Detecting Selection Signals In *Plasmodium falciparum* Using Identity-by-Descent Analysis

Lyndal Henden[1,2], Stuart Lee[1,2], Ivo Mueller[1,2], Alyssa Barry[1,2], Melanie Bahlo[1,2,*]

1. Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville VIC 3052, Australia
2. Department of Medical Biology, University of Melbourne, Parkville VIC 3010, Australia

* Corresponding author
E-mail: bahlo@wehi.edu.au (MB)

Abstract

Detection of selection signals using genomic data is vital for tracking drug resistance loci in microorganisms that cause disease, such as malaria, and monitoring of such loci is crucial for disease control efforts. Here we present a novel method of detecting relatively recent selection using identity-by-descent approaches suitable for multiclonal, recombining microorganisms. Application of this new method to a large whole genome sequencing study of *Plasmodium falciparum* identifies many well-known signatures such as *crt* and *k13*, associated with chloroquine resistance and artemisinin resistance respectively, and through relatedness networks shows how these signatures are distributed in Southeast Asia, Africa and Oceania. Using these networks, we confirmed an independent origin of chloroquine resistance in Papua New Guinea and the spread of multiple artemisinin resistance mutations in Southeast Asia. We also found two novel signals of selection not yet investigated in detail.

Introduction

Antimicrobial drugs have been largely successful in reducing the global burden of diseases such as malaria, HIV and tuberculosis, which are responsible for hundreds of thousands of deaths and hundred of millions of clinical cases annually (World Health Organization, 2015). Despite their success, combinations of prolonged drug use, high levels of genetic diversity and high mutation rates have enabled these, and other microorganisms, to rapidly develop resistance to first-line antimicrobial drugs (Neafsey et al. 2008; Talisuna et al. 2004, zur Wiesch et al. 2011). Adding to concerns is the emergence of multi-drug resistance, whereby microorganisms are resistant to a number of drugs, further limiting treatment options resulting in increased morbidity and mortality, hindering advances made in combating the diseases they cause. Antimicrobial drug resistance poses an enormous threat to the effectiveness of treatments and prevention regimes, and identifying the genetic mechanisms underlying such resistance is essential for monitoring and controlling the spread of resistance as well as rational use of existing and development of novel antimicrobial treatments.

Resistance to antimicrobial drugs occurs when a microorganism develops either single or multiple variants, such as point mutations, copy number variations or chromosomal rearrangements that reduce the organism's sensitivity to one or more drugs used to treat infections. These variants introduce strong genetic signatures into the organism's genome that are reflective of natural selection (Mu et al. 2010). For recently selected variants arising in recombining organisms these signatures are characterized by unusually long haplotypes that have risen to high frequency within the population. The long haplotypes arise because genomically neighboring variants hitchhike with the variant being selected, due to a lack of exposure to recombination over the short period of time that the haplotype has been under selection. Many of the existing methods for detecting signatures of positive selection exploit these long haplotypes using statistical methods such as the extended haplotype homozygosity

(EHH) test, which calculates the probability that two randomly selected chromosomes have identical haplotypes adjoining an identical core haplotype (Sabeti et al. 2002; Voight et al. 2006; Sabeti et al. 2007). Such methods can detect signatures of selection due to novel variants with high power, however have limited power to detect selection on standing variation. This is due to multiple founder haplotypes carrying the variant at similar frequencies in the population and longer exposure to recombination events leading to smaller haplotypes (Albrechtsen et al. 2010; Vitti et al. 2013). The ability to detect selection on standing variation is a desirable property when investigating drug resistance as multiple instances of such selection have been identified (Nair et al. 2007; MalariaGEN *Plasmodium falciparum* Community Project, 2015).

An alternative method to identify signatures of selection, including selection on standing variation, is to detect regions of the genome that have been inherited from a common ancestor (Albrechtsen et al. 2010). Such regions are said to be identical by descent (IBD) and genomic regions with many shared haplotypes provide evidence of recent positive selection. IBD analyses investigating signatures of selection have been successfully applied to human studies, identifying the HLA region which regulates the human immune system (Albrechtsen et al. 2010), and LCT gene, enabling digestion of lactase post-infancy (Han and Abney 2013). There are currently no reported attempts of IBD analyses performed on microorganisms, such as *Plasmodium*, to identify recent positive selection. This is in part due to the surprising lack of methodologies for haploid species (Henden et al. 2016), in addition to the complexity of infections often containing more than one genetically distinct strain.

Unlike allele sharing based methods such as EHH and XP-EHH (Sabeti et al. 2002; Sabeti et al. 2007), IBD mapping of microorganisms can also be used to infer fine-scale population structure and allows the ability to monitor disease control and transmission, as well as to determine if an antimicrobial drug-resistant haplotype has spread or arisen independently between geographic locations. Furthermore,

IBD mapping has the potential to uncover multidrug resistance and, for diseases that experience relapse infections such as malaria caused by *Plasmodium vivax*, may be able to distinguish between new or relapsing infections in drug efficacy and cohort studies.

Here we introduce isoRelate, a freely available R package that performs IBD mapping on recombining haploid species, such as the malaria-causing parasite *Plasmodium* and the bacterium *Staphylococcus aureus*, that also allows for multiple infections. Our methodology is based on earlier work for the human male X chromosome (Henden et al. 2016) and implements a first order hidden Markov model to detect relatedness between pairs of isolates, where an isolate refers the microorganisms extracted from an infection. The model uses unphased genotype data from single nucleotide polymorphisms (SNPs), which can be obtained from either array data or sequencing data that randomly samples SNP variation throughout the genome.

Using isoRelate, we demonstrate the ability of IBD analyses to detect signals of recent positive selection using whole genome sequencing (WGS) data for a previously published global *Plasmodium falciparum* dataset of 2,550 isolates (MalariaGEN *Plasmodium falciparum* Community Project, 2016). We make comparisons with other popular methodologies that also try to detect recent positive selection. Additionally, we use isoRelate to explore *P. falciparum* population structure between geographical regions; confirm the global spread of resistance to the antimalarial drug chloroquine as well as explore resistance to artemisinin as a soft selective sweep, and investigate the ability of IBD to detect multidrug resistance.

Results

Validation of isoRelate

We validated our methodology by applying isoRelate to the MalariaGEN Pf3k genetic cross dataset (Miles et al. 2016) to detect known recombination events. This dataset contains the parents and offspring of three *P. falciparum* strain crosses; 3D7 x HB3, 7G8 x GB4, and HB3 x Dd2. There are 21, 40 and 37 isolates for the three crosses respectively, and 11,612 SNPs, 10,903 SNPs and 10,637 informative SNPs remaining following filtering procedures (Supplementary Table 1). We combined the results for all three crosses and found that isoRelate detected 98% of all reported IBD segments, with an average concordance between inferred and reported segments of 99%. Additionally, isoRelate detected segments with 99% accuracy; meaning only 1% of segments were likely to be false positives. We did not infer IBD between any of the founders. This is expected given the documented origins of these three strains, which were derived from very different geographic regions (Manske et al 2012). False negatives, where IBD was not inferred between parents and offspring, were observed predominantly in genomic regions located between recombination events. Moreover, identical segment boundaries were detected between all replicate isolates. We note that our methodology has been extensively tested on simulated data for the human X chromosome and as such we have not performed simulation studies here (Henden et al. 2016).

Population analysis of *P. falciparum*

To demonstrate the ability of isoRelate to investigate a haploid species with known selection signals, we performed IBD mapping of 2,550 *P. falciparum* isolates from 14 countries across Africa, Southeast Asia and Papua New Guinea as part of the MalariaGEN Pf3K dataset. The samples in this dataset were collected during the years 2001 to 2014 (Supplementary Table 2) and details of the collection process

and sequencing protocols have been described elsewhere (Manske et al. 2012; MalariaGEN *Plasmodium falciparum* Community Project 2016). We define within-country analyses as all pairwise IBD comparisons between isolates from the same country (14 analyses in total) while between-country analyses as all pairwise-country comparisons (91 analyses in total) where pairs of isolates contain one isolate from each country.

After all filtering procedures were complete, 2,377 isolates remained for analysis with 994 isolates (42%) classified as having multiple infections (Supplementary Tables 2 and 3). The mean number of SNPs remaining post filtering for within-country analyses was 31,018 SNPs with the least number of SNPs in the analysis of Papua New Guinea (18,270 SNPs) and the largest number of SNPs in the analysis of Guinea (44,528 SNPs) (Supplementary Table 2). SNPs for between-country analyses were selected if they appeared in both countries at similar frequencies, which resulted in an average of 12,271 SNPs per analysis with the smallest number of SNPs in the analysis between Mali and Papua New Guinea (1,945 SNPs) and the largest number of SNPs in the analysis between Guinea and Malawi (29,138 SNPs) (Supplementary Table 4). These highly varying numbers of informative SNPs largely reflect geographical isolation and distance but are also influenced by the quality of the WGS data with poorer quality sequencing leading to fewer SNPs. Analyses with so few SNPs, such as Mali and Papua New Guinea, are unlikely to detect selection signatures since smaller IBD segments will fail to be detected, however are still useful for identifying closely related isolates that are expected to share large IBD segments over many SNPs.

Investigating levels of relatedness

We calculated the proportion of pairs IBD at each SNP and investigated the distributions of these statistics across the genome (Figure 1, Supplementary Table 5, Supplementary Figure 1). We identified higher levels of relatedness in Southeast Asia than in Africa or in Papua New Guinea, with isolates from Cambodia displaying

the highest average sharing across the genome (5%). The Cambodian dataset consists of isolates collected from four study locations; therefore we stratified the relatedness proportions by study location to identify sites with extremely high amounts of relatedness. We detected high relatedness between 87% (2,890/3,321) of pairs from the Pailin Province of Cambodia, with on average 29% of pairs IBD per SNP (Supplementary Tables 6 and 7, Supplementary Figures 2 and 3). Isolates from Pailin make up 16% of the Cambodian dataset and inflate the overall signal seen in Cambodia. We also detected high amounts of relatedness, including many clonal isolates, in the Thai Province of Sisakhet, which borders Cambodia, reflecting similar transmission dynamics between regions in close proximity.

Relatedness proportions can also be used to identify genomic regions with particularly high amounts of sharing that may be under positive selection as previously shown for IBD studies in human populations (Figure 1) (Albrechtsen et al. 2010; Han and Abney 2013). We observe higher levels of relatedness over several known *P. falciparum* antimalarial drug resistance genes such as *Pfcrt* (chloroquine resistance transporter) and *Pfdhfr* (dihydrofolate reductase) in addition to several regions suspected of being associated with antimalarial drug resistance. In particular, a large proportion of sharing occurs towards the right telomere of chromosome 6, which contains a number of promising candidate genes suspected of being associated with pyrimethamine resistance (Amambua-Ngwa et al. 2012; Park et al. 2012). Many of these signals also show substantial continent and/or country variation (Figure 1).

Relatedness-networks can be created using clustering techniques to identify groups of isolates sharing a common haplotype. We constructed a relatedness-network to investigate clusters of isolates sharing near-identical genomes, reflecting identical infections or 'duplicate' samples (Figure 2). Southeast Asia has a number of large clusters containing highly related isolates with the five largest clusters belonging to Cambodia, containing between 12 and 68 isolates, indicative of clonal expansions. The largest cluster contains mostly isolates from the Pursat Province of Cambodia,

however the remaining isolates are from the Pailin Province and the Ratanakiri Province of Cambodia, suggesting common haplotypes between western and eastern Cambodia. In contrast, we did not find any isolates within Guinea or Mali to be highly related, nor did we find isolates from different countries to be highly related (Supplementary Table 8, Supplementary Figures 4 - 9).

Some clusters would separate into multiple disjoint clusters if even a single isolate were removed from the group. Isolates which, if removed, would result in disjoint clusters were generally observed to have MOI > 1, where their genome data consists of at least two genetically distinct haplotypes. Such isolates have potentially come from individuals who are travelling between geographical locations and become infected with *P. falciparum* strains unique to those regions, resulting in IBD that connects multiple, otherwise unconnected sub clusters of isolates.

Analysis of selection signals over the chloroquine resistance locus, *Pfcrt*

To assess the significance of a selection signature we transformed the IBD results for each analysis to account for variations in relatedness between isolates and SNP allele frequencies, then performed normalization allowing us to calculate a new summary metric for each SNP, $-\log_{10}$ P-values. The genome-wide distributions of the $-\log_{10}$ P-values for within-country analyses are shown in Figure 3 and the top five signals of selection for each country are reported in Supplementary Table 9. We examined in detail the selection signals overlapping the known *P. falciparum* chloroquine resistance transporter gene, *Pfcrt*, located on chromosome 7 at 403,222-406,317 (Figure 4).

All countries except Malawi and Myanmar have at least one significant SNP within 12kb of *Pfcrt* based on a 5% genome-wide significance threshold. Malawi withdrew the use of chloroquine as an antimalarial drug in 1993, which resulted in the disappearance of the molecular marker of chloroquine resistance (K76T mutation) in Malawian *P. falciparum* populations (Laufer et al. 2006). Thus we would not

expect to see a signature of selection over *Pfcrt* in Malawi. Additionally, none of the between-country analyses involving isolates from Malawi reach significance within 60kb of the *Pfcrt* locus.

Surprisingly, an increase in IBD proportions is observed over *Pfcrt* in Myanmar however the closest significant SNP is located 45kb downstream of *Pfcrt*. In contrast, little to no increase in IBD is observed in the region surrounding *Pfcrt* in Cambodia and Laos, although significant SNPs are identified within close proximity to *Pfcrt*. Both Cambodia and Laos have many isolates sharing large proportions of their genome IBD; potentially adding noise to the summary statistics resulting in inflated significance.

In most countries the highest proportion of IBD on chromosome 7 occurs downstream of *Pfcrt*, with higher levels of IBD extending further downstream of *Pfcrt* than upstream, including over a known set of *var* genes, which were excluded from the IBD analysis due to their complex genetic structure which leads to significant mapping problems. This potentially indicates that *Pfcrt* is regulating a gene downstream or alternatively a second region in close proximity to *Pfcrt* is under selection. A secondary signal immediately downstream of the *var* genes cluster on chromosome 7 has been previously identified in isolates sampled from The Gambia (Nwakanma et at. 2014).

We investigated relatedness over *Pfcrt* between isolates from different countries and confirmed the spread of chloroquine resistance throughout Southeast Asia and Africa, while also confirming an independent origin of chloroquine resistance in Papua New Guinea (Figure 5) (Mehlotra et al. 2001; Wootton et al. 2002). However we were unable to determine the exact haplotypes at codons 72-76 of the *Pcfrt* gene, of which CVIET and SVMNT have been associated with chloroquine resistance (Mehlotra et al. 2001; Wootton et al. 2002), due to low quality data resulting in missing genotype calls for many isolates in addition to unknown haplotype phase for MOI>1 isolates.

In particular the largest cluster in Figure 3 contains 48% of all isolates, of which 78% have missing genotype calls at codons 73-75 collectively. All isolates in this cluster have the wild type C allele at the C72S variant codon 72. Additionally 95% of these isolates have the chloroquine resistant K76T mutation (codon 76). Thus we speculate the dominant haplotype in the largest cluster to be CVIET, which arose in Southeast Asia and spread to Africa (Wootton et al. 2002). All isolates from Papua New Guinea have the C72S mutation and K76T mutation (and missing genotype calls at codons 73-75) consistent with the presence of the SVMNT haplotype (Mehlotra et al. 2001).

Analysis of selection signals over the artemisinin resistance locus, *Pfk13*

Parasite resistance to the antimalarial drug artemisinin has been associated with mutations in the *P. falciparum* kelch 13 gene, *k13*, located on chromosome 13 at 1,724,817-1,726,997 (Miotto et. al. 2013; Ariey et al. 2014). We detected selection signals of marginal significance over *Pfk13* in Cambodia and Thailand (Figure 3), which is not surprising given that artermisin resistance has only recently been identified in Cambodia in 2007 and is currently confined to Southeast Asia (Maude et al. 2014). Given the samples from Cambodia and Thailand were collected between 2009 to 2013 (Supplementary Table 2), the resistance mutations are expected to be at low frequencies within these populations, producing very weak signals of selection.

Artemisinin resistance has arisen as a soft selective sweep, involving at least 20 independent *Pfk13* mutations (MalariaGEN *Plasmodium falciparum* Community Project). Relatedness networks over *Pfk13* identify many disjoint clusters of related isolates, with at least 9 clusters containing isolates that carry the most common mutation associated with artemisinin resistance, C580Y (MalariaGEN *Plasmodium falciparum* Community Project) (Figure 6). We identified isolates from Cambodia, Thailand and Vietnam as carriers of this mutation at frequencies of 40%, 26% and

1% respectively. Additionally, relatedness is detected between isolates from Cambodia and Thailand that have the C580Y mutation as well as isolates from Cambodia and Vietnam with this mutation, suggesting that some resistance-haplotypes have swept between countries (Takala-Harrison et al. 2015).

Investigating global inheritance of genomic locations

We investigated the IBD analyses results between countries to determine if any other genomic locations had experienced a global spread like that of chloroquine. We identified a signal on chromosome 6 as having done so, not only between Africa and Southeast Asia, but also Papua New Guinea. In fact, significant IBD sharing is detected in all pairwise-country analyses over the interval chr6: 1,102,005-1,283,312. This interval contains 32 genes of which several have been identified as promising drug resistance candidates (Amambua-Ngwa et al. 2012; Park et al. 2012, Amambua-Ngwa et al. 2016). The cause of this selection pressure remains unknown.

Detection of multidrug resistance from selection signatures

We explored selection signatures to determine if multidrug resistance could be identified. Specifically, we investigated the *P. falciparum* multidrug resistance gene 1 (*Pfmdr1),* which has been associated with chloroquine resistance and amodiaquine resistance when the *Pfmdr1* N86Y mutation is present along with the *Pfcrt* K76T mutation (Veiga et al. 2016). Figure 7 displays genome-wide selection signals in Ghana, stratified by pairs who are IBD over *Pfmdr1* and pairs who are not IBD over *Pfmdr1*. A significant signal of selection is observed over *Pfcrt* in both stratified groups, suggesting *Pfcrt* is under selection jointly with *Pfmdr1* as well as independently of *Pfmdr1*. Of the isolate pairs who are IBD over *Pfmdr1*, 13% are also IBD over *Pfcrt* while 6% are IBD over *Pfcrt* and carry both the N86Y mutation and the K76T mutation. The median proportion of genome inferred IBD between these pairs is 1%, alleviating concerns that joint inheritance of both variants is due to highly related pairs. An additional selection signal is identified over *Pfglurp*

(glutamine-rich protein, a candidate vaccine antigen) also suggesting joint selection of both *Pfmdr1* and *Pfglurp*.

Analysis of selection signal methodologies

We compared the selection signatures generated by isoRelate within countries to those detected by the integrated haplotype score (iHS), an algorithm that makes use of the EHH, designed to identify strong signals of recent positive selection (Voight et al. 2006). The EHH algorithm requires knowledge of haplotype phase, which is currently not possible for isolates with MOI > 1 as the number of strains in an infection and the proportions they contribute to the mixed infection must be known, in addition to having quality data sequenced at high coverage. In contrast, haplotype phase is trivial for isolates with MOI = 1, therefore we performed comparisons of isoRelate and iHS using only isolates with MOI = 1, on the same SNPs (Supplementary Tables 10, Supplementary Figure 10). The largest $-\log_{10}$ P-value for a single SNP within each of 12 interesting genes is reported in Supplementary Table 11.

Although there is some overlap in the selection signatures produced by iHS and isoRelate, there is a surprising dissimilarity between the results. iHS detects selection at *Pfglurp*, *Pfama1* and *Pftrap* more frequently than isoRelate, however has difficulty detecting selection in Southeast Asia and Papua New Guinea. In contrast isoRelate commonly detects selection over *Pfdhfr*, *Pfmdr1*, *Pfcrt* and *Pfdhps*. Additionally prominent signals are also detected on chromosome 6 and chromosome 12 by isoRelate, in regions that, as yet, have no reported candidate genes.

The genes *Pfglurp*, *Pfama1* and *Pftrap* detected by iHS encode surface proteins that undergo balancing selection (Ochola-Oyier et al. 2016; Ohashi et al. 2014) and hence have been investigated as vaccine targets. We anticipate selection on extremely recent mutations in these genes that are at low frequency within the population, in

which case iHS is more likely to detect this selection than isoRelate. This is simply because iHS profiles are calculated relative to the number of isolates in a population while isoRelate profiles are calculated relative to the number of pairwise combinations in the population, which heavily dilutes excess IBD sharing of low frequency haplotypes.

Additionally, iHS assumes that all samples are independent, meaning there is no relatedness between isolates. This assumption is violated in all countries, particularly in Southeast Asia where there are many highly related isolates, preventing iHS from decaying to a threshold at some SNPs, resulting in missing iHS values. On average 84% of SNPs in African countries have missing iHS values, while 94% of SNPs in Southeast Asian countries have missing values, contributing to the lack of signals detected in Asia. To avoid such loss of information, related isolates could be removed from iHS analyses at the risk of reduced power due to smaller sample sizes. However in some instances the sample size would reduce significantly, as is the case with Cambodia, which would experience an 80% reduction in sample size if isolates sharing more than 10% of their genome IBD were removed. Considering the number of SNPs with missing values, iHS does surprisingly well in African countries.

Similarly, we examined the selection signatures generated by isoRelate between countries and those of the cross-population EHH (XP-EHH) methodology, which compares the integrated EHH profiles between two populations at the same SNP (Sabeti et al. 2007), using isolates with MOI = 1 (Supplementary Table 12). XP-EHH is designed to detect selection that is near fixation in one population but not in the other, while isoRelate will detect genomic regions where at least some haplotypes under selection are shared between the two populations. The two methods treat the data in fundamentally different ways, therefore signatures detected by XP-EHH and isoRelate reflect different signals and are difficult to compare. We performed an artificial analysis between XP-EHH and isoRelate, combining isoRelate results from

with-country analyses and between-country analyses. Details are provided in the Supplementary Material.

Discussion

Relatedness mapping of microorganisms is extremely useful for investigating the genetic mechanisms involved in diseases. We demonstrate this on a global whole-genome sequenced *P. falciparum* dataset using a new IBD methodology, isoRelate, which allows for novel insights into the geographical spread of antimalarial drug resistance, including multidrug resistance, as well as population structure.

IBD inference of *P. falciparum* genomes allows us to compare different levels of relatedness between geographical regions. Here we identified the Pailin Province of Cambodia as having many highly related isolates, either as a result of intensified malaria control efforts following the emergence of artemisinin resistance in 2007 (Maude et al. 2014) or as an artifact of the sampling collection procedures, in which case greater efforts may need to be made to attain independence for population genetic studies. As such, we propose genome wide IBD summaries as a means of monitoring malaria control, whereby intensified control regimes reduce malaria transmission and genetic diversity (Anderson et al. 2000; Daniels et al. 2015), resulting in more relatedness between strains and higher proportions of IBD.

Our algorithm allows us to infer IBD status at any genomic location, which lead us to develop a new summary measure of IBD sharing in populations at genomic locations, resulting in a novel measure for detecting selection. We developed a statistical framework to test the significance of selection signatures, which, unlike iHS and XP-EHH, accounts for the level of relatedness between isolates. Using the IBD approach we were able to identify both known resistance loci, underpinned by known resistance genes, including *Pfcrt* and *Pfk13*, and several novel signals of selection, one of which has been previously reported on chromosome 6 (Amambua-

Ngwa et al. 2012; Park et al. 2012, Amambua-Ngwa et al. 2016). Quantifying relatedness is important in analyses wishing to investigate selection, as highly related isolates add noise to the results, making it harder to identify selection signatures. Ideally related isolates would be excluded from analyses, however as disease control reduces transmission, highly related isolates will become prominent (Daniels et al. 2015) and removing these isolates could greatly reduce the power of the analysis.

We generated relatedness networks to provide insights into the number of haplotypes within a genomic interval as well as their origin, which has immediate applications for monitoring the geographic spread of antimicrobial drug resistant haplotypes. We visualized the spread of chloroquine resistance across Southeast Asia and Africa using such networks, confirming an independent origin of resistance in Papua New Guinea (Mehlotra et al. 2001; Wootton et al. 2002). We also examined relatedness over *Pfk13* and were able to visualize a number of founder haplotypes carrying the C580Y mutation, associated with artemisinin resistance, also confirming that resistance to artemisinin has arisen as a soft selective sweep (MalariaGEN *Plasmodium falciparum* Community Project 2016).

IBD analyses require several criteria to be met. This includes the availability of a good quality reference genome and the fact that the organism must recombine as one of its main sources of creating genetic variation. As such these methods do not appear to be applicable to *Mycobacterium tuberculosis* for example, but will work with any other organism that shares these criteria with *P. falciparum*. Amongst these are *P. vivax* (Carlton et al. 2008) and some species of *Staphylococcus* (Feil et al. 2001). Thus isoRelate will have broader application than just *P. falciparum*. Furthermore, isoRelate can be applied to any dense genomic data that produces SNP genotypes, which includes WGS, RNA sequencing and SNP arrays.

In summary, isoRelate is the first algorithm to implement an IBD-based selection detection approach applicable for field isolates with possible multi-clonality. We

have shown that our approach can dissect complex signals of selection, including selection on standing variation. This method will be invaluable for the identification and genomic surveillance of drug resistance loci in many microorganisms.

Materials and Methods

Data processing

MalariaGEN genetic crosses dataset

To validate our method's ability to recapitulate recombination events and thus IBD sharing we made use of a previously published *P. falciparum* genetic cross. Whole genome sequencing (WGS) data was retrieved for 98 *P. falciparum* lab isolates that were generated as part of the MalariaGEN consortium Pf3k project (Miles et al., 2016). This dataset included the parent and progeny (first generation) of crosses between the pairs of parent clones 3D7 and HB3, 7G8 and GB4, and HB3 and Dd2. We retrieved all available Pf3k data in VCF file format from data release 5 (https://www.malariagen.net/data/pf3k-5). SNPs were excluded if they were not in a 'core' region of the genome (Miles et al., 2016), or if they had QD ≤ 15 or MQ ≤ 50, or if less than 90% of samples were not covered by at least 5 reads, or they were not polymorphic or if their MAF was less than 1% (using a read depth estimator). Samples were also excluded if less than 90% of their SNPs were not covered by at least 5 reads. Supplementary Table 1 shows the number of isolates and SNPs before and after filtering of each genetic cross.

We visualized parental recombination breakpoints in the progeny's haplotypes using the GATK genotype data with default settings in the online app (https://www.malariagen.net/apps/pf-crosses/1.0/). This allowed us to produced 'truth' IBD datasets with known recombination events. We then assessed isoRelate's inferred IBD segment locations against this dataset.

MalariaGEN global P. falciparum dataset

WGS was performed on 2,512 *P. falciparum* field isolates sampled from 14 countries across Africa and Southeast Asia as part of the MalariaGEN consortium Pf3k project (Manske et al. 2012; MalariaGEN *Plasmodium falciparum* Community Project 2016). We retrieved all available Pf3k data in VCF file format from release 5. We merged all nuclear chromosome VCF files and applied filters to the 2,512 samples and 1,057,870 biallelic SNPs.

Variants were filtered using GATK's SelectVariants and VariantFiltration modules (DePristo et al., 2011). SNPs were excluded if there were more than 3 SNPs within a 30 base pair window, or if they were not in a 'core' region of the genome, or if they had Variant Quality Score Recalibration (VQSR) < 0. Moreover, to reduce the possibility of spurious SNP calls further filters for Quality of Depth (QD), Strand Odds Ratio (SOR), Mapping Quality (MQ) and MQ Rank Sum (MQRankSum) were applied (QD > 15, SOR < 1, MQ > 50, MQRankSum > -2). This filtering left 561,695 SNPs in the dataset.

Next, separating the data by country of origin, SNPs were excluded if less than 90% of samples were not covered by at least 5 reads or they were not polymorphic. Samples were also excluded if less than 90% of their SNPs were not covered by at least 5 reads. Following this, countries were grouped into broader geographical regions of West Africa, Central Africa or Southeast Asia, and the intersection of SNPs within a region was taken. Lastly, within each country, SNPs with minor allele frequencies (MAF) less than 1% (using read depths) were removed. Supplementary Table 2 displays the number of isolates and SNPs before and after filtering of each country. Nigeria was excluded from all downstream analyses due to the low number of SNPs remaining after filtering.

Papua New Guinea dataset

WGS data was available for 38 *P. falciparum* isolates from Madang, Papua New Guinea (PNG), sampled in 2007 and sequenced at the Wellcome Trust Sanger Institute (WTSI), Hinxton, UK as part of the MalariaGEN consortium (http: //www.malariagen.net/about; study ID: 1021-PF-PG-MUELLER). The sequencing data was processed by replicating the analysis processing steps of the MalariaGen Pf3k field isolates for compatibility (Supplementary material).

Assessing clonality and extracting data for IBD analysis

We applied the F_{ws} metric to within-country SNP sets to determine isolates that had multiple infections (Manske et al. 2012). An isolate was classified as having multiple infections if $F_{ws} < 0.95$. For each country PED and MAP files for downstream analysis were extracted using moimix (Lee, 2016). Heterozygous SNP calls were retained for isolates assigned as having multiplicity of infection (MOI) greater than 1, otherwise heterozygous SNPs were set to having a missing value at those SNPs to signify the likelihood of a genotyping error.

Detecting relatedness between isolates

We extend a first order hidden Markov model (HMM) that detects IBD segments between pairs of human samples to allow detection of IBD between pairs of non-human, haploid samples (Henden et al, 2016). The assumption of a first order HMM is unlikely to hold in the presence of dense datasets containing linkage disequilibrium, however we do not consider this to be an issue with *P. falciparum* due to the short LD segments in its genome (Mu et al. 2005; Volkman et al. 2012).

Genotype calls are used to determine the number of alleles shared IBD at each SNP between a pair of isolates. The potential number of shared alleles at a SNP defines the state space in the HMM and is dependent on the MOI of the pair under

consideration. An isolate with MOI = 1 consists of a single strain and is analyzed as if it were haploid; thus sharing either 0 or 1 allele IBD with any other isolate. An isolate with MOI > 1 consists of multiple genetically distinct (and possibly related) strains, and is considered diploid; sharing 0, 1 or at most 2 alleles IBD with other isolates. Here we make the assumption that an isolate with MOI > 1 actually has MOI = 2, arguing that the current coverage of WGS data struggles to identify more than two clones contributing to an isolate. This assumption will be incorrect for some isolates; however, the progress of malaria control efforts has lead to a decrease in the number of multiple infections, with the majority of multiple infections consisting of two strains (Galinsky et al., 2015).

Initial probabilities, emission probabilities and transition probabilities are calculated as in Henden et al. (2016) and are described in the Supplementary material. Both the initial probabilities and the emission probabilities require population allele frequencies and we compute these frequencies for each country separately for *P. falciparum*. This is necessary due to the highly divergent sets of SNPs observed in *P. falciparum* globally (Neafsey et al. 2008). To perform IBD analyses between isolate from different countries, SNPs were included in the analysis if the population allele frequencies between the pair of countries differed by less than 0.3. Population allele frequencies for the combined countries were then calculated using all isolates from pairs of countries being examined. SNPs with MAF less than 1% were removed from the analysis along with SNPs with missing genotype data for more than 10% of isolates. Similarly, isolates with missing genotype data for more than 10% of SNPs were removed and a genotyping error rate of 1% was included in the model. Supplementary Tables 2 and 4 give the number of isolates and SNPs before and after filtering for each country and pairwise-country dataset.

IBD segments are reported based on the results from the Viterbi algorithm (Rabiner, 1989) and segments that contain less than 20 SNPs or have lengths less than 50,000bp are excluded, as they are likely to represent distant population sharing

that is not relevant to recent selection. IBD analyses were performed between all pairs of isolates that remained once filtering procedures had been applied.

The algorithm has been developed as an R package, isoRelate, and can be downloaded from https://github.com/bahlolab/isoRelate.

Identifying selection signals and assessing significance from IBD

We created a matrix of binary IBD status with rows corresponding to SNPs and columns corresponding to isolate pairs. For each column, we subtract the column mean from all rows to account for the amount of relatedness between each pair. Following this we subtract the row mean from each row and divide by the square root of $p_i(1-p_i)$, where p_i is the population allele frequency of SNP i. This adjusts for differences in SNP allele frequencies, which can affect the ability to detect IBD. Next we calculate row sums and divide these values by the square root of the number of pairs. These summary statistics are then normalized genome-wide such that they follow a standard normal distribution with a mean of 0 and standard deviation of 1. Negative z-scores are difficult to interpret when investigating positive selection; therefore we square the z-scores such that the new summary statistics follow a chi-squared distribution with 1 degree of freedom. This produces a set of genome wide test statistics $\{X_{iR,s}\}$, where $X_{iR,s}$ is the chisquare distributed test statistic for IBD sharing from isoRelate at SNP s. Q-Q plots indicate that this normalization procedure produces test statistics that follow the normal distribution to a good approximation over all within and between country comparisons (data not shown).

We calculate p-values for $\{X_{iR,s}\}$, after which we perform a $-\log_{10}$ transformation of the p-values to produce our final summary statistics, used to investigate the significance of selection signatures. Finally, a 5% genome-wide significance threshold was used to assess evidence of positive selection.

Relatedness networks

To examine the haplotype sharing between isolates within and between countries, both as genome-wide averages and at a regional level, we generated relatedness networks using the R package igraph (Csardi and Nepusz, 2006). Each node in the network represents a unique isolate and an edge is drawn between two nodes if the isolates are IBD anywhere within interval. Isolates with MOI = 1 are represented by circle nodes while isolates with MOI > 1 are represented by squares. Node colors are unique for isolates from different countries.

Detecting multidrug resistance

To investigate multidrug resistance we extract all pairs who are IBD over a drug resistant gene of interest. Here a pair is classified as IBD if they have an IBD segment that partially or completely overlaps the specified interval. From this subset of pairs we calculate our selection signal as per usual and investigate the distribution of these statistics across the genome. All selection signatures that reach significance provide evidence of co-inheritance and thus mutual-selection in these pairs. Therefore we examine joint selection of an antimalarial drug resistant gene with other drug resistant genes for evidence of multidrug resistance.

Comparing methods for the detection of selection

We performed a standard analysis of selection signals using the scikit-allel v0.201.1 package in Python 2.7 (Python Software 2016; Miles and Harding 2016). To compute within-country selection statistics we calculated the integrated haplotype score (iHS) for SNPs passing a MAF filter of 1% on a per country basis (Voight et al. 2006). We report the iHS if the EHH decays to 0.05 before reaching the final SNP examined within a maximum gap distance of 2 Mb spanning the EHH region, otherwise iHS was set to missing. To standardize iHS we binned all SNPs into 100 equally sized bins partitioned on allele frequencies and then subtracted the mean

and divided by the standard deviation of iHS within that bin. We computed $\log_{10}$ P-values using the normalized iHS from a standard normal distribution.

To compute between-country selection statistics we computed the cross-population extended haplotype score (XP-EHH) for all pairwise combinations of countries in the Pf3k field isolates dataset (Sabeti et al. 2002; Sabeti et al. 2007). The same normalization and filtering procedures were applied as in the iHS computation.

Acknowledgements

This publication uses data generated by the Pf3k project (www.malariagen.net/pf3k) and in MalariaGEN *Plasmodium falciparum* Community Project. (2016). We thank the MalariaGEN Consortium for allowing the use of this data. This work was supported by National Health and Medical Research Council (NHMRC) Program Grant (APP1054618) and NHMRC Senior Research Fellowship (1002098) to M.B and a NHMRC Project Grant (APP1027108) awarded to A.E.B. L.H. was supported by the John and Patricia Farrant Scholarship and the Australian Postgraduate Award Scholarship. This work was also supported by Victorian State Government Operational Infrastructure Support and the Australian Government NHMRC IRISS funding.

References

World Health Organization. Global action plan on antimicrobial resistance 2015. WHO (online 2016) http://www.who.int/antimicrobial-resistance/publications/global-action-plan/

Neafsey, D. et al. (2008) Genome-wide SNP genotyping highlights the role of natural selection in *Plasmodium falciparum* population divergence. *Genome Biol*, 9:R171

zur Wiesch, P. A. et al. (2011) Population biological principles od drug-resistance evolution in infectious diseases. *Lancet Infect Dis*, 11(3): 236-247

Talisuna, A. O., Bloland, P. and D'Alessandro, U. (2004) History, dynamics, and public health importance of malaria parasite resistance, *Clin Microbiol Rev*, 17(1): 235-254

Mu, J. et al. (2010) *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet*, 42(3): 268-271

Sabeti, P. C. et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature,* 419:832-837

Voight, B. F., Kudaravalli, S., Wen, X. and Pritchard, J. K. (2006) A map of recent positive selection in the human genome. *PLoS Biol,* 4:e72.

Sabeti, P. C. et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature,* 449:913-918

Albrechtsen, A., Moltke, I. and Nielsen, R. (2010) Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186: 295-308

Vitti, J. J., Grossman, S. R. and Sabeti, P. C. (2013) Detecting natural selection in genomic data. *Annu Rev Genet*, 47: 97-120

Nair, S. et al. (2007) Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Mol Biol Evol*, 24(2): 562-573

MalariaGEN *Plasmodium falciparum* Community Project. (2016) Genomic epidemiology of artemisinin resistant malaria. *eLife*, 5: e08714

Han, L. and Abney, M. (2013) Using identity by descent estimation with dense genotype data to detect positive selection. *Euro J Hum Genet*, 21: 205-211

Henden, L., Wakeham, D. and Bahlo, M. (2016) XIBD:software for inferring pairwise identity by descent on the X chromosome. *Bioinformatics*, 32(15): 2389-2391

Miles, A. et al. (2016) Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Res*, 26:1-12

Manske, M. et al. (2012) Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, 487(7407): 375-379

Amambua-Ngwa, A. et al. (2012) SNP genotyping identifies new signatures of selection in a deep sample of West African *Plasmodium falciparum* malaria parasite. *Mol Biol Evol*, 29(11): 3249-3253, doi: 10.1093/molbev/mss151

Park, D. et al. (2012) Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. *PNAS*, 109(32):13052-13057, doi: 10.1073/pnas.1210585109

Laufer M. K. et al. (2006) Return of chloroquine antimalarial efficacy in Malawi. *N Engl J Med*, 355(19):1959-1966, doi: 10.1056/NEJMoa062032

Nwakanma, D. et al. (2014) Changes in malaria parasite drug resistance in an endemic population over a 25-year period with resulting genomic evidence of selection. *J Infect Dis*, 209:1126-1135

Mehlotra, R. K. et al. (2001) Evolution of a unique *Plasmodium falciparum* chloroquine-resistance phenotype in association with pfcrt polymorphism in Papua New Guinea and South America. *Proc Natl Acad Sci U S A*, 98(22): 12689-12694

Wootton, J. C. et al. (2002) Genetic diversity and chloroqine selective sweeps in Plasmodium falciparum. *Nature*, 418(6895): 320-323

Miotto, O. et al. (2013) Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nat Genet*, 45(6): 648-655

Ariey, F. et al. (2014) A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature*, 505: 50-55

Maude, R. J. et al. (2014) Spatial and temporal epidemiology of clinical malaria in Cambodia 2004-2013. *Malar J*, 13: 385, doi: 10.1186/1475-2875-13-3885

Takala-Harrison, S. et al. (2015) Independent emergence of artemisinin resistance mutations among Plasmodium falciparum in Southeast Asia. J Infect Dis, 211: 670-679

Amambua-Ngwa, A. et al. (2016) Exceptionally long-range haplotypes in *Plasmodium falciparum* chromosome 6 maintained in an endemic African population. *Malar J*, 15: 515-525

Veiga, M. I. et al. (2016) Globally prevalent PfMDR1 mutations modulate Plasmodium falciparum susceptibility to artemisinin-based combination therapies. *Nat Commun*, 7:11553, doi: 10.1038/ncomms11553

Ochola-Oyier, L. I. et al. (2016) Comparison of allele frequencies in Plasmodium falciparum merozoite antigens in malaria infections sampled in different years in a Kenyan population. *Malar J*, 15(1): 261-272

Ohashi, J. et al. (2014) Diversifying selection on the thrombospondin-related adhesive protein (TRAP) gene of *Plasmodium falciparum* in Thailand. PLoS ONE, 9(2): e90522

Anderson, T. J. C. (2000) Microsatellite markers reveal a spectrum of population structure in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol*, 17(10): 1467-1482

Daniels, R. F. (2015) Modeling malaria genomics reveals transmission decline and rebound in Senegal. *PNAS*, 112(22): 7067-7072

Carlton, J. M. et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium* vivax. *Nature*, 455: 757-763

Feil, E. J. et al. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *PNAS*, 98(1): 182-187

DePristo et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491-498

Lee, S. et al. (2016) Assessing clonality in malaria parasites using massively parallel sequencing data. In preparation.

Mu, J. et al. (2005) Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biol*, 3(10): 1734-1741

Volkman, S. K., Neafsey, D. E., Schaffner S. F., Park, D. J. and Wirth, D. F. (2012) Harnessing genomics and genome biology to understand malaria biology. *Nat Rev Genet*, 13: 315-320

Galinsky, K. et al. (2015) COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malar J,* 14(4)

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceeding of the IEEE*, 77(2):257-286

Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal*, Complex Systems 1695

Python Software Foundation. *Python Language Reference version 2.7,* https://www.python.org/

Miles, A. and Harding, N. scikit-allel: v0.20.3. (Zenodo, 2016). doi:10.5281/zenodo.45420

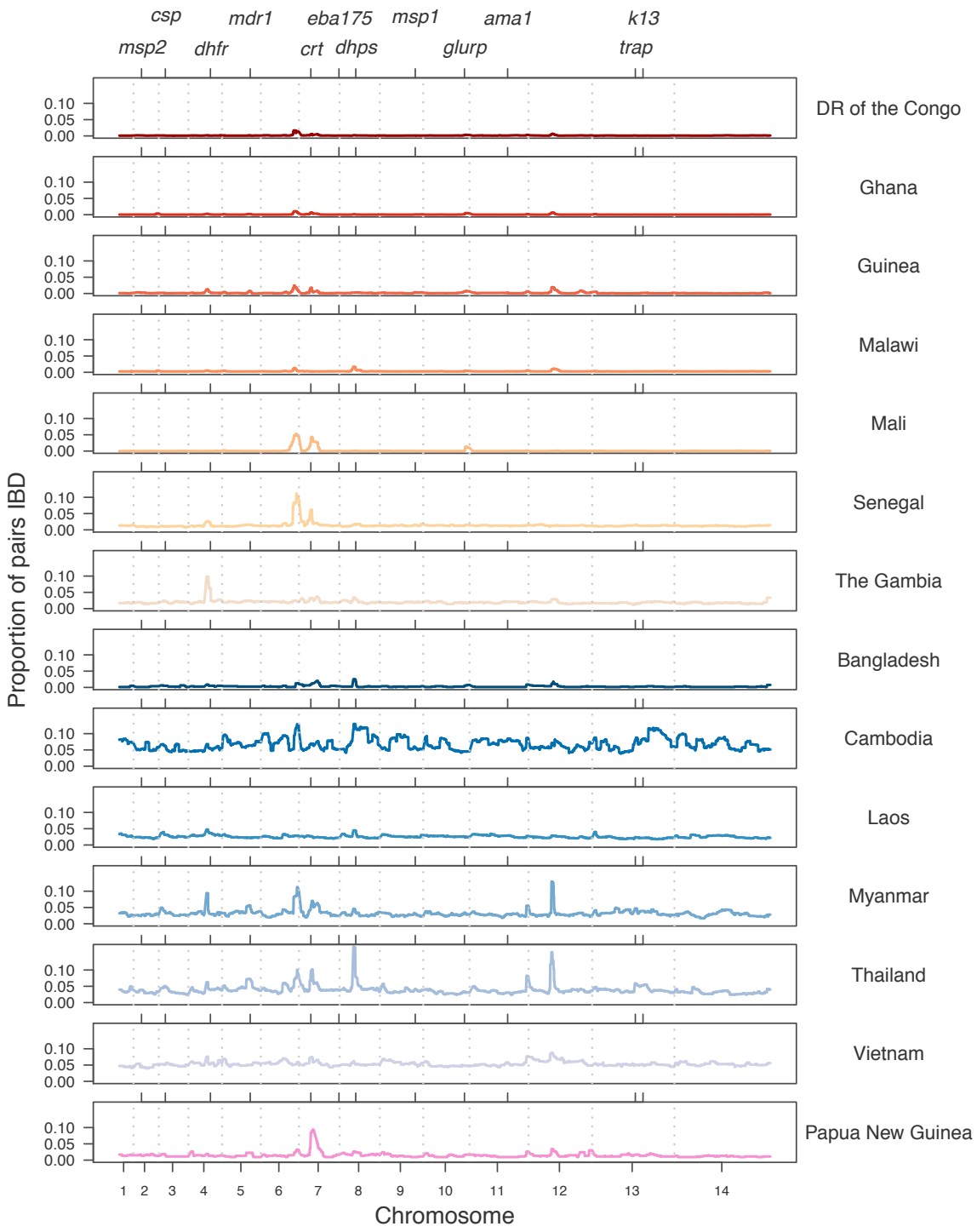

Figure 1. The proportion of pairs within each country who are IBD at each SNP, displayed genome wide. Chromosome boundaries are indicated by grey dashed vertical lines and positive control genes are identified gene symbols by tick marks on the top x-axis. Countries that are part of the African continent are shades of red and orange while countries in Southeast Asia are shades of blue and Papua New Guinea is pink.

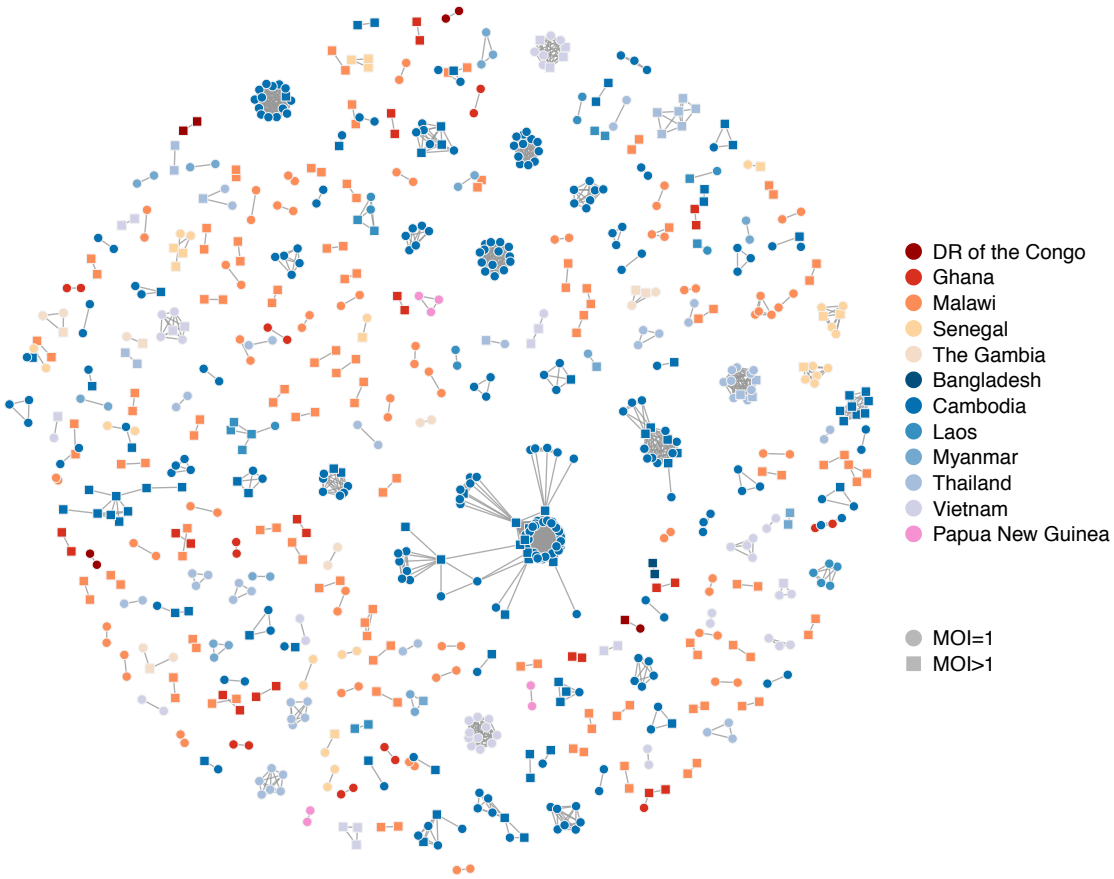

Figure 2. Relatedness network showing connections between pairs of isolates identified as having high proportions of IBD sharing. Each node identifies a unique isolates and an edge is drawn between two isolates if they share more than 90% of their genome IBD. Isolates with MOI = 1 are represented by circles while isolates with MOI > 1 are represented by squares. There are 264 clusters in this network comprising 805 isolates (out of 2,377 isolates) in total. Isolates that do not share more than 90% of their genome IBD with any other isolate are omitted from the network.

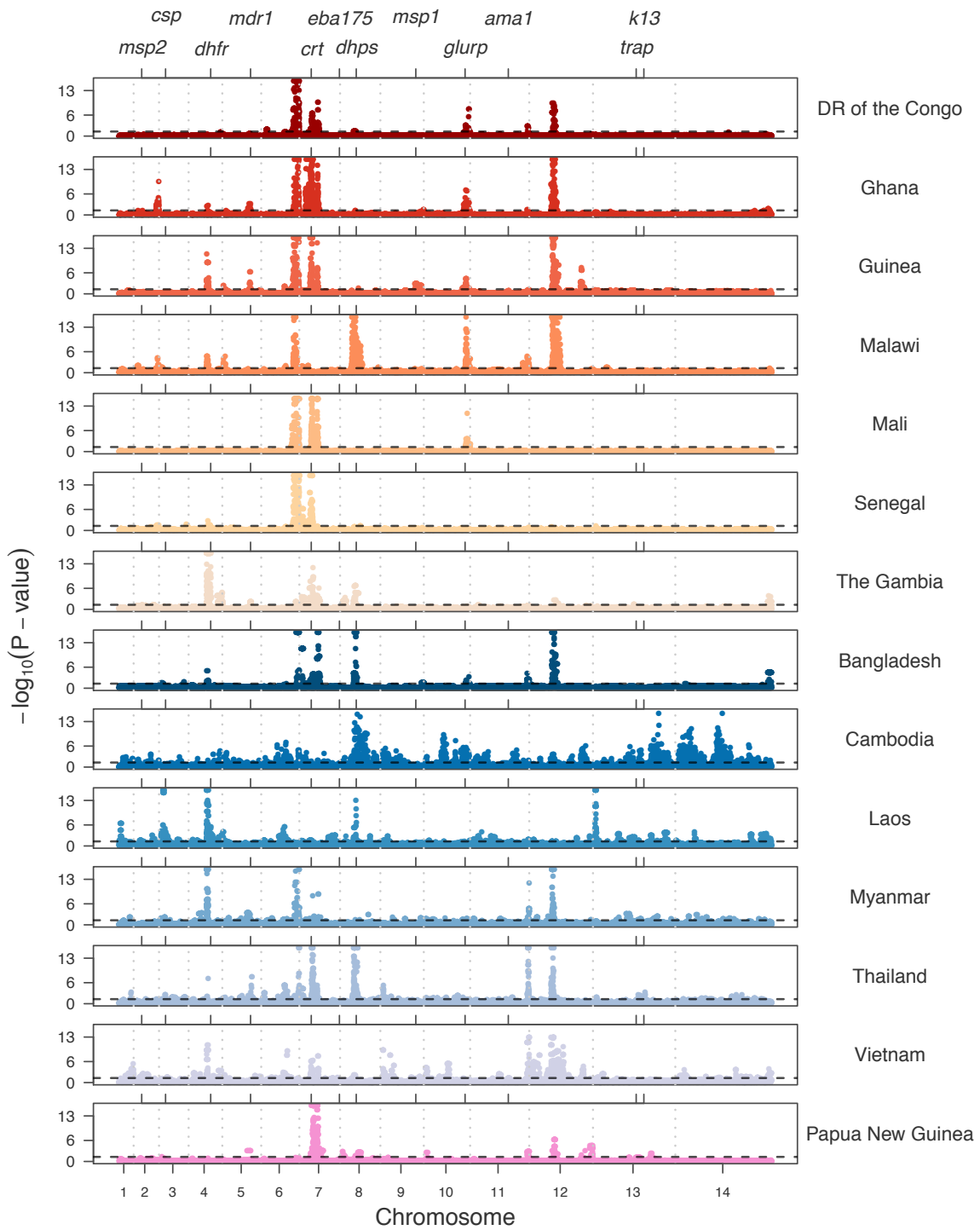

Figure 3. $-\log_{10}$(P-values) of X_{iR} calculated by transforming and normalizing the IBD proportions within each country. Dashed horizontal lines represent a 5% singnificance threshold. Grey dashed vertical lines indicate chromosome boundaries. Positive control genes are identified by gene symbol and tick marks on the upper x-axis.

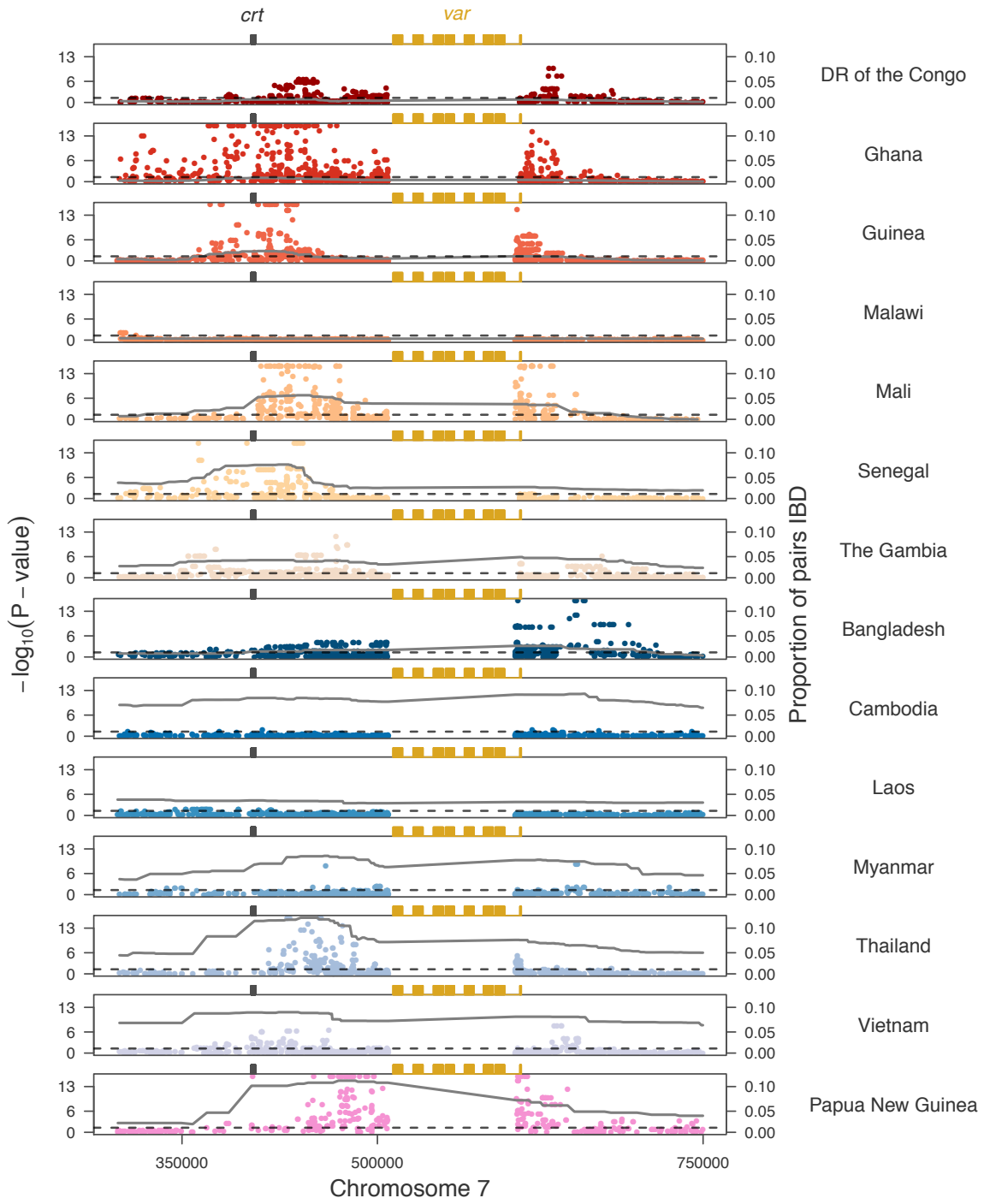

Figure 4. Selection signals within each country on chromosome 7 between 300,000-750,000bp, surrounding the *Pfcrt* locus. Coloured points correspond to $-\log_{10}$(p-values) of the X_{iR} test statistic for all SNPs and the dashed horizontal lines represent the 5% singnificance threshold. Solid lines are the proportion of pairs IBD over the interval, extrapolated between adjacent SNPs. The *Pfcrt* gene and the *var* gene locations are indicated on the upper x-axis. Note that the *var* gene clusters are blacklisted for the IBD interrogation.

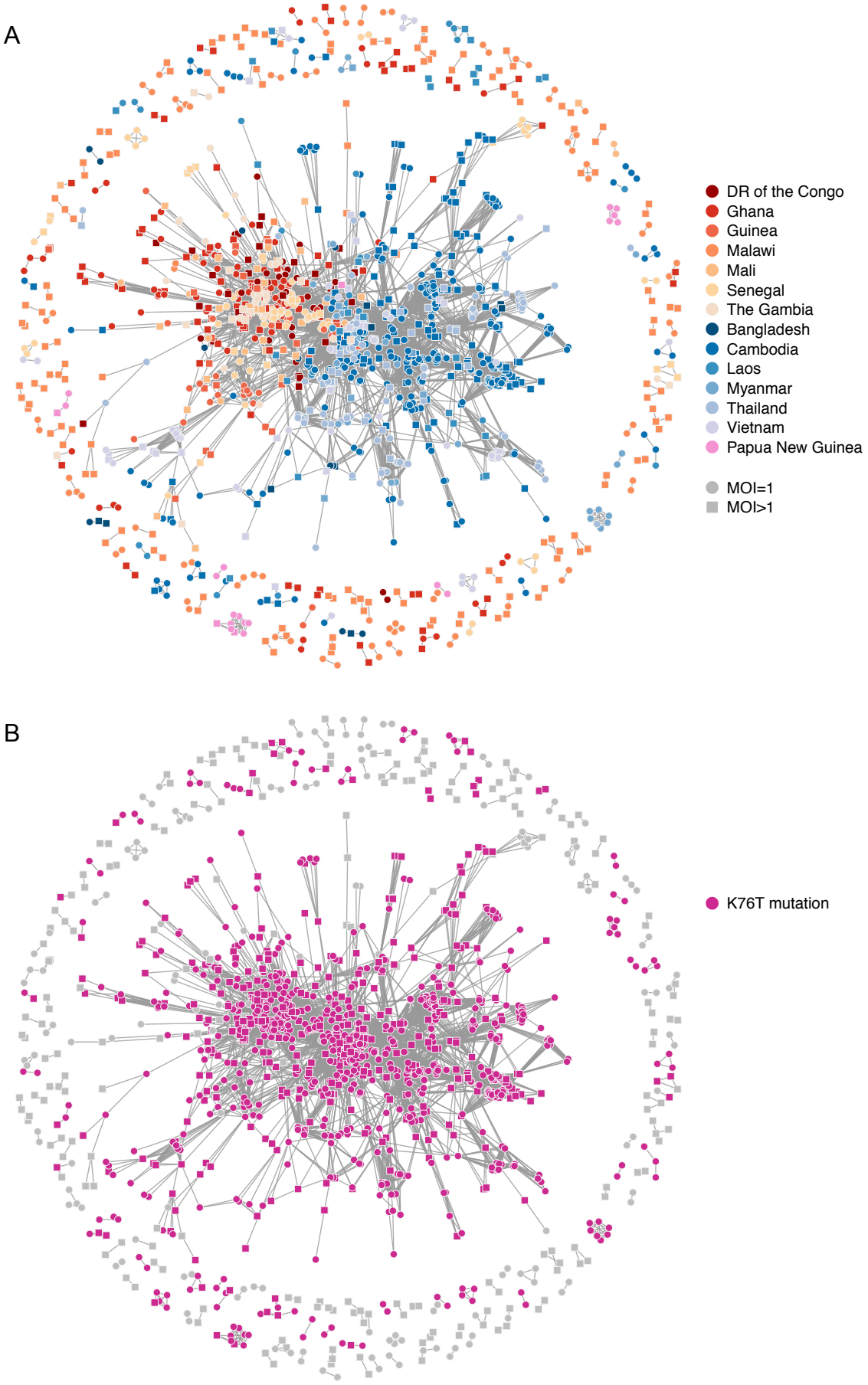
A
DR of the Congo
Ghana
Guinea
Malawi
Mali
Senegal
The Gambia
Bangladesh
Cambodia
Laos
Myanmar
Thailand
Vietnam
Papua New Guinea
MOI=1
MOI>1
B
K76T mutation

Figure 5. Relatedness network showing connections between pairs of isolates inferred IBD over *Pfcrt* (chr7: 403,222-406,317). Each node identifies a unique isolates and an edge is drawn between two isolates if they were inferred IBD anywhere over *Pfcrt*. Isolates with MOI = 1 are represented by circles while isolates with MOI > 1 are represented by squares. There are 178 clusters in this network comprising of 1,563 isolates in total, with the largest cluster containing 1,134 isolates. Isolates that are not IBD over *Pfcrt* are omitted from the network. **A.** Isolates are coloured according to country. **B**. Isolates are coloured if they carry the K78T mutation associated with chloroquine resistance.

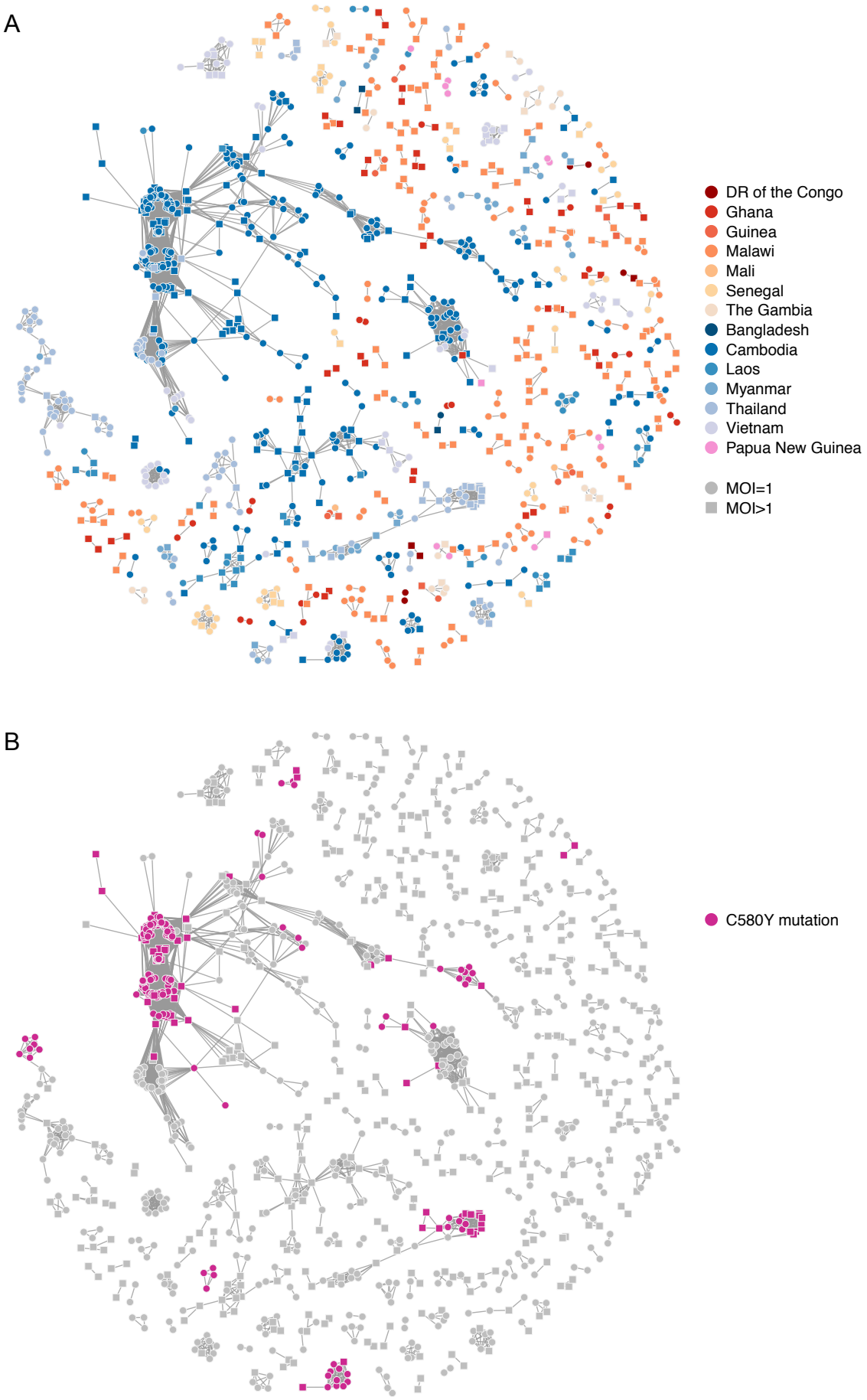
A
DR of the Congo
Ghana
Guinea
Malawi
Mali
Senegal
The Gambia
Bangladesh
Cambodia
Laos
Myanmar
Thailand
Vietnam
Papua New Guinea
MOI=1
MOI>1
B
C580Y mutation

Figure 6. Relatedness network showing connections between pairs of isolates inferred IBD over *Pfk13* (chr13: 1,724,817-1,726,997). Each node identifies a unique isolates and an edge is drawn between two isolates if they were inferred IBD anywhere over *Pfk13*. Isolates with MOI = 1 are represented by circles while isolates with MOI > 1 are represented by squares. There are 242 clusters in this network comprising of 1,148 isolates in total, with the largest cluster containing 335 isolates. Isolates that are not IBD over *Pfk13*are omitted from the network. **A.** Isolates are coloured according to country. **B**. Isolates are coloured if they carry the C580Y mutation associated with artemisinin resistance.

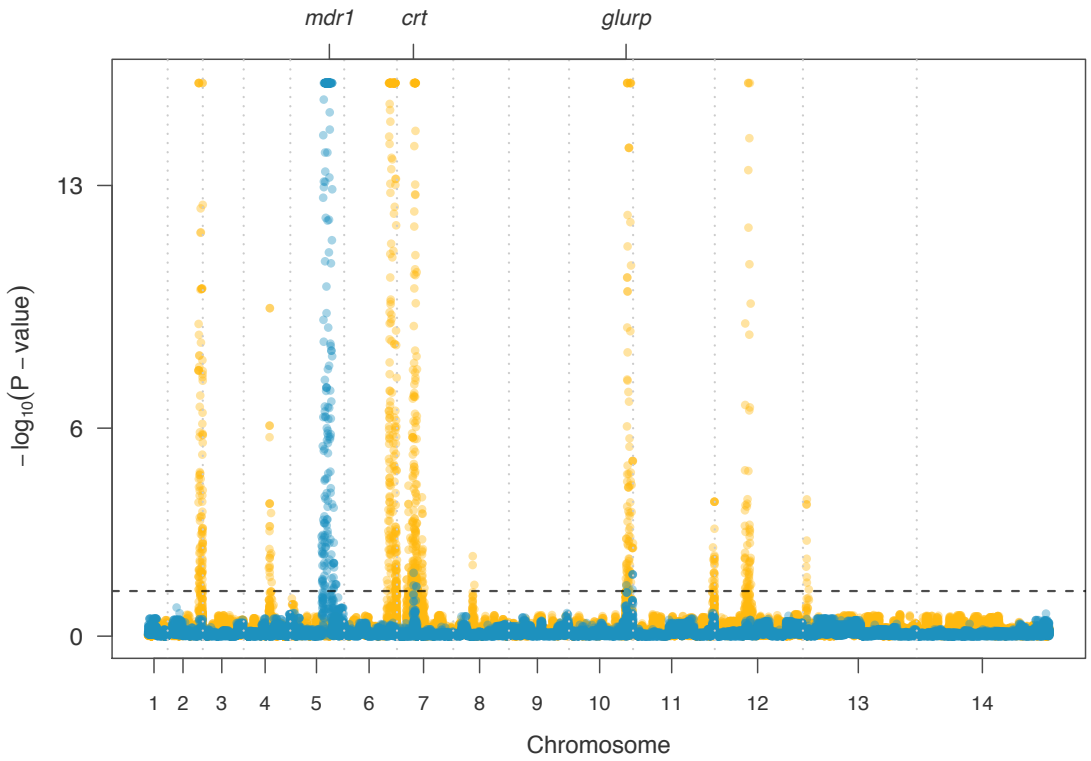

Figure 7. Genome-wide selection signatures within Ghana stratified by pairs who are IBD (blue points) or non-IBD (yellow points) over *Pfmdr1* (chr5: 957,890-962,149). The dashed horizontal line represent a 5% singificance threshold and the dashed vertical lines idenitfies the chromosome boundaries. Genomic positions of the genes *Pfmdr1*, *Pfcrt* and *Pfglurp* are indiated by ticks on the upper x-axis.