exist multiple haplotypes that are associated with selected variants, it may become easier to identify causal SNPs for a particular phenotype or to characterize markers associated with antimalarial resistance.

To build on the current global understanding of *P. vivax* population genetics and drug resistance, we have assembled and analyzed 148 whole genome sequences (WGS) of *P. vivax* from Southeast Asia, the most representative collection of samples compiled to date. Leveraging this multi-population data, we have identified several selective markers unique to particular populations as well as broader signatures of selection. For several alleles, we present haplotype networks and associated statistics that may be indicative of the spread of selected alleles across populations. Broadly, our results provide the necessary groundwork for future studies to further delineate how selected alleles have originated and subsequently propagated across populations.

METHODS

Preparation of variants

We acquired WGS of *P. vivax* from three previously published studies [12, 20, 23]. However, not all of these samples demonstrated sufficient coverage to be used in our analyses. To determine which samples were appropriate for the purposes of our study, we first aligned all samples to the PVP01 *P. vivax* reference genome with *bwa mem* and called variants using *HaplotypeCaller* from the *GATK* suite [24-26]. Our choice of using the PVP01 reference over the older Sal1 was based on both its high quality as well as its origin from Papua New Indonesia, which is much more geographically proximal to our samples compared to El Salvador, where Sal1 was sourced. After examining this initial call set, we chose to discard samples with fewer than 80% of variant sites genotyped.

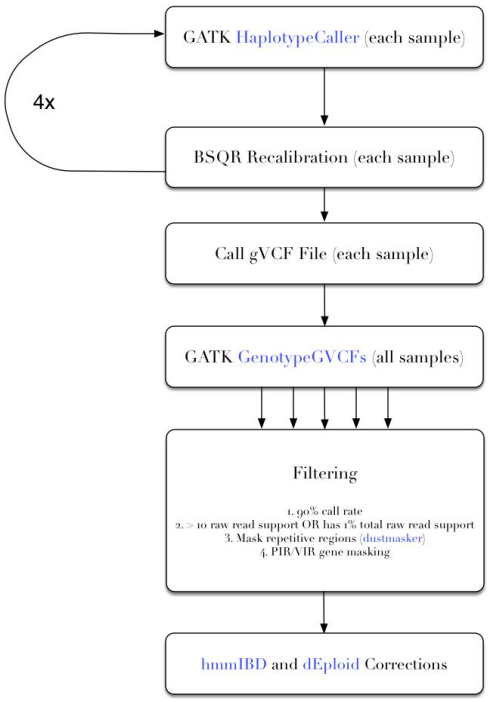

Figure 2: A summary of our pipeline to identify variants. The first five of our steps were adopted from previously published best practices. Our *hmmIBD* and *dEploid* steps enabled us to correct heterozygous genotype calls for haploid samples.

Since *P. vivax* is a non-model organism with no gold standard variant set with which to calibrate variant filtering, we next adapted best practices suggested by the developers of *GATK*, performing four steps of Base Quality Score Recalibration. These sample-specific variants were then coalesced into a single, master variant set using the GATK tool *GenotypeGVCFs*, excluding any indels and only retaining SNPs. From these samples, we again removed those featuring less than 80% of sites genotyped across the genome. Notably, these variants were processed as

originating from a diploid organism rather than a haploid organism in order to characterize mixed infections, as we later describe.

We performed several steps to trim down our variants into a high-confidence set. First, we filtered out any variants that were localized within repetitive regions of the *P. vivax* genome, as these are often sources of error in variant calling. Our candidate repetitive regions were obtained using the tool *dustmasker* with default settings [27]. Additionally, we filtered any variants that were within PIR or VIR genes—a set of genes that are highly repetitive in sequence—as they are similarly prone to genotyping errors [20]. We also filtered variants that were called in less than 90% of our samples or had greater than one alternate allele, since most genomic analysis tools are designed to accommodate a maximum of only two alleles. Furthermore, we removed variants that were supported by less than 1% of all reads at that locus or variants that were supported by less than 10 reads in total, as was suggested by a prior malaria study conducted on a similar scale [28]. These steps removed variants that were relatively rare in our samples and thus had a higher likelihood of being false positives. We designated the resultant variants as being of high-confidence and used them for the purposes of our analyses. A summary of our pipeline is depicted in **Figure 2**.

Detecting clonal isolates

Although the *P. vivax* genome is haploid throughout its residence in the human body, sequencing efforts for blood samples of *P. vivax* have frequently identified diploid sequences or other, higher forms of ploidy. These results can be attributed to multiple infections of *P. vivax* within a single host. Since the presence of multiple infections could have confounded our population genetic analyses, we performed a series of steps to correct for this possibility and ensure that the variants associated with our samples featured only a single *P. vivax* isolate.

We first clustered each of our samples by their country of origin and then applied the tool *hmmIBD*—which is optimized to pinpoint tracts of identity-by-descent in haploid species—within each population to identify which samples were most closely related to each other [29]. We next constructed a panel for each sample, derived from its ten most closely samples as identified by *hmmIBD*. The panels were then processed by the tool *dEploid*, a recently-developed program to deconvolute *Plasmodium* mixed-infection isolates based on haplotype representation [30]. These calculations provided us with isolate proportions within each sample based on the sample's genotyping. If there existed an isolate within each sample with a proportional representation of 70% or more, we modified that sample to reflect only the majority isolate's genotype. If there were no isolates within a sample whose haplotypes were represented in a proportion greater than 70%, we discarded that sample from our later analyses.

Characterizing population structure

After filtering our variants and identifying mixed infections, we performed a series of analyses to detect basic population structure. We used the R package *SNPRelate* to conduct a principal components analysis on our SNPs, filtering them for linkage disequilibrium based on an r^2 value of 0.5 [31]. In addition, we identified the population relatedness between each group using the tool *ADMIXTURE*, similarly filtering SNPs using an LD threshold of 0.5 [32]. To determine the best K value denoting the number of population clusters, we performed ten-fold cross-validation for our several runs of admixture. We used the population clusters identified through the PCA and *ADMIXTURE* in our later analyses, as these clusters were largely concordant with geographic origin.

Since LD is frequently associated with selective sweeps, we characterized the extent of LD present within the *P. vivax* genome. We determined LD decay across populations through the

program *PopLDdecay* [36]. To identify haplotype blocks, we first applied the program *Haploview* with a 20,000 base pair window surrounding selected sites using default parameters [37]. We additionally used the *PLINK* software package—which performs functions similar to *Haploview*—using a minimum r^2 parameter of 0.2 [38].

Identifying signatures of selection within and across populations

We used the program *selscan* to calculate the nSL statistic for our variants [39]. These results were normalized using a tool included in the *selscan* package. Since we were unable to ascertain which of our alleles were ancestral or derived, we used the absolute values of these statistics for the purposes of our analyses, as was done in a prior study [20]. We also calculated the H12 statistic—which is optimized to pinpoint soft selective sweeps that might be missed by other methods—with scripts provided by Nandita Garud [40].

Notably, out of all the selection statistics commonly used, only the nSL statistic does not require a recombination map, which has not yet been identified for *P. vivax*. Furthermore, nSL also has been demonstrated to be effective at identifying soft selective sweeps on standing genetic variation in addition to hard selective sweeps [41]. Given these two advantages, we chose the nSL results to be our primary reference point for experiments involving selection. We calculated these statistics across all populations as well as between populations, choosing nSL score cutoffs as 99.9^{th} percentile for the entire variant set. To determine if sample size might bias nSL results between populations, we performed a permutation test in which we determined the likelihood of recovering a similar set of candidates for selected variants using a reduced sample size compared to a larger size.

Determining transmission characteristics of selected variants and drug resistance alleles

We employed several approaches to better understand the nature of haplotypes associated with positively selected variants across populations. First, we performed TCS network analyses around variants of interest, with window sizes of 40,000, 20,000, 10,000, and 2,000 base pairs [42]. These networks were visualized using the software package *popart* [43]. Additionally, we used the software package associated with the H12 statistic to cluster haplotypes—an approach that had been previously deployed for analyses of soft selective sweeps in Drosophila and domesticated dogs—using window sizes of 400, 200, 100, and 50 SNPs [40, 44].

We also performed pairwise comparisons of the haplotypes surrounding selected loci in individual samples to infer similarity, with a maximum of 2 missing sites. Using window sizes ranging from 500 to 5000 base pairs, these pairwise comparisons were then processed with the *DASH* algorithm, which leverages shared sequence identity to infer identity-by-descent haplotype clusters [45]. Furthermore, we employed the hierarchical clustering-based UPGMA algorithm to develop phylogenetic trees for sequences surrounding selected variants, as suggested in a prior malaria study [41, 46].

Characterizing the epigenetic landscape of P. vivax isolates

To answer whether the epigenetic makeup of *P. vivax* was under selection, we downloaded ChIP-seq data from a study surveying histone modifications within *P. vivax* sporozoites [47]. We aligned reads from that data using *bamtools* and called histone peaks using the *MACS2* software suite with the broad peaks option [48, 49]. Collectively, we were able to ascertain whole genome profiles for the histone marks H3K9ac, H3K4me3, and H3K9me3.

RESULTS

Although we originally inferred genotypes across 281 samples for analyses, our *dEploid* correction and filtering protocol reduced this set to 148 samples. These samples originated across the Southeast Asian countries Myanmar, Thailand, Cambodia, Vietnam, Malaysia, and Papua New Guinea (PNG). Through our variant calling approach, we were able to identify 271,334 high-confidence variants, which we continued with in our later analyses. In our PCA, we characterized three major population clusters, with 2.68 percent of the genetic variation captured by the first eigenvector and 1.89 captured by the second (**Figure 3A)**. We classified these clusters as the Myanmar-Thailand (MT), Cambodia-Vietnam (CV), and Malaysia-PNG (MP) populations, respectively. We found F_{st} between MT and CV to be 0.03 CV and MP as 0.11, and MT and MP to be 0.15.

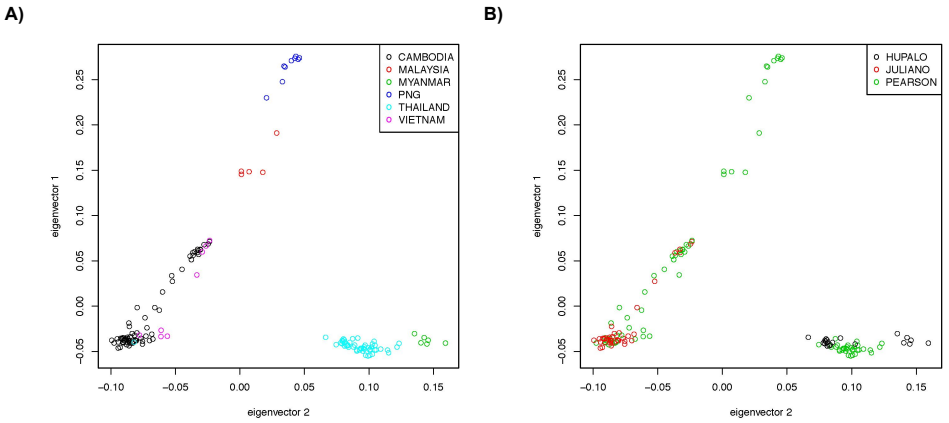

Figure 3: Principal components analysis of 148 *P. vivax* samples across Southeast Asia. Samples were filtered for linkage disequilibrium and subjected to PCA. Panel A depicts samples by their country of origin and demonstrates that they are differentiated largely based on their region of origin. In Panel B, samples are labelled by the experiment in which they were originally sequenced to verify that batch effects did not biased our genotyping process.

Since it was possible that our use of sequence data from three different experiments could have biased our variant calling strategy through batch effects, we also labelled our PCA using the experiments from which samples were derived. We found that samples derived from different experiments but localized to the same country largely overlapped, with the partial exception of Thai samples (**Figure 3B**). To validate the clustering suggested by our PCA, we identified clusters through ADMIXTURE (**Figure 4**). Using 10-fold cross-validation, we determined that

the appropriate cluster size for our samples was K=3, with the constituent members in agreement with those suggested by our PCA analysis.

We initially performed the nSL test across the entirety of our samples as well as individually, within populations. However, since the MP population contained significantly fewer samples than the other two populations, we used a permutation test to determine whether population size influenced the nSL test's ability to recover selected sites. We found that a sample size of 14—the extent of the MP population—was not sufficient to recover most selected sites in the two larger populations (p=0.0). Consequently, we only included samples from the MP population when determining selected sites across all populations, excluding it for the rest. A Manhattan plot of selected regions across both all three populations is shown in **Figure 5A**.

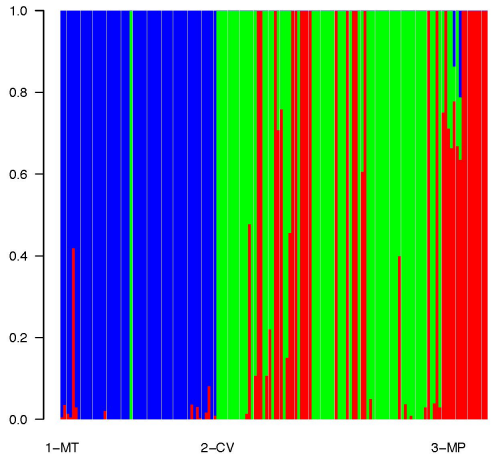

Figure 4: ADMIXTURE plot of population clusters. Using the program *ADMIXTURE,* we identified three high-confidence populations across our sample set. These included the Myanmar-Thailand (MT), Cambodia-Vietnam (CV), and Malaysia-PNG (MP)populations. Each line represents a single sample while the proportions of blue, green, and red represent the proportions of MT, CV, and MP ancestry identified within each sample, respectively.

Of these sites, 19% of the 99.9th percentile SNPs were found to be missense, protein-coding mutations. Within the MT and CV populations, we found 25% and 21% of highly selected SNPs to be missense mutations, respectively. Notable genes close to or directly implicated in these selected sites that have previously gone unreported included karyopherin alpha, which plays a major role in nuclear import, and GPI ethanolamine phosphate transferase 1, implicated in posttranslational modification pathways.

Using Hudson's F_{st} statistic, we also identified sites that appeared to be differentially selected between the Cambodia-Vietnam and Myanmar-Thailand populations. Notable genes implicated in or associated with these regions included histone-arginine methyltransferase CARM1, which could facilitate epigenetic adaptations to specific environments, and CCR4-associated factor 1, which has a fundamental role in coordinating the expression and egress of parasite invasion proteins [50]. A Manhattan plot depicting values with high F_{st} differentiation between the Cambodia-Vietnam and Myanmar-Thailand populations is shown in **Figure 5B**.

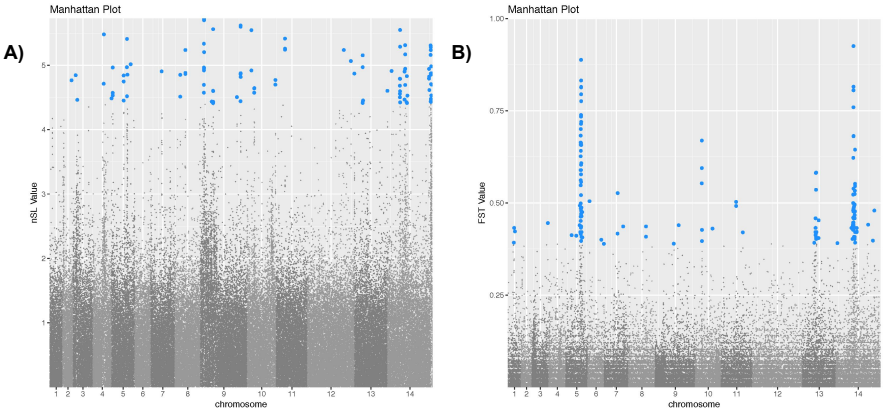

Figure 5: Manhattan plots of whole genome scans for selection and population differentiation. Panel A identifies the top 99.9th percentile of genomic sites across all samples which displayed evidence of selection based on the nSL statistic. Panel B depicts sites the top 99.9th percentile of sites that are highly differentiated by the Fst statistic between the Cambodia-Vietnam and Myanmar-Thailand populations.

We examined LD within our populations and found that it was extremely low **(Figure 6)**. In Myanmar-Thailand, LD decayed to an r^2 of 0.1 on average within 180 base pairs, while in Cambodia-Vietnam it decayed to the same amount within 170 base pairs and in Malaysia-PNG within 2600 base pairs. We subjected several our candidate selected SNPs and their encompassing genomic regions to haplotype block

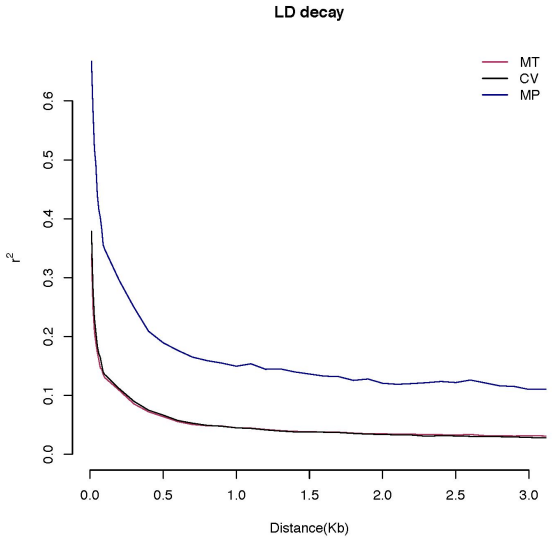

Figure 6: Average decay of linkage disequilibrium across populations. Decay within the Myanmar-Thailand and Cambodia-Vietnam populations is rapid (~200 bp to reach 0.1) while it is slightly stronger in Malaysia-PNG (~2600 bp to reach 0.1).

analysis with *Haploview* and *PLINK*. However, *Haploview* was unable to identify any significant haplotype blocks within these regions. Although *PLINK* did produce estimations of haplotype block size, the majority of its estimates included haplotype blocks of only several hundred base pairs. Moreover, most of its estimates did not include our positively selected SNPs of interest within haplotype blocks.

We applied several approaches in an attempt to identify distinct or similar haplotype backgrounds associated with selected alleles. First, we developed haplotype networks around putatively selected variants. For variants that were largely fixed across all populations, we largely found no evidence that the haplotypes bearing these alleles were clustered by population (**Figure 7A**; **Figure 7B**). For variants that were associated with high nSL scores but were not entirely fixed across all populations, we found that selected variants largely clustered together, regardless of the window size used to construct each network (**Figure 7C**). Although it cannot be ruled out that these variants were the result of recurrent mutations, their clustering implied a singular origin of the

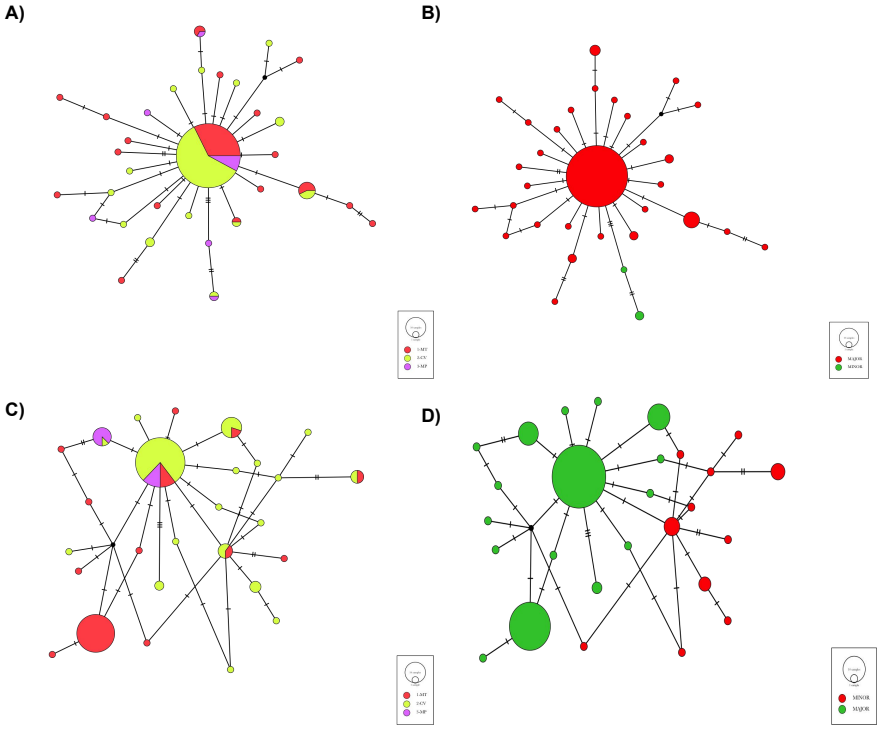

Figure 7: TCS haplotype networks for selected variants. The haplotypes depicted are 10,000 base pairs in length, centered around the selected allele. Panel A, we include the haplotype network for the variant *LT635619_377458*—which has largely reached fixation—with labellings corresponding to population of origin. Panel B includes this same network, but with labels presenting both the major and minor alleles. Panels C and D display similar networks for the selected variant *LT635616_1046912*, which has not yet reached fixation. Notably, both selected alleles do not display any major population-level organization in their networks.

alleles observed. Moreover, both the major and minor alleles did not display any evidence of population-specific clustering (**Figure 7D)**. These results largely suggest that selected variants do not remain exclusive to individual populations but instead appear to propagate widely.

The software package associated with the H12 statistic is capable of both identifying regions under selection as well as clustering haplotypes at a particular locus. Despite using multiple window sizes for this analysis, we were unable to identify any candidate variants that had potentially been subjected to soft selective sweeps. Using the H12 clustering tool, we found that haplotypes featuring selected alleles did not appear to be significantly differentiated from those without, implying that most selected variants within the *P. vivax* genome are not captured by extensive selective sweeps, hard or soft.

We further confirmed this interpretation by using the *DASH* algorithm, which clustered haplotypes based on their pairwise distance. Samples which featured selected alleles largely clustered together in multiple independent clusters. However, our ability to definitively determine whether the members of each cluster were genetically related was hindered because many clusters overlapped in their

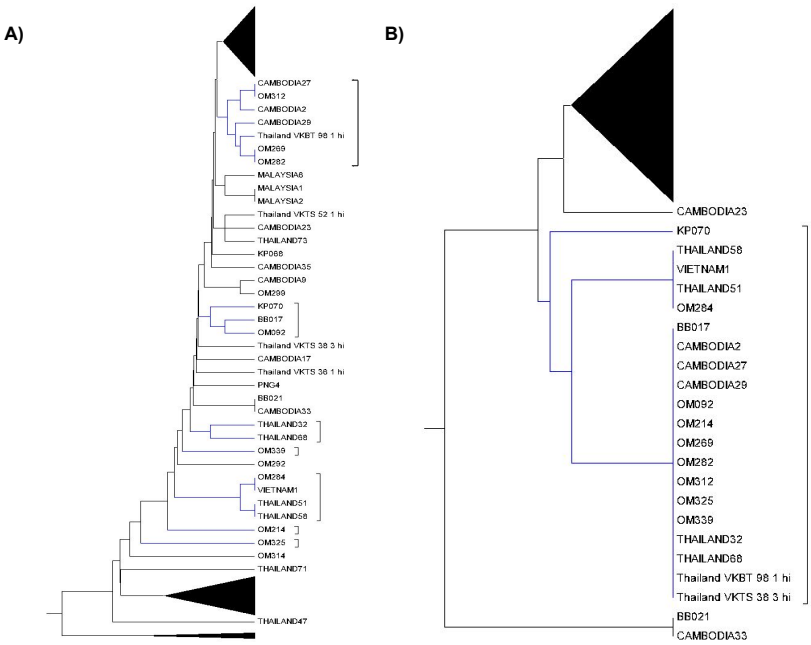

Figure 8: UPGMA tree of a missense mutation selected for in a histone acetyltransferase. The minor allele is depicted in blue with bracketing, while the rest represent samples featuring haplotypes with the major allele. Some clades have been collapsed for clarity. Panel A depicts the phylogenetic relationship amongst haplotypes encompassing 40,000 base pairs around the major and minor alleles, while Panel B depicts the relationships of 2,000 base pair haplotypes. Although minor allele haplotypes appear to have originated independently in Panel A, these haplotypes largely coalesce in Panel B, thus limiting any conclusions.

constituent members. Notably, however, almost all haplotype clusters identified through *DASH* featured members from different populations, supporting the notion that selected alleles do not remain sequestered to distinct populations but instead are found distributed throughout all.

Additionally, we performed UPGMA phylogenetic clustering on our putatively selected variants since this phylogenetic method incorporates hierarchical clustering. Although several haplotypes associated with selected variants appeared to cluster independently when analyzing trees encompassing 40,000 base pairs around these variants, we found that these clusters coalesced when comparing trees associated with smaller, 2,000 base pair windows. In agreement with our prior analyses, we did not find any clusters that exclusively featured a single population, regardless of window size (**Figure 8**). Although our analysis through this approach cannot entirely rule out independent origins of these mutations, these results in conjunction with prior evidence suggest that mutations which display evidence of selection do not occur recurrently. Moreover, the short length of haplotype blocks surrounding selected SNPs suggest a scenario in which selection began to act on standing variation after it had a chance to rise to intermediate frequency and to spread throughout multiple populations.

We also mapped selected alleles to regions associated with sporozoite epigenetic modification. However, we were unable to identify any regional overlap. This particular finding could suggest that mutations within *P. vivax* regulatory regions have not undergone any strong selection. Given our limited dataset on epigenetic modification, however, we believe a more plausible explanation is that our underlying data is insufficient to infer any conclusion.

DISCUSSION

Our study provides several promising avenues for future research that could have widespread applications for developing elimination strategies across the globe. These findings also imply that improvements in methodology are necessary for continued research in *P. vivax*. One particular advantage of our study is that we applied the *HaplotypeCaller* algorithm from GATK to call variants, as opposed to *UnifiedGenotyper*. While the latter has been used extensively in prior studies of *P. vivax*, *HaplotypeCaller* is better suited at calling variants in close proximity to each other, as suggested by its documentation. Given the extensive genetic variation previously observed for *P. vivax* and in our study, we believe this new approach could be more effective for the purposes of genotyping [11, 51].

Our study also represents the first attempt to use the *dEploid* tool to correct for polyclonal infections within *P. vivax* samples. Given the unique relapse nature of *P. vivax*, it was unsurprising to find that 124 of our original 281 samples were too mixed to be properly deconvoluted, in line with prior estimates in the region [12, 20, 23]. Furthermore, the *dEploid* package was powerful by enabling us to properly identify the correct genotype for high-confidence haploid samples in which heterozygote calls were made. Although future experimental work through single-cell sequencing of *Plasmodium* parasites may one day make tools such as *dEploid* obsolete, their capacity to ascertain the genotypes of mixed infections is valuable by reducing the degree to which samples are discarded because they are mixed, a frequent limitation for population genetic studies of malaria [52].

Our estimates for population differentiation and structure are in agreement with prior studies. Just as Hupalo et al. identified CV samples to be more closely related to MT samples than to MP samples, our inclusion of a wider set of samples from Thailand and Cambodia

through Parobek et al. and Pearson et al.'s studies suggested a similar story. The large differentiation we observed between Cambodia-Vietnam and Malaysia-PNG is surprising since the two populations are in close geographic proximity. Moreover, we expected to find a larger differentiation between the Cambodia-Vietnam and Myanmar-Thailand populations. Although Cambodia borders Thailand on the west, a malaria-free corridor found in the middle of Thailand has been suggested to reduce the transmission of *P. vivax* and *P. falciparum* isolates between Western Thailand—which is beyond the malaria-free corridor—and Cambodia [53]. Given this limitation to gene flow, we were surprised to find a lack of substantial isolation by distance.

Since we used samples sequenced across several independent experiments, it was important to determine whether sample ascertainment influenced genetic differentiation. However, as depicted in our PCA plot, samples derived from independent experiments within Cambodia showed no major genetic differentiation, supporting our strategy to integrate WGS from independent experiments together. One notable exception were samples from Thailand, which demonstrated a partial degree of differentiation upon PCA. However, we determined that the samples in the Hupalo et al. study were ascertained from a different geographic region of Thailand compared to those sequenced in the Pearson et al. study. Since *P. vivax* in Thailand has been previously demonstrated to carry strong population substructure largely based on geography, we attribute this observed split in our PCA to sample location rather than sequencing bias [54].

Because our study incorporated significantly more samples than used in prior analyses of *P. vivax* WGS data, we found multiple previously-unidentified signatures of selection, many of which were located within functionally important genes. These data could prove valuable as potential targets for vaccine development or strategies toward future antimalarial development.

In the future, we plan to verify some of these candidates by checking their degree of selection in South American populations, since it is likely that they will be selected for there was well based on similar selective pressures.

Our results investigating haplotype architecture suggest that selected alleles have become widespread throughout populations rather than remaining sequestered exclusively within a single population. Furthermore, we determined that the majority of selected alleles likely originated once rather than recurrently. Additionally, the lack of the lack of large haplotype blocks around selected SNPs suggests that the haplotypes associated with selected traits are the product of selection on standing genetic variation rather than recent mutations. However, there are several caveats to our conclusions. Since the UPGMA algorithm assumes constant rates of molecular evolution across all clades, it has been demonstrated to be susceptible to systematic error in forming clades [58]. As such, our results in using that analysis to characterize single versus multiple origins could have been influenced by that same bias. Moreover, none of our analyses were specifically powered to detect the timeline of origin for our variants. Consequently, we plan to incorporate several more experiments to further support our conclusions.

First, we intend to determine the properties of migration and effective population size between our identified populations using the tool *MIGRATE-N* [55]. This analysis will better equip us to understand the extent to which gene flow occurs between our populations and whether the expansion of selected alleles falls in line with or exceeds population-wide expectations. Additionally, we expect to perform experiments comparing the allele frequency spectrums of each population using the tools *Moments* and *δaδi* to determine how recent or ancient demographic events may have influenced the spread of selected alleles, or in the case of selection on standing variation, their spread before selection [56, 57].

Furthermore, we will use the coalescent-based software packages *argweaver* and *RENT+* to infer fine-grained genealogies associated with the selected traits we have identified [59, 60]. Since these programs are equipped to identify small regions of recombination that may go undetected by more mainstream tools, we expect they will be better suited relative to the UPGMA algorithm to determine whether mutations associated with selected alleles have occurred recurrently. Moreover, they may also reveal whether selected alleles were founded on standing genetic variation versus recent *de novo* mutations, based on the characteristics of their associated phylogenetic tree and estimates for the time to most recent common ancestor (TMRCA).

Finally, we believe incorporating samples from South America will improve our capacity to determine the genetic origins and transmission dynamics of selected alleles within Southeast Asia. Since populations of *P. vivax* across South America remain relatively isolated, prior studies have demonstrated that its samples exhibit more extensive linkage disequilibrium and identity-by-descent relative to samples from Southeast Asia [11, 23, 51]. These particular properties have likely given rise to more stable and longer haplotypes for alleles under selection. Thus, populations from South America could provide a valuable model under which to compare haplotype backgrounds for selected alleles within Southeast Asia. Furthermore, samples from South America might offer a window into how selected alleles within Southeast Asian populations might respond to reduced gene flow and parasite transmission across borders.

BIBLIOGRAPHY AND REFERENCES CITED

1. Mendis K, Sina BJ, Marchesini P, Carter R: **The neglected burden of Plasmodium vivax malaria**. *The American journal of tropical medicine and hygiene* 2001, **64**(1_suppl):97-106.
2. Thomson-Luque R, Saliba KS, Kocken CH, Pasini EM: **A Continuous, Long-Term Plasmodium vivax In Vitro Blood-Stage Culture: What Are We Missing?** *Trends in parasitology* 2017, **33**(12):921-924.
3. Kochar DK, Saxena V, Singh N, Kochar SK, Kumar SV, Das A: **Plasmodium vivax malaria**. *Emerging infectious diseases* 2005, **11**(1):132.
4. Anstey NM, Handojo T, Pain MC, Kenangalem E, Tjitra E, Price RN, Maguire GP: **Lung injury in vivax malaria: pathophysiological evidence for pulmonary vascular sequestration and posttreatment alveolar-capillary inflammation**. *The Journal of infectious diseases* 2007, **195**(4):589-596.
5. Alexandre MA, Ferreira CO, Siqueira AM, Magalhães BL, Mourão MPG, Lacerda MV, Alecrim MdGC: **Severe plasmodium vivax malaria, Brazilian Amazon**. *Emerging infectious diseases* 2010, **16**(10):1611.
6. Douglas NM, Anstey NM, Buffet PA, Poespoprodjo JR, Yeo TW, White NJ, Price RN: **The anaemia of Plasmodium vivax malaria**. *Malaria journal* 2012, **11**(1):135.
7. Kute VB, Trivedi HL, Vanikar AV, Shah PR, Gumber MR, Patel HV, Goswami JG, Kanodia KV: **Plasmodium vivax malaria–associated acute kidney injury, India, 2010–2011**. *Emerging infectious diseases* 2012, **18**(5):842.
8. Gething PW, Elyazar IR, Moyes CL, Smith DL, Battle KE, Guerra CA, Patil AP, Tatem AJ, Howes RE, Myers MF: **A long neglected world malaria map: Plasmodium vivax endemicity in 2010**. *PLoS neglected tropical diseases* 2012, **6**(9):e1814.
9. Koepfli C, Rodrigues PT, Antao T, Orjuela-Sánchez P, Van den Eede P, Gamboa D, Van Hong N, Bendezu J, Erhart A, Barnadas C: **Plasmodium vivax diversity and population structure across four continents**. *PLoS neglected tropical diseases* 2015, **9**(6):e0003872.
10. Taylor JE, Pacheco MA, Bacon DJ, Beg MA, Machado RL, Fairhurst RM, Herrera S, Kim J-Y, Menard D, Póvoa MM: **The evolutionary history of Plasmodium vivax as inferred from mitochondrial genomes: parasite genetic diversity in the Americas**. *Molecular biology and evolution* 2013, **30**(9):2050-2064.
11. de Oliveira TC, Rodrigues PT, Menezes MJ, Gonçalves-Lopes RM, Bastos MS, Lima NF, Barbosa S, Gerber AL, de Morais GL, Berná L: **Genome-wide diversity and differentiation in New World populations of the human malaria parasite Plasmodium vivax**. *PLoS neglected tropical diseases* 2017, **11**(7):e0005824.
12. Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amaratunga C, Suon S, Mao S, Noviyanti R, Trimarsanto H: **Genomic analysis of local variation and recent evolution in Plasmodium vivax**. *Nature genetics* 2016, **48**(8):959-964.
13. Dose A: **Guidelines for Treatment of Malaria in the United States**. *women*, **9**(10):11-12.
14. Baird JK: **Resistance to therapies for infection by Plasmodium vivax**. *Clinical Microbiology Reviews* 2009, **22**(3):508-534.

15. Gogtay N, Kannan S, Thatte UM, Olliaro PL, Sinclair D: **Artemisinin-based combination therapy for treating uncomplicated Plasmodium vivax malaria**. *The Cochrane Library* 2013.
16. Alam MT, Vinayak S, Congpuong K, Wongsrichanalai C, Satimai W, Slutsker L, Escalante AA, Barnwell JW, Udhayakumar V: **Tracking origins and spread of sulfadoxine-resistant Plasmodium falciparum dhps alleles in Thailand**. *Antimicrobial agents and chemotherapy* 2011, **55**(1):155-164.
17. Betuela I, Robinson LJ, Hetzel MW, Laman M, Siba PM, Bassat Q, Mueller I: **Primaquine treatment for Plasmodium vivax-an essential tool for malaria control and elimination in Papua New Guinea**. *Papua New Guinea Medical Journal* 2014, **57**(1/4):68.
18. Ashley EA, Dhorda M, Fairhurst RM, Amaratunga C, Lim P, Suon S, Sreng S, Anderson JM, Mao S, Sam B: **Spread of artemisinin resistance in Plasmodium falciparum malaria**. *New England Journal of Medicine* 2014, **371**(5):411-423.
19. Payne D: **Spread of chloroquine resistance in Plasmodium falciparum**. *Parasitology today* 1987, **3**(8):241-246.
20. Parobek CM, Lin JT, Saunders DL, Barnett EJ, Lon C, Lanteri CA, Balasubramanian S, Brazeau N, DeConti DK, Garba DL: **Selective sweep suggests transcriptional regulation may underlie Plasmodium vivax resilience to malaria control measures in Cambodia**. *Proceedings of the National Academy of Sciences* 2016, **113**(50):E8096-E8105.
21. Barry AE, Waltmann A, Koepfli C, Barnadas C, Mueller I: **Uncovering the transmission dynamics of Plasmodium vivax using population genetics**. *Pathogens and global health* 2015, **109**(3):142-152.
22. Anthony TG, Conway DJ, Cox-Singh J, Matusop A, Ratnam S, Shamsul S, Singh B: **Fragmented population structure of Plasmodium falciparum in a region of declining endemicity**. *The Journal of infectious diseases* 2005, **191**(9):1558-1564.
23. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, Vallejo AF, Herrera S, Arévalo-Herrera M, Fan Q: **Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax**. *Nature genetics* 2016, **48**(8):953-958.
24. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D: **Scaling accurate genetic variant discovery to tens of thousands of samples**. *bioRxiv* 2017:201178.
25. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. *arXiv preprint arXiv:13033997* 2013.
26. Auburn S, Böhme U, Steinbiss S, Trimarsanto H, Hostetler J, Sanders M, Gao Q, Nosten F, Newbold CI, Berriman M: **A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes**. *Wellcome open research* 2016, **1**.
27. Morgulis A, Gertz EM, Schäffer AA, Agarwala R: **A fast and symmetric DUST implementation to mask low-complexity DNA sequences**. *Journal of Computational Biology* 2006, **13**(5):1028-1040.
28. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O'Brien J, Djimde A, Doumbo O, Zongo I: **Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing**. *Nature* 2012, **487**(7407):375.

29. Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE: **hmmIBD: software to infer pairwise identity by descent between haploid genotypes**. *bioRxiv* 2017:188078.
30. Zhu SJ, Almagro-Garcia J, McVean G: **Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data**. *Bioinformatics* 2017, **34**(1):9-15.
31. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS: **A high-performance computing toolset for relatedness and principal component analysis of SNP data**. *Bioinformatics* 2012, **28**(24):3326-3328.
32. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**(2):945-959.
33. Cook DE, Andersen EC: **VCF-kit: assorted utilities for the variant call format**. *Bioinformatics* 2017, **33**(10):1581-1582.
34. Miles AH, N.: **scikit-allel - Explore and analyse genetic variation. 0.21.2 ed.** 2018.
35. Lunter G, Marth G, Sherry S: **The variant call format and VCFtools**. *Bioinformatics* 2011, **378**.
36. [https://github.com/BGI-shenzhen/PopLDdecay]
37. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2004, **21**(2):263-265.
38. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets**. *Gigascience* 2015, **4**(1):7.
39. Szpiech ZA, Hernandez RD: **selscan: an efficient multithreaded program to perform EHH-based scans for positive selection**. *Molecular biology and evolution* 2014, **31**(10):2824-2827.
40. Garud NR, Messer PW, Buzbas EO, Petrov DA: **Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps**. *PLoS genetics* 2015, **11**(2):e1005004.
41. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R: **On detecting incomplete soft or hard selective sweeps using haplotype structure**. *Molecular biology and evolution* 2014, **31**(5):1275-1291.
42. Clement M, Posada D, Crandall KA: **TCS: a computer program to estimate gene genealogies**. *Molecular ecology* 2000, **9**(10):1657-1659.
43. Leigh JW, Bryant D: **popart: full-feature software for haplotype network construction**. *Methods in Ecology and Evolution* 2015, **6**(9):1110-1116.
44. Cagan A, Blass T: **Identification of genomic variants putatively targeted by selection during dog domestication**. *BMC evolutionary biology* 2016, **16**(1):10.
45. Gusev A, Kenny EE, Lowe JK, Salit J, Saxena R, Kathiresan S, Altshuler DM, Friedman JM, Breslow JL, Pe'er I: **DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation**. *The American Journal of Human Genetics* 2011, **88**(6):706-717.
46. Schliep K, Paradis E, de Oliveira Martins L, Potts A, White TW, Stachniss C, Kendall M: **Package 'phangorn'**. 2018.
47. Jex A, Mueller I, Kappe S, Mikolajcjak S, Sattabongkot J, Patrapuvich R, Lindner S, Flannery E, Koepfli C, Ansell B: **Integrated transcriptomic, proteomic and epigenomic analysis of Plasmodium vivax salivary-gland sporozoites**. *bioRxiv* 2017:145250.

48. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT: **BamTools: a C++ API and toolkit for analyzing and managing BAM files**. *Bioinformatics* 2011, **27**(12):1691-1692.
49. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS**. *Nature protocols* 2012, **7**(9):1728.
50. Balu B, Maher SP, Pance A, Chauhan C, Naumov AV, Andrews RM, Ellis PD, Khan SM, Lin J-w, Janse CJ: **CCR4-associated factor 1 coordinates the expression of Plasmodium falciparum egress and invasion proteins**. *Eukaryotic cell* 2011, **10**(9):1257-1263.
51. Lo E, Lam N, Hemming-Schroeder E, Nguyen J, Zhou G, Lee M-C, Yang Z, Cui L, Yan G: **Frequent Spread of Plasmodium vivax Malaria Maintains High Genetic Diversity at the Myanmar-China Border, Without Distance and Landscape Barriers**. *The Journal of infectious diseases* 2017, **216**(10):1254-1263.
52. Trevino SG, Nkhoma SC, Nair S, Daniel BJ, Moncada K, Khoswe S, Banda RL, Nosten F, Cheeseman IH: **High-resolution single-cell sequencing of malaria parasites**. *Genome biology and evolution* 2017, **9**(12):3373-3383.
53. Parker DM, Carrara VI, Pukrittayakamee S, McGready R, Nosten FH: **Malaria ecology along the Thailand–Myanmar border**. *Malaria journal* 2015, **14**(1):388.
54. Kittichai V, Koepfli C, Nguitragool W, Sattabongkot J, Cui L: **Substantial population structure of Plasmodium vivax in Thailand facilitates identification of the sources of residual transmission**. *PLoS neglected tropical diseases* 2017, **11**(10):e0005930.
55. Beerli P: **MIGRATE-N: estimation of population sizes and gene flow using the coalescent**. *Available at popgen sc fsu edu/Migrate/Download html* 2008.
56. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD: **Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data**. *PLoS genetics* 2009, **5**(10):e1000695.
57. Jouganous J, Long W, Ragsdale AP, Gravel S: **Inferring the joint demographic history of multiple populations: beyond the diffusion approximation**. *Genetics* 2017:genetics. 117.200493.
58. Husmeier D, Dybowski R, Roberts S: **Probabilistic modeling in bioinformatics and medical informatics**: Springer Science & Business Media; 2006.
59. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A: **Genome-wide inference of ancestral recombination graphs**. *PLoS genetics* 2014, **10**(5):e1004342.
60. Mirzaei S, Wu Y: **RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination**. *Bioinformatics* 2016, **33**(7):1021-1030.