A computational method for predicting the most likely evolutionary trajectories in the stepwise accumulation of resistance mutations

R. Charlotte Eccleston[a,1], Emilia Manko[a], Susana Campino[a], Taane G. Clarke[a,b], Nicholas Furnham[a]

[a]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom; [b]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

[1]R. Charlotte Eccleston

Email: charlotte.eccleston@lshtm.ac.uk

Author Contributions: R.C.E., S.C., T.G.C. and N.F. conceived and designed the study. E.M. performed the bioinformatic processing of the raw sequencing data. R.C.E performed the research. R.C.E. analyzed the data. R.C.E. wrote the first draft of paper. All authors commented on and edited the manuscript and approved the final version. R.C.E. compiled the final manuscript.

Competing Interest Statement: None.

Classification: Major: Biological Sciences, Minor: Biophysics and Computational Biology, Evolution

Keywords: Antimicrobial resistance, epistasis, pathogen evolution

This PDF file includes:

Main Text
Figures 1 to 8
Tables 1 to 3

Abstract

Pathogen evolution of drug resistance often occurs in a stepwise manner via accumulation of multiple mutations, which in combination have a non-additive impact on fitness, a phenomenon known as epistasis. Epistasis complicates the sequence-structure-function relationship and undermines our ability to predict evolution. We present a computational method to predict evolutionary trajectories that accounts for epistasis, using the Rosetta Flex ddG protocol to estimate drug binding free energy changes upon mutation and an evolutionary model based in thermodynamics and statistical mechanics. We apply this method to predict evolutionary trajectories to known multiple mutations associated with resistant phenotypes in malaria. Resistance to the combination drug sulfadoxine-pyrimethamine (SP) in malaria-causing species *Plasmodium falciparum (Pf)* and *Plasmodium vivax (Pv)* has arisen via the accumulation of multiple point mutations in the DHFR and DHPS genes. Four known *Pf*DHFR pyrimethamine resistance mutations are highly prevalent in field-isolates and multiple studies have shown epistatic interactions between these mutations determine the accessible evolutionary trajectories to the highly resistant quadruple mutation N51I,C59R,S108N,I164L. We simulated the possible evolutionary trajectories to this quadruple *Pf*DHFR mutation as well as the homologous *Pv*DHFR mutations. In both cases, our most probable pathways agreed well with those determined experimentally. We also applied this method to predict the most likely evolutionary pathways to observed multiple mutations associated with sulfadoxine resistance in *Pf*DHPS and *Pv*DHPS. This novel method can be applied to any drug-target system where the drug acts by binding to the target.

Significance Statement

Antimicrobial resistance (AMR) is a major public health threat, resulting from the overuse and misuse of antimicrobial drugs. Our ability to monitor emerging resistance and direct the most appropriate treatment strategy (stewardship) would be strengthened by the development of methods to accurately predict the mutational pathways that lead to resistance. We present a computational method to predict the most likely resistance pathways that result from mutations that alter drug binding affinity. We utilized the Rosetta Flex ddG protocol and a thermodynamic evolutionary model. This novel approach can accurately capture known resistance trajectories and can be applied to any system where resistance arises via changes in drug binding affinity.

Main Text

Introduction

The development of new antimicrobial drugs and therapeutics, and the design of more successful drug deployment strategies to reduce the prevalence of resistance, requires an understanding of the underlying molecular evolution.

Antimicrobial resistance (AMR) poses a huge global health threat by a wide range of mechanisms including: the protection of drug targets by post-translational modifications such as methylation; the presence and increased expression of proteins that bind to targets preventing drug interactions; limiting the physical entry of the drug compound to the cell or increasing the efficiency of drug efflux through the type or number of efflux pumps, or in the case of bacteria, horizontal gene transfer [1-3]. One of the major routes to resistance, and a focus of the work presented here, is genomic variation within protein coding regions. Of particular significance are single-nucleotide polymorphisms (SNPs) in the antimicrobial target gene that alter the protein structure and prevent efficient binding of the antimicrobial drug. Provided these SNPs do not prevent the target from carrying out its function, the resistant strains will proliferate within the population [4].

The evolution of resistance alleles is affected by the evolutionary interplay between selection for resistance and selection for protein function, as well as drug concentration, epistasis between resistant mutations and mutational bias [5-8]. Prime examples of this are the *Plasmodium falciparum* and *vivax* parasites that cause the majority of malaria infections and which have evolved strong resistance to many antimalarial drugs, including pyrimethamine [9], chloroquinine [10] and sulfadoxine [11].

There were an estimated 229 million new cases of malaria world-wide in 2019, resulting in approximately 409,000 deaths predominately among children under 5 [12]. *P. falciparum* accounts for the majority of cases in Africa, Southeast Asia, the Eastern Mediterranean and Western Pacific, whilst *P. vivax* is predominant in Central and South America.

P. falciparum malaria has been treated with the combination drug sulfadoxine-pyrimethamine (SP) since 1970s, which targets the folate metabolic pathway. Numerous resistance mutations have arisen within its genome as a result of SNPs in *P. falciparum* dihydrofolate reductase (*Pf*DHFR) and dihydropteroate synthase (*Pf*DHPS) genes, which are the targets of pyrimethamine and sulfadoxine respectively [11, 13, 14]. Although SP is not usually used to treat *P. vivax*, co-infections with *P. falciparum* have meant SP resistance mutations have also arisen in the *P. vivax* genome [15]. The enzymes of the folate pathway are largely conserved across *Plasmodium* species, and so polymorphisms in equivalent positions have been observed in *P. vivax* DHFR (*Pv*DHFR) and DHPS (*Pv*DHPS) and are thought to confer resistance to SP [16-19]

The DHFR gene encodes an enzyme that uses NADPH to synthesize tetrahydrofolate, a co-factor in the synthesis of amino acids [20], and pyrimethamine acts to disrupt this process, thereby blocking DNA synthesis and slowing down growth. Stepwise acquisition of multiple mutations leading to resistance to pyrimethamine has been observed in both *Pf*DHFR [7, 9] and *Pv*DHFR [8]. DHPS exists as a bifunctional enzyme with hydroxymethyldihydropterin pyrophosphokinase (HPPK) and performs a key step in folate synthesis by catalysing the conversion of para-aminobenzoate (pABA) to dihydropteroate. The anti-folate drug sulfadoxine

targets DHPS by preventing this conversion and mutations in DHPS have arisen that confer resistance to sulfadoxine [21, 22].

In *Pf*DHFR, a combination of four mutations – Asn-51 to Ile (N51I), Cys-59 to Arg (C59R), Ser-108 to Asn (S108N) and Ile-164 to Leu (I164L) – has been reported to result in resistance to pyrimethamine [23] by altering the binding pocket and reducing the affinity for the drug [24]. Non-additivity in both pyrimethamine binding free energy and the concentration required to inhibit cell growth by 50% (IC50) has been observed experimentally for combinations of these four mutations [7, 9]. This non-additivity – known as epistasis – means that mutations which on their own are not associated with a resistance phenotype, could do so when in combination with other mutations. Epistasis between these mutations has been shown to determine the evolutionary trajectories to the quadruple mutation N51I,C59R,S108N,I164L, which is strongly associated with pyrimethamine resistance.

A similar investigation was conducted into the homologous set of *Pv*DHFR mutations – Asn-50 to Ile (N50I), Ser-59 to Arg (S58R), Ser-117 to Asn (S117N) and Ile-173 to Leu (I173L) - and the accessible evolutionary trajectories to the quadruple mutation [8], some combinations of which have been observed to result in pyrimethamine resistance both *in vivo* and *in vitro* [19, 25]. Evolutionary simulations accounting for growth rates, IC50 measurements for increasing concentrations of pyrimethamine and nucleotide bias predicted the most likely pathways to the quadruple mutation for different drug concentrations. They observed the trajectories at each concentration were influenced by epistasis between the mutations and the adaptive conflict between endogenous function and acquisition of drug resistance.

These studies, along with other investigations [5, 26] have highlighted the prevalence of epistasis among resistance mutations and the importance of considering epistatic interactions between mutations when predicting evolutionary trajectories of drug resistance.

Epistasis between mutations within the same protein arises due to energetic interactions between the amino acids, where the impact of a mutation on either function, fitness or a physical property

depends upon the protein sequence [27]. When epistasis occurs between two or more mutations, their combined impact on protein fitness or a physical trait such as stability or binding affinity, does not equal the sum of their independent impacts. As well as the aforementioned studies, epistasis been observed in the evolution of many pathogens, such as *Escherichia coli* [28], Influenza [29], and RNA viruses [30]. Epistasis may also have important consequences for the success of AMR management strategies that aim to reduce resistance via the cessation of use of a particular drug. In theory, stopping the use of a drug should result in reversion of resistance mutations, because most resistance mutations incur a large fitness cost in the absence of drug [31-33]. However, the success of this strategy has been mixed, and in some cases bacterial populations remained resistant [34-36] , which may be the result of compensatory mutations (a type of epistasis), which mitigate the deleterious impact of resistance mutations, allowing them to remain in a population and thus retain resistance even in the absence of drug selection pressures [37].

Therefore, understanding how epistasis arises and predicting which mutations will interact epistatically, is important for anticipating future mutations, designing new drugs and developing strategies to minimize resistance and predicting the most likely evolutionary trajectories to resistance.

Whilst experimental methods have been successful in characterising the interactions between resistance mutations and predicting evolutionary trajectories, they are expensive and time consuming. Therefore, identifying computational methods that can accurately characterise how mutations interact could potentially enable fast and reliable predictions of evolutionary trajectories to resistance.

One of the main mechanisms of drug resistance is selection for mutations that reduce drug binding free energy. Several computational tools exist to predict changes in binding free energy upon mutation, however their ability to predict epistasis is limited or unknown. The Rosetta protocol Flex ddG [38] is the current state-of-the art method for predicting changes in protein-

protein and protein-ligand binding free energy. Rosetta is a software suite for macromolecular modelling and design that uses all-atom mixed physics- and knowledge-based potentials, and provides a diverse set of protocols to perform specific tasks, such as structure prediction, molecular docking and homology modelling [39]. The Flex ddG protocol has been found to perform better than machine learning methods and comparably to molecular dynamics methods when tested on a large dataset of ligand binding free energy changes upon protein mutation [40, 41]. However, its ability to capture epistasis has not yet been tested. Therefore, we investigated how well Flex ddG can capture epistasis between mutations and observed a good correlation with experimental data.

Next, we developed a method for predicting evolutionary trajectories to multiple resistance mutations, using predicted binding free energy changes and an evolutionary model based in thermodynamics and statistical mechanics. We tested this method by simulating the accessible evolutionary trajectories to known resistant quadruple mutants in both *Pf*DHFR and *Pv*DHFR, as these targets have been studied experimentally. We observed good agreement between the most likely trajectories predicted by our model and those predicted experimentally.

Once we had demonstrated our method for predicting evolutionary trajectories agrees well with experimentally determined trajectories, we extended our analysis to the DHPS gene of *P. falciparum* and *P.vivax*. This work represents the first attempt to predict evolutionary trajectories to resistance in these genes, either experimentally or computationally.

We compared our predictions to the frequency of mutations found in isolate data from countries from several geographical regions including South America, West and East Africa, South and Southeast Asia and Melanesia.

Results

Rosetta Flex-ddG accurately predicts binding free energy changes and epistasis for *Pf*DHFR mutations

We investigated if Flex ddG, was able to capture experimentally measured non-additivity in binding free energy. The change in binding free energy was predicted for the combinatorically complete set of the four *Pf*DHFR pyrimethamine resistance mutations N51I, C59R, S108N and I164L. The authors of the Flex ddG protocol suggest conducting a minimum of 35 runs and taking the average of the distribution as the prediction for that mutation. We observed that the distributions were not well characterised by only 35 runs and so we undertook 150 runs for each mutation. Similarly, many of the distributions were not Normal and so we chose to use the peak of the distribution as a prediction for each mutation.

We compared our predictions with the data of Sirawaraporn et al. (1997) [9] who experimentally determined the change in binding free energy for a subset of the combinatorically complete set of the four known *Pf*DHFR resistance mutations (Table 1), where a positive ΔG value indicates a destabilising mutation and a negative ΔG value indicates a stabilising mutation (Note: Rosetta Flex ddG calculates the change in binding free energy as $\Delta\Delta G = \Delta G_{mut} - \Delta G_{WT}$, whereas Sirawaraporn et al. (1997) calculated the change as the reverse, $\Delta\Delta G = \Delta G_{WT} - \Delta G_{mut}$, where *WT* indicates the wild-type free energy and *mut* indicates the mutant free energy. Therefore, in the original manuscript, a mutation that destabilised the binding corresponded to a negative $\Delta\Delta G$, whilst here we have reversed the signs of the original data to enable comparison with our predictions. Furthermore, Flex ddG predictions are in Rosetta Energy Units (R.E.U), whereas data from [9] is in units kcal/mol).

Our predictions had a Pearson's correlation of 0.63 with the experimental dataset and 8/9 of the mutations were correctly classified as either stabilising or destabilising. These metrics show that the predictions capture the experimental data satisfactorily. The single mutation S108N was found to be the only destabilising single mutation in both the experimental data and the

predictions, with all other single mutations having a stabilising impact on pyrimethamine binding. There are, however, some discrepancies, namely in the experimental data the double mutation N51I/S108N is more destabilizing to binding than single mutation S108N, however the peak of the Flex ddG distribution was stabilizing, with only a few of the 150 runs predicting the mutation was destabilizing. The triple mutation C59R/S108N/I164L was found experimentally to be the most destabilizing of the triple mutations, however Flex ddG predicted it to be only mildly destabilizing and the least destabilizing of the triple mutations. Furthermore, the quadruple mutation was found experimentally to have the most destabilising impact out of all combinations of single and multiple mutations, however, Flex ddG predicted it to be less destabilising than the double mutation C59R/S108N.

Sirawaraporn et al. (1997) determined the ‘interaction energy’, which they defined as the difference between the change in binding free energy due to a multiple mutation and the sum of the changes in binding free energy when the mutations occur on their own. For a double mutation, the interaction energy is the same as the pairwise epistasis between the pair of mutations. For triple and quadruple mutations, the interaction energy quantifies the non-additivity between the mutations and indicates the level of epistasis between the mutations.

We determined the interaction energy by finding the difference between the predicted change in binding free energy of a multiple mutation and the sum of the predictions of their independent binding free energy changes (Table 1). A positive value of the interaction energy indicates the sum of the independent impacts is more destabilizing than the impact of the multiple mutation and a negative value indicates the sum is less destabilizing than the combined impact.

A correlation of 0.765 with the data was observed and 4/5 values were correctly classified as having either a positive or negative interaction energy. The incorrectly classified mutation was again N51I/S108N, where the interaction energy was predicted to be positive, but found experimentally to be negative, because the sum of the individual predictions was destabilising but the double mutation itself was predicted to have a stabilising impact.

Both the experimental data and our predictions found that the quadruple mutation had the largest magnitude interaction energy reflecting the greatest difference between the stabilising impact of the sum of the individual mutations and the destabilising impact of the quadruple mutation itself.

The triple mutation C59R,S108N,I164L was found to have a large negative interaction energy in both the data and our predictions, where the triple mutation was found to be destabilizing whilst the sum of the individual impacts was stabilising to pyrimethamine binding. We also observed large negative interaction energy between S108N and C59R, where C59R is stabilizing in the wildtype background but destabilizing in the background of S108N, an example of sign epistasis and in agreement with the observations of both Sirawaraporn et al. (1997) and Lozovosky et al. (2009).

However, whilst the interaction energy of the triple mutation N51I,C59R,S108N was negative for both data and predictions, in our predictions its magnitude was much smaller compared to the data. Both single mutations N51I and C59R were predicted to be only marginally stabilising – almost neutral - to pyrimethamine binding, whilst in the experimental data both mutations have a large stabilising impact. Furthermore, the triple mutation was predicted to be only marginally more destabilising than single mutation S108N, resulting in the small negative interaction energy.

A thermodynamic evolutionary model can predict the most likely evolutionary trajectories to quadruple mutations in both *Pf*DHFR and *Pv*DHFR

We simulated the evolutionary trajectories to quadruple mutant N51I,C59R,S108N,I164L, using an evolutionary model, adapted from previous studies [42-44], in which selection acts to reduce the binding affinity between *Pf*DHFR and pyrimethamine. Briefly, starting from the wild-type protein, we randomly sample a value from the Flex ddG distributions for each of the four single mutations and calculate the fitness of the mutated protein (Eq. 1), and the fixation probability (Eq. 2). A mutation is then chosen with a probability proportional to the fixation probability and this is

repeated until the quadruple mutation is reached. We carried out 10,000 runs and determined the probability of each trajectory.

The top two most likely trajectories were S108N/C59R/N51I/I164L, S108N/C59R/I164L/N51I, respectively (Figure 1a), corresponding to the top two most likely pathways determined in [7]. The fourth most likely pathway predicted by our model, S108N/N51I/C59R/I164L, corresponds to the third most likely pathway predicted by [7].

We compared our predictions to the frequency of mutations observed in our isolate data (Figure 1b). In the isolate data, S108N was the most frequent single mutation, C59R,S108N the most frequent double mutation and N51I,C59R,S108N the most frequent triple mutation. This suggests the pathway proceeds in the order S108N/C59R/N51I/I164L, in agreement with the most likely pathway from both our evolutionary simulations and those using experimental data.

Single mutations C59R and N51I are the second and third most prevalent single mutations, respectively, in both our isolate data and our pathway predictions and I164L is absent from the isolate data and has a zero probability of being selected as the first step of our evolutionary trajectories. Whilst there is agreement between the model and the data regarding the most common double mutation (C59R,S108N), the agreement is less convincing for the remaining double mutations. For example, N51I,S108N is the second most frequent double mutation in the isolate data but the fourth most likely double mutation predicted by the model when summing up all pathways. There is better agreement with the triple mutations, where C59R,S108N,I164L is the second most frequent triple in the data and the second most likely triple in the simulations when summing over all possible pathways.

We also compared our pathway predictions to the distributions of mutations from isolates from different geographical areas. We grouped the isolate data in to five geographical regions: South America (Brazil, Colombia and Peru), West Africa (Benin, Burkina Faso, Cameroon, Cape Verde, Cote d'Ivoire, Gabon, Gambia, Ghana, Guinea, Mali, Mauritania, Nigeria and Senegal), Central and East Africa (Congo, Eritrea, Ethiopia, Kenya, Madagascar, Malawi, Tanzania, Uganda),

Southeast Asia (Bangladesh, Cambodia, Indonesia, Laos, Myanmar, Thailand and Vietnam) and Melanesia (Papua New Guinea).

Examining the geographical distribution of *Pf*DHFR mutations (Figure 2), we can see that some combinations of the aforementioned set of four resistance mutations are observed in high frequencies in most regions (See Supplementary data 'PfDHFR_pcnt.csv'). By considering the frequencies of the possible combinations of the four mutations N51I, C59R, S108N and I164L, we can infer the likely evolutionary pathways to pyrimethamine resistance followed in each region and compare with our predicted pathways.

The evolution of pyrimethamine resistance in West Africa and Southeast Asia appears to have followed the two main pathways predicted here and in previous studies, S108N/C59R/N51I/I164L and/or S108N/C59R/I164L/N51I.

In West Africa, the most frequent mutations were single mutation S108N, double mutation C59R,S108N and triple mutation N51I,C59R,S108N, which were found in 1.2%, 8.1%, 67.0% of isolates respectively. The quadruple mutation N51I,C59R,S108N,I164L, single mutations C59R, and N51I and double mutations N51I,C59R and N51I,S108N were each found in >1% of isolates. This suggests West Africa is following the main trajectory S108N/C59R/N51I/164L.

In Southeast Asia, single mutation S108N, double mutation C59R,S108N, triple mutation N51I,C59R,S108N and quadruple mutation N51I,C59R,S108N,I164L were observed in >1%, 8.2%, 34.2% and 52.5% of isolates, respectively. Additionally, double mutants N51I,C59R, N51I,S108N and S108N,I164L were each observed in >1% of isolates and triple mutants N51I,C59R,I164L and C59R,S108N,I164L were observed in >1% and 3.7% of isolates, respectively. This suggests the two most likely pathways to pyrimethamine resistance in this region are S108N/C59R/N51I/I164L and S108N/C59R/I164L/N51I.

However, not all regions follow the most likely trajectories predicted by our evolutionary model. The evolution of pyrimethamine resistance in Central and East Africa appears to have followed a slightly different pathway. The three most common mutations were double mutations N51I,S108N and C59R,S108N and triple mutation N51I,C59R,S108N which were found in 12.1%, 2.9% and 82.3% of isolates respectively. Single mutation S108N, double mutation N51I,C59R, triple mutation N51I,S108N,I164L and quadruple mutation N51I,C59R,S108N,I164L were each found in >1% of isolates. All other combinations of the four mutations were absent. This suggests the main trajectory in this region is S108N/N51I/C59R/I164L, in addition to a second less likely trajectory S108N/C59R/N51I/I164L.

The distribution of mutations in South America also suggests alternative routes to pyrimethamine resistance. The most frequent mutations were single mutation S108N, double mutation N51I,S108N and triple mutation N51I,S108N,I164L which accounted for 52.0%, 24.0% and 8.0% of the isolates, respectively. All other combinations of the four mutations were not observed in South America. This suggests this region is following the trajectory S108N/N51I/I164L to the triple mutation.

In Papua New Guinea, however, out of all possible combinations of the four mutations, only single mutation S108N and double mutation C59R,S108N were observed in >1% and 19.3% of isolates, respectively. Triple mutation C59R,S108N,S306F was observed in the remaining 79.8% of isolates.

From the geographical distribution we can see that there are a small number of other mutations that appear at non-negligible frequencies. In South America, the single mutation A16V was observed in 2.0% of isolates and the triple mutation C50R,N51I,S108N is observed in 8.0% of isolates.

Next, we generated Flex ddG predictions of the change in binding free energy for each mutation in the combinatorically complete set of four mutations in *Pv*DHFR, (N50I, S58R, S117N and

I173L), which are homologous to the four in *P. falciparum*. We used these predictions to simulate the evolutionary trajectories to the quadruple mutation (Figure 3a) and compared our results to those presented in Jiang et al. (2013) [8]. In [8] they consider the wild-type allele to have a Serine at codons 58 and 117 and therefore we do the same for ease of comparison. The most likely first step in our model is S58R, which Flex ddG predicts to be the only single mutation to reduce the binding affinity when considering both the average and the peak of the distribution (Figure S1). Single mutation S117N reduces the binding affinity on average, but the peak of the distribution is a mild stabilizing effect (Figure S2).

The most likely pathway predicted by our simulations (S58R/S117N/I173L/N50I) corresponds to the second most likely pathway predicted by Jiang et al. (2013) [8] for the highest pyrimethamine concentration. Our second most likely pathway (S58R/S117N/N50I/I173L) corresponds to the first most likely pathway predicted in [8] for the highest pyrimethamine concentration.

We compared the frequency of the 16 possible combinations of mutations in the pathway to the frequency found in our *P. vivax* isolate data (Figure 3b). Mutations S117N and S58R were the top two most frequent single mutations in our isolate data, respectively. The double mutant S58R,S117N was the most frequent mutation of the set found in our isolate data and was the most likely double mutation in our simulations when considering all possible routes to each of the six possible double mutants. The only triple mutant observed in the data was S58R,S117N,I173L and this was the most likely of the four possible triple mutations in our simulations when considering all possible routes. The quadruple mutation N50I,S58R,S117N,I173L is not observed in our isolate data, and has not been observed in the literature either.

We grouped the data into five broad geographical regions: South America (Brazil, Colombia, Guyana, Panama, Peru), East Africa (Ethiopia, Eritrea, Madagascar, Sudan, Uganda), South Asia (Afghanistan, Bangladesh, India, Pakistan, Sri Lanka), Southeast Asia (Cambodia, China, Laos, Malaysia, Myanmar, Philippines, Thailand and Vietnam) and Melanesia (Papua New Guinea).

If we consider the geographical distribution of mutations in *Pv*DHFR (Figure 4), we can see the dominant mutations differs between regions and compared to *Pf*DHFR (Figure 2) there is a more diverse set of mutations that occur at non-negligible frequencies. As mentioned previously, the quadruple mutation is not observed in the isolate data and so we will infer trajectories up to triple mutant combinations of the four mutations N50I, S58R, S117N, I173L.

In South America, single mutations S117N and S58R and double mutation S58R,S117N were observed in 13.2%, 3.5% and 28.8% of isolates respectively. Double mutation S117N,I173L and triple mutation S58R,S117N,I173L were both observed in >1% of isolates. This suggests the evolution to triple mutation S58R,S117N,I173L follows a main trajectory S117N/S58R/I173L and two less likely alternative trajectories S58R/S117N/I173L and S117N/I173L/S58R.

However, in East Africa, South Asia, Southeast Asia and Papua New Guinea, only S117N, S58R and S58R,S117N are observed out of the possible mutational combinations.

In addition to the set of four mutations studied here, we observed other *Pv*DHFR mutations in our isolate data. The double mutation S58R,S117N is now so frequent in the population that both mutations are considered fixed and the allele containing the substitutions at those residues is considered the major allele [45]. Therefore, we also observe mutations N117T and R58K at low frequencies in our dataset as these are mutations that occur following the Asparagine substitution at position 117 and the Arginine substitution at position 58.

Two quadruple mutations F57L,S58R,T61M,N117T and F57I,S58R,T61M,N117T were also found in high frequency in our isolate data. Mutation F57L,S58R,T61M,N117T was found in 38.5% of isolates from Papua New Guinea and 15.1% of isolates from Southeast Asia, whilst F57I,S58R,T61M,N117T was found in 29.7% of isolates from Southeast Asia. Intermediates of these two quadruple mutations were also found in our isolate data. The high frequency of these

quadruple mutations suggests it may result in a resistant phenotype and future work could study the trajectories to these mutations.

Predicting the most likely evolutionary pathways to multiple resistant mutations in *Pf*DHPS

The *Pf*DHPS mutations Ser-436 to Ala (S436A), Ala-437 to Gly (A437G), Lys-540 to Glu (K540E) and Ala-581 to Gly (A581G) are highly prevalent in isolate data and have been shown to confer resistance to sulfadoxine, both independently and in combination. In our data, the quadruple mutation S436A,A437G,K540E,A581G is observed, along with high frequency intermediate single, double and triple mutant combinations.

We generated Flex ddG predictions of the change in binding free energy upon mutation for the combinatorically complete set of these four mutations. Triglia et al. (1997) [46] reported a 10-fold decrease in sulfadoxine binding affinity for single mutation A437G, but only a 4-fold decrease in binding affinity for A581G. Double mutations A437G/A581G and S436A/A437G were reported to result in a further 10-fold decrease in binding affinity, with a 70-fold decrease for A437G/K540E, relative to A437G alone. These results suggest these mutations are interacting epistatically.

We compared the peak of our predicted distributions for each mutation to the data presented in Triglia et al. (1997) [46] (Table 2). We observed Pearson's correlation of 0.469 with 5/5 correctly classified as having a destabilizing impact. In the data, single mutation A437G was found to have a greater destabilizing impact than A581G. If we consider the average of the distributions, A437G was found to be around 50 times more destabilizing than A581G ($< \Delta\Delta G_{A437G} >= 0.259$, $< \Delta\Delta G_{A581G} >= 0.005$ (R.E.U)). However, if we consider the peak of the distributions, A437G has a smaller destabilizing impact than A581G ($\Delta\Delta G^{*}_{A437G} = 0.025$, $\Delta\Delta G^{*}_{A581G} = 0.11$ (R.E.U)). If we analyze the overall distribution of A437G (Fig. S3) we can see that whilst the peak of the distribution occurs around small destabilizing changes in free energy and there is a long tail where many of the runs have predicted a much larger destabilizing impact, with a maximum of

1.58 (R.E.U), and only 2/150 runs predicted a stabilizing impact. The A581G distribution (Fig. S4), however, has a maximum destabilizing prediction of just 0.246, with almost half (65/150) of the runs predicting a stabilizing effect. This again highlights the need to consider the entire distribution of predictions for a specific mutation, instead of a summarizing statistic, when predicting the most likely evolutionary pathways.

We simulated the possible evolutionary trajectories to known *Pf*DHPS sulfadoxine quadruple mutation S436A,A437G,K540E,A581G (Figure 5a). The most likely first step was single mutation A437G, in agreement with previous studies which found A437G was the most frequent single mutation sulfadoxine-resistant *Pf*DHPS isolates [13, 47] and the data presented in [46]. Furthermore, A437G is now considered fixed in the population [48] and A437G is also the most common mutation in our isolate data, followed by S436A (Figure 5b). The most likely double mutation in our simulations was A437G/S436A, followed by A437G/K540E. This corresponds well to our isolate data in which the two most likely double mutations were A437G,K540E and S436A,A437G, respectively. The most likely triple mutation from our simulations was A437G/K540E/A581G followed by A437G/A581G/K540E and the two most frequent triple mutations in our isolate data were A437G,K540E,A581G and S436A,A437G,K540E, respectively. The most likely pathway to the quadruple mutation was A437G/S436A/A581G/K540E, followed by A437G/K540E/A581G/S436A. In our isolate data, the quadruple mutation is not very common and is only found in 1/6038 isolates.

We grouped the isolate data in to five geographical regions: South America (Brazil, Colombia and Peru), West Africa (Benin, Burkina Faso, Cameroon, Cape Verde, Cote d'Ivoire, Gabon, Gambia, Ghana, Guinea, Mali, Mauritania, Nigeria and Senegal), East Africa (Congo, Eritrea, Ethiopia, Kenya, Madagascar, Malawi, Tanzania, Uganda), Southeast Asia (Bangladesh, Cambodia, Indonesia, Laos, Myanmar, Thailand and Vietnam) and Melanesia (Papua New Guinea).

In Southeast Asia, the most common mutations were single mutation A437G, double mutations S436A,A437G, A437G,K540E and A437G,A581G and triple mutations S436A,A437G,K540E and

A437G,K540E,A581G which were observed in 4.6%, 5.8%, 6%, 2.3%, 21.5% and 28.6% of isolates, respectively. Single mutations S436A and K540E were observed in 1.1% and >1% of isolates, respectively and the quadruple mutation S436A,A437G,K540E,A581G was observed in >1% of isolates. This suggests three possible trajectories to the quadruple mutation with A437G/K540E/A581G/S436A being the most likely pathway, followed by A437G/K540E/S436A/A581G and A437G/S436A/K540E/A581G as the second and third most likely pathways, respectively.

In South America, the most common mutations were single mutation A437G and double mutation A437G,A581G which were both found in 10.0% of isolates and triple mutation A437G,K540E,A581G which was observed in 16.0% of isolates. All other combinations of the four mutations were not observed in the South America isolates. This suggests the pathway followed to the triple mutation is A437G/A581G/K540E.

In West Africa, single mutations A437G and S436A were observed in 41.9% and 17.2% of isolates, respectively, double mutations S436A,A437G and A437G,K540E were observed in 18.0% and >1% of isolates, respectively and triple mutation A437G,K540E,A581G was observed in >1% of isolates. This suggests separate evolutionary trajectories to double mutation S436A,A437G and triple mutation A437G,K540E,A581G. The main trajectory to the S436G,A437G is A437G/S436A with a second slightly less likely trajectory S436A/A437G, whilst the trajectory to the triple mutation likely occurs A437G/K540E/A581G.

In East Africa, the most common mutations were single mutation A437G, double mutation A437G,K540E and triple mutation A437G,K540E,A581G which were observed in 24.8%, 57% and 7.3% of isolates, respectively. Single mutations S436A and K540E were observed in 1.5% and >1% of alleles, respectively and double mutations S436A,A437G and A437G,A581G were observed in 1.1% and >1% of alleles respectively and triple mutations S436A,A437G,K540E was observed in >1% of isolates. This suggests separate trajectories to triple mutants S436A,A437G,K540E and A437G,K540E,A581G. The main trajectory to S436A,A437G,K540E

occurs via pathway A437G/K540E/S436A with a second less likely pathway A437G/S436A/K540E. The main trajectory to A437G,K540E,A581G occurs via pathway A437G/K540E/A581G.

In Papua New Guinea, most isolates (56.3%) were the wild-type allele. The most common mutations were single mutation A437G and double mutation A437G,K540E, which were observed in 5.9% and 29.4% of isolates, respectively. All other combinations of the four mutations were absent from the isolates. This suggests the trajectory to the double mutation follows the pathway A437G/K540E.

Predicting the most likely evolutionary pathways to resistant mutations in *Pv*DHPS

Homologous mutations in *Pv*DHPS to those in *Pf*DHPS have been reported that confer resistance to sulfadoxine, such as S382F/A/C, A383G, K512E/M/T and A553G, which correspond to *Pf*DHPS mutations S436F/A, A437G, K540E and A581G respectively [16, 49]. It has been suggested that these homologous mutations, most notably A383G and A553G, confer sulfadoxine resistance in *P. vivax* [50]. Pornthanakasem et al. (2016) [51] reported 30-fold decrease in sulfadoxine binding affinity upon single mutation A383G relative to the wild-type allele, and a further almost 4-fold decrease for the double mutation A383G,A553G relative to A383G, and an almost 5-fold decrease for the triple mutation S328A,A383G,A553G relative to A383G.

The single mutation A383G, along with double A383G,A553G and triple mutant S328A,A383G,A553G have been observed in 90% of mutations in areas where malaria is endemic in Thailand [52]. It is also thought that, like A437G, A383G is likely to be the initial mutation and may be necessary for the appearance of subsequent mutations, as it is most often found in field isolates [49, 52]. In fact, A383G is now so frequent, it its considered fixed in the population, much like its homologous mutation A437G. In our isolate data, the wild-type allele contains a Glycine at position 383, however to test our method we chose Ala-383 as the wild-type residue, and simulated evolutionary trajectories including the A383G mutation.

In our isolate data, mutation Met-205 to Ile (M205I) occurs frequently in combination with A383G, A553G and S382A/C. M205I has been previously identified by other studies as a common polymorphism occurring *P. vivax* [49, 51], however it is unknown if this mutation confers resistance to sulfadoxine as it occurs in the PPPK gene of the bifunction enzyme PPPK-DHPS. It may be possible, however, that M205I interacts via long-range epistasis with the three other aforementioned *Pv*DHPS sulfadoxine resistance mutations and so we chose to predict the most likely evolutionary pathways to quadruple mutant M205I,S382A,A383G,A553G which was observed at high frequency in our isolate data

We generated Flex ddG predictions of the change in sulfadoxine binding free energy for the combinatorically complete set of the four aforementioned mutations (M205I, S382A, A383G, A553G) and also the single mutation V585A. We then compared the peak of the predicted distributions for single mutations A383G and V585A, double mutation A383G,A553G and triple mutation S328A,A383G,A553G to the data presented in [51], and observed a Pearson's correlation of 0.94 with ¾ of the mutations correctly classified as stabilizing or destabilizing (Table 3). Despite the high correlation, the predicted destabilizing effect on sulfadoxine binding free energy for the triple mutation S382A,A383G,A553G was less than both single mutation A383G and double mutation A383G,A553G. This is at odds with the data, where the triple mutation is the most destabilizing mutation of the set studied.

Comparing Flex ddG predictions, A383G was the most destabilizing single mutation compared to M205I, S382A and A553G, and was predicted to be 7-times more destabilizing than the second most destabilizing mutation, A553G. Double mutation A383G,A553G was the most destabilizing double mutation and triple mutation M205I,S382A,A553G was the most destabilizing triple mutation.

We determined the non-additivity between the impact of multiple mutations and the sum of their independent impacts to investigate epistasis between the mutations (see Supplementary data, 'PvDHPS_epistasis.xlsx'). Only one mutation, M205I,A553G appears to be additive, whilst all

other multiple mutations were predicted to be non-additive, suggesting epistasis. All multiple mutations except double mutation S382A,A553G, triple mutation S382A,A383G,A553G and quadruple mutation M205I,S382A,A383G,A553G were more destabilizing than expected. The largest non-additive interactions were for mutations M205I,S382A,A553G and M205I,S382A,A383G,A553G, where M205I,S382A,A553G was predicted to be 9-times more destabilizing than expected, and M205I,S382A,A383A,A553G was predicted to be almost 10-times less destabilizing than expected. Interestingly, quadruple mutation M205I,S382A,A383G,A553G, was predicted to be less destabilizing than single mutation A383G alone.

Single mutation M205I occurs in the PPPK gene and is not in direct contact with sulfadoxine and on its own Flex ddG predicts it to have a mostly neutral impact on sulfadoxine binding affinity ($\Delta\Delta G^{*}_{M205I}$=-0.03 (R.E.U)). However, it is predicted to interact epistatically with all mutations, except A553G and A383G,A553G although there are no direct interactions between M205I and the other mutations, suggesting long-range epistasis.

We simulated possible evolutionary trajectories to the quadruple mutation M205I,S382A,A383G,A553G (Figure 7a). The most likely first step in the pathway is single mutation A383G, in agreement with the observation that A383G is now a major allele. The two most likely pathways to the quadruple mutation are A383G/M205I/S382A/A553G and A383G/M205I/A553G/S382A respectively.

Analyzing the frequency of each mutation in our isolate data (Figure 7b), M205I is the most frequent single mutation and A383G is the second most frequent, whereas A383G and A553G are the two most likely single mutations in our simulations (Supplementary data 'PvDHPS_pathway_probabilities.csv'). The most frequent double mutations in the isolate data and the most likely double mutations in our evolutionary simulations were M205I,A383G and A383G,A553G, when summing over all possible pathways. The most frequent triple mutation in the isolate data, M205I,A383G,A553G, was the second most likely triple mutation in our

simulations, considering all pathways. The most likely triple mutation in our simulations, M205I,S382A,A383G, corresponds to the second most frequent triple mutation in the isolate data.

To compare our simulations to the geographical distribution of mutations, we grouped the data into five rough regions: South America (Brazil, Colombia, Guyana, Mexico, Panama, Peru), East Africa (Ethiopia, Eritrea, Madagascar, Sudan, Uganda), South Asia (Afghanistan, Bangladesh, India, Pakistan, Sri Lanka), Southeast Asia (Cambodia, China, Laos, Malaysia, Myanmar, Philippines, Thailand and Vietnam) and Melanesia (Papua New Guinea).

If we consider the geographical distribution of *Pv*DHPS mutations (Figure 8), we can see that the dominant mutations differ between regions. In Southeast Asia, many combinations of the set of four mutations M205I, S382A, A383G and A553G are observed. Single mutations M205I, A383G and A553G were and observed in 1.2%, >1% and 5.9% of isolates, respectively. Double mutations M205I,A383G, S382A,A383G and A383G,A553G were observed in 28.2%, >1% and >1% of isolates, respectively. Triple mutations M205I,S382A,A383G, M205I,A383G,A553G and S382A,A383G,A553G were observed in 1.2%, 26.1% and >1%, respectively, and the quadruple mutation was observed in 11.3% of isolates. This suggests the evolutionary trajectory to the quadruple mutation occurs main pathways M205I/A383G/A553G/S382A, with possible alternative pathway A383G/M205I/A553G/S382A.

However, only a small number of the possible combinations of the four mutations were observed in South Asia, South America, East Africa and Papua New Guinea. In South Asia, mutations A383G, A553G and A383G,A553G were observed in 2.6%, 1.8% and 12.3% of isolates, respectively. In South America, mutations A383G, M205I and M205I,A383G were observed in 6.5%, 15.5% and 26.4%, respectively. In East Africa, mutations A383G and M205I were observed in 3.6% and 5.9% of isolates, respectively, and in Papua New Guinea, single mutation M205I was observed in 15.4% of isolates.

Other frequent multiple mutations include M205I,S382C,A383G which was found in 11.2% of South America isolates, E142G,M205I and E142G,M205I,A383G,A647V which were found in

36.9% and 21.4% of East African isolates, respectively and E132G,A383G,A553G which was found in 8.9% of Southeast Asian isolates.

Discussion

We have presented a method for predicting the most likely evolutionary trajectories to multiple mutants, utilizing Rosetta Flex ddG protocol and a thermodynamic evolutionary model. The most likely pathways predicted by our model to the pyrimethamine-resistant quadruple *Pf*DHFR mutant correspond well to those predicted by Lozovsky et al. (2009) [7]. They used experimentally determined IC50 values of *Pf*DHFR pyrimethamine binding for the combinatorically complete set of the four *Pf*DHFR mutations (N51I, C59R, S108N, I164L), combined with information regarding the nucleotide bias of the *P. falciparum* genome, to simulate the evolutionary trajectories. The three most likely pathways based on experimental IC50 values were found in the top four most likely pathways based on our simulations using predictions of binding free energy.

We also simulated the most likely evolutionary trajectories to the *Pv*DHFR quadruple mutation N50I,S58R,S117N,I173L and compared our results to those of Jiang et al. (2013) [8]. They considered the relative growth rates of the different alleles at different drug concentrations when simulating evolutionary trajectories, which incorporate both change in pyrimethamine binding affinity (K_i) and catalytic activity (k_{cat}). Our top two most likely pathways correspond to their top two most likely pathways for the highest pyrimethamine concentration they consider, albeit in reverse order. At high pyrimethamine concentrations, it is likely alleles which significantly reduce binding affinity will be selectively favoured even if there is a slight reduction in catalytic activity. This may be why our predictions agree well their predictions for high pyrimethamine concentration, but not for low-to-middle pyrimethamine concentrations, because even though ligand concentration is included in our equation for protein fitness (Eq. 1), our model cannot account for adaptive conflict between K_i and k_{cat}.

This highlights a limitation of our method as it only accounts for changes in binding affinity and does not account for changes in protein function. As previously mentioned, DHFR catalyzes the reduction of substrate DHF via oxidation of cofactor NADPH. Therefore, in the case of the DHFR enzyme, a future iteration of the model could include the impact resistance mutations have on binding of these two ligands, as a proxy for changes to enzyme function. However, this would require a much more complex model of protein fitness and would be much more computationally expensive.

Mutations occurring at a drug-binding site may also reduce the protein's thermodynamic stability [53] and therefore may not be selected for, even if they improve the resistance phenotype. Therefore, our model may also be improved by including selection for mutations that do not reduce thermodynamic stability relative to the wild-type enzyme. There are several computational methods to predict changes in protein stability upon mutation, including mCSM [54], Rosetta Cartesian ddG [55] and FoldX [56]. However, it must also be noted that most proteins are marginally stable [57-59], a property which may have evolved either as an evolutionary spandrel [60, 61] (a characteristic that arises as a result of non-adaptive processes which is then used for adaptive purposes [62]) or due to selection for increased flexibility to improve certain functionalities [63, 64]. Therefore, the model would also have to account for the fact that a resistance mutation that increases protein stability relative to the wild-type stability may also result in a reduction in fitness.

The quadruple *Pv*DHFR mutation (N50I,S58R,S117N,I173L) has not been noted in the literature or in our clinical isolate data. This may be because *P. vivax* is only exposed to pyrimethamine when present in co-infections with *P. falciparum* and so is not under continued selection for pyrimethamine resistance. Furthermore, the quadruple mutation only reached fixation in the evolutionary simulations described in [8] for the highest pyrimethamine concentrations, therefore clinical dosages of pyrimethamine may not be high enough to select for the quadruple mutation.

Additionally, the quadruple mutation may result in a fitness impairment that requires compensatory mutations.

Despite its limitations, the evolutionary trajectories predicted using our method agree well with experimentally predicted trajectories for both *Pf*DHFR and *Pv*DHFR, and so we applied our method to predict the evolutionary trajectories to quadruple mutants in both *Pf*DHPS and *Pv*DHPS that may confer resistance to sulfadoxine. Evolutionary trajectories to multiple resistance mutations in these enzymes have not yet been investigated fully, either experimentally or computationally, although attempts have been made to infer the first step in sulfadoxine resistance. Therefore, our work presents the first attempt to predict evolutionary trajectories in these enzymes.

As we can see from the geographical distributions of mutations in *Pf*DHPS (Figure 6) and *Pv*DHPS (Figure 8), there is much more variation between different regions than was observed for the DHFR gene from both species (Figure 2 and Figure 4 for *Pf*DHFR and *Pv*DHFR, respectively). For both species, the quadruple mutations to which we simulated evolutionary trajectories are only observed in Southeast Asia, a region of high drug pressure Furthermore, whilst we have only considered the trajectories for one multiple mutant for each gene, it may be beneficial to consider one per region for the DHPS gene from both species.

Oguike et al. (2016) [65] reported the emergence of a further *Pf*DHPS sulfadoxine resistance mutation Ile-431 to Val (I431V) in Nigeria and found it occurred most commonly as a quintuple mutant with S436A, A437G, A581G and A613S. This quintuple mutation is also found frequently in our isolate data from West Africa, along with single, double, triple and quadruple combinations of the mutations involved. Therefore, it may also be useful to study the evolutionary trajectories to this quintuple mutation.

We might also consider quadruple *Pv*DHPS mutation E142G,M205I,A383G,A647V, which is highly prevalent in East Africa. However, the mutations at codons 142 and 205 occur in the *Pv*PPPK gene of the bifunctional enzyme PPPK-DHPS and are not directly involved in sulfadoxine binding. Therefore, this mutation may not result in a resistant phenotype.

We also inferred evolutionary pathways from our clinical isolate data from different geographical regions for each drug-target combination studied here. This analysis suggested that different regions often follow different evolutionary trajectories and that the most likely evolutionary trajectories predicted by our model are not always the most prevalent.

Geographical differences in the distribution of resistant alleles may be the result of drug regimens and gene flow in parasite populations. If we consider the geographical distribution of *Pf*DHFR mutations (Figure 2), we can see that triple mutant N51I,C59R,S108N is common in West and East Africa as well as Southeast Asia, whereas the quadruple mutation N51I,C59R,S108N,I164L is only observed at high frequency in Southeast Asia. Combination drug SP was first used in 1967 to treat *P. falciparum* in Southeast Asia, and resistance was first noted that same year on the Thai-Cambodia and Thai-Myanmar borders [66]. In Africa, SP was first used in the 1980s, with resistance occurring later that decade. However, analysis of *Pf*DHFR genotypes and microsatellite haplotypes surrounding the DHFR gene in Southeast Asia and Africa suggest a single resistant lineage that appeared in Southeast Asia accumulated multiple mutations, including the triple N51I,C59R,S108N [67-69], migrated to Africa and spread throughout the continent [70-72]. Variation in the frequency of *Pf*DHFR (and *Pf*DHPS) mutants across Africa occurs because of differences in the timing of chloroquine withdrawal and introduction of SP, as well as continued use of SP for intermittent preventive treatment (IPTp) in pregnant women residing in areas of moderate to high malaria transmission intensity [48, 73].

Pyrimethamine resistance increased in West Papua in the early 1960s following the introduction of mass drug administration [74]. In our data, the distribution in Papua New Guinea is made up

mainly by the double mutant C59R,S108N and triple mutant C59R,S108N,S306F. Microsatellite haplotype analysis suggests C59R,S108N in Melanesia has two lineages, one of which originated in Southeast Asia whilst the other evolved indigenously [69].

Pyrimethamine resistance in South America looks surprisingly different from the distributions in Africa and Southeast Asia. SP was introduced in South America and low-level resistance was first noted in Colombia in 1981 [75]. Microsatellite haplotype analysis suggests pyrimethamine resistance evolved indigenously in South America, with at least two distinct lineages detected. A triple mutant lineage (C50I,N51I,S108N) was identified in Venezuela that possibly evolved from double mutant N51I,S108N [71]. A second triple mutant lineage (N51I,S108N,I164L) was identified in Peru and Bolivia which also possibly evolved from a distinct double mutant (N51I,S108N) lineage [76].

The distribution of *Pv*DHFR pyrimethamine-resistance mutations (Figure 4) is much more diverse than the *Pf*DHFR distribution (Figure 2), despite the lower number of *Pv*DHFR isolates. However, the opposite is true for the DHPS gene, with a greater diversity of mutations found the *P. falciparum* isolates.

In general, the *Pv*DHFR gene is much more polymorphic than *Pf*DHFR gene, with over 20 alleles observed in a limited geographical sampling [25], whereas fewer *Pf*DHFR alleles have been observed despite much more extensive surveillance with non-synonymous changes and insertions/deletions occurring rarely [77]. It also appears that the origin of *Pv*DHFR pyrimethamine resistance mutation is much more diverse than *Pf*DHFR. Hawkins et al. (2008) [78] investigated isolates from Colombia, India, Indonesia, Papua New Guinea, Sri Lanka, Thailand and Vanuatu. They found multiple origins of the double *Pv*DHFR mutant 58R,117N and three independent origins of triple mutant 58R,61M,117T and quadruple mutants 57I,61M,117T,173F and 57L,58R,61M,117T in Thailand, Indonesia and Papua New Guinea/Vanuatu. Shaukat et al. (2021) [79] assessed the evolutionary origin of *Pv*DHFR pyrimethamine resistance mutations in Punjab, Pakistan and found multiple origins of single

mutation S117N and a common origin of double mutant 58R,S117N and triple mutants 57L,58R,117N, 58R,61M,117N and 58R,118N,I173L. This is in contrast to the evolutionary origin of pyrimethamine resistance in *Pf*DHFR, where mutations in Africa shared a common origin with a resistance lineage from Asia.

The distribution of *Pf*DHPS sulfadoxine resistance mutations is strikingly different across different geographical regions (Figure 6). Microsatellite haplotype analysis of resistant alleles revealed single, double and triple sulfadoxine resistance mutations arose independently in Cambodia, Kenya, Cameroon and Venezuela [71, 80]. Microsatellite analysis of *Pf*DHPS in Africa indicated multiple origins of single mutant (A437G) and double mutant (S436A,A437G and A437G,K540E) and different resistance lineages when comparing east and west Africa, resulting in differing distributions of resistance mutations [81].

Similarly, multiple origins of *Pv*DHPS resistance mutations at codons 382, 383 and 553 were observed in Cambodia, Indonesia, Sri Lanka and Thailand [49], which may account for the differences in the distribution of mutations between geographical regions (Figure 8).

This highlights the need to distinguish between geographical regions and account for existing resistance alleles within that region and trace their lineages when attempting to predict the next step in evolutionary trajectories to highly resistant multiple mutants. Given the current dominant resistance allele from a specific region, our method could be used to predict the most likely next steps from a subset of likely mutations.

We have presented a new computational method for predicting the most likely evolutionary trajectories that has demonstrated good agreement with trajectories predicted experimentally and has the advantage of being much quicker and more cost-effective. This method can be applied to any system in which a drug binds to a target molecule, provided a structure of the complex exists or can be produced via structural modelling. Given the threat antimicrobial resistance poses, methods to accurately and efficiently predict future trajectories are vital and can inform treatment strategies and aid drug development.

Materials and Methods

Homology modelling

Homology modelling was carried out in Modeller [82] to produce complete structures of the target proteins bound to their drug molecules.

Several crystal structures of *Pf*DHFR exist in the Protein Data Bank (PDB). The entry 3QGT provides the crystal structure of wild-type *Pf*DHFR complexed with NADPH, dUMP and pyrimethamine, however residues in the ranges 86-95 and 232-282 are missing from the structure. Homology modelling was used to complete the structure using a second wild-type *Pf*DHFR structure PDB entry 1J3I along with a wild-type *Pv*DHFR structure PDB entry 2BLB.

To produce a complete structure of *Pv*DHFR, PDB entry 2BLB was used as a template, which provides the X-ray crystal structure of wild-type *P. vivax* DHFR in complex with pyrimethamine. This structure was only missing a loop section between residues 87-105 and so Modeller was used to build this missing loop.

To produce a complete structure of *Pf*DHPS, PDB entry 6JWX was used as a template, which provides the crystal structure of wild-type *Pf*DHPS-HPPK in complex with sulfadoxine. This structure is missing a number of loop sections, so Modeller was used to build these loops using additional wild-type *Pf*DHPS template 6JWR and *P. vivax* HPPK-DHPS template 5Z79.

To produce a complete structure of *Pv*DHPS, PDB entry 5Z79 was used as a template, which provides the crystal structure of wild-type *P. vivax* HPPK-DHPS. This structure does not include sulfadoxine, however, and is missing several important loop sections. Modeller was used to build the structure complete with SDX-DHP and completed loops using additional templates 6JWX and 6JWA.

Flex ddG binding free energy predictions

The Rosetta Flex ddG protocol was used to estimate the change in binding free energy upon mutation, $\Delta\Delta G = \Delta G_{mut} - \Delta G_{WT}$, for each step in all possible mutational trajectories for a set of stepwise resistance mutations (see supplementary data 'Flex_ddG' folder for example Rosetta script, example resfile and example command line. The protein-ligand structure files and ligand parameter files can be found in the folders named for the specific targets). To predict the change in binding free energy for a single or multiple mutation, we used the structure of the target protein with the drug molecule bound as input to Flex ddG and ran the protocol for 150 times per mutation to produce a distribution of predictions of the change in the free energy of binding. We then found the peak of the distribution to produce a single estimate of the change in the binding free energy for the mutation, denoted $\Delta\Delta G_X^*$ for mutation *X*.

To predict the stepwise evolutionary trajectories, we must consider the interactions between the mutations in the pathway. The interaction energy (or epistasis) in the binding free energy between two mutations *X* and *Y,* can be written $\varepsilon_{XY} = \Delta\Delta G_{X,Y} - (\Delta\Delta G_X + \Delta\Delta G_Y)$. This quantifies by how much the change in binding free energy of the double mutant *X,Y* deviates from additivity of the single mutants, where each are calculated with respect to the wild-type. Therefore, the change in binding free energy when mutation *Y* occurs in the background of mutation *X* can be written $\Delta\Delta G_{X/Y} = \Delta\Delta G_{X,Y} - \Delta\Delta G_X$, where $\Delta\Delta G_{X/Y} = \Delta\Delta G_Y + \varepsilon_{XY}$.

For a third mutation, *Z*, occurring in the background of double mutation *X,Y*, the interaction energy between *Z* and *X,Y* is $\varepsilon_{XY,Z} = \Delta\Delta G_{X,Y,Z} - (\Delta\Delta G_{X,Y} + \Delta\Delta G_Z)$. The quantity $\varepsilon_{XY,Z}$ is not the same as the third order epistasis between mutations *X*, *Y*, and *Z,* or the interaction energy $\varepsilon_{XYZ} = \Delta\Delta G_{X,Y,Z} - (\Delta\Delta G_X + \Delta\Delta G_Y + \Delta\Delta G_Z)$, as it does not account for the interaction between *X* and *Y,* rather it only quantifies the interaction between *Z* and the two mutations *X* and *Y.*.

Therefore, the change in binding free energy when mutation *Z* occurs in the background of double mutant *X,Y* can be calculated as $\Delta\Delta G_{X,Y/Z} = \Delta\Delta G_{X,Y,Z} - \Delta\Delta G_{X,Y}$, where $\Delta\Delta G_{X,Y/Z} = \Delta\Delta G_Z + \varepsilon_{XY,Z}$.

To estimate the change in binding free energy when mutation *Y* occurs in the background of mutation *X*, $\Delta\Delta G_{X/Y}$ for stepwise pathway *X/Y*, we subtracted the predictions $\Delta\Delta G_X^i$ for the first mutation *X,* from the predictions for the double mutation *X,Y,* $\Delta\Delta G_{X,Y}^i$, to create a set of 150 'predictions' for the change in binding free energy when *Y* occurs in the background of *X,* $\Delta\Delta G_{X/Y}^i$ i.e. $\Delta\Delta G_{X/Y}^i = \Delta\Delta G_{X,Y}^i - \Delta\Delta G_X^i$ for $i = \{1, \dots, 150\}$. To estimate the change in binding free energy when mutation *Z* occurs in the background of mutations *X* and *Y* we calculated. $\Delta\Delta G_{X,Y/Z}^i = \Delta\Delta G_{X,Y,Z}^i - \Delta\Delta G_{X,Y}^i$. We applied a similar method for the quadruple mutations, so that we had a set of 'predictions' for each step in the possible evolutionary trajectories.

Simulating Evolutionary Trajectories

To predict the most likely evolutionary trajectories to reach a quadruple mutant we used a model based in thermodynamics and statistical mechanics where the fitness of a protein is determined by the probability it would not be bound to a ligand, $P_{unbound}$. We consider a two-state system in which the protein can either be bound or unbound and do not explicitly account for if the protein is folded or unfolded in either the bound or unbound state. For ligand concentration $[L]$ it can be shown that the probability a protein is unbound is

$$P_{unbound} = \frac{1}{\frac{[L]}{K_d} + 1} \tag{1}$$

where K_d is the protein-ligand dissociation constant and can be calculated as $c_0 e^{\Delta G/kT}$ where c_0 is a reference ligand concentration (set here arbitrarily to 0.6M), ΔG is the protein-ligand binding free energy, *k* is the Boltzmann constant and *T* is the temperature in Kelvin.

Starting from the wild-type protein, with binding free energy ΔG_{WT} and fitness $P_{unbound}^{WT}$, we extract one sample *i* from the 150 values of the predicted binding affinity changes for the single mutations to determine the binding free energy after mutation *X,* $\Delta G_X^i = \Delta G_{WT} + \Delta\Delta G_X^i$, and

calculate the fitness of each single mutant protein $P_{unbound}^{X(i)}$. We can calculate the probability the mutation will fix in the population using the Kimura fixation probability

$$p_{fix} = \frac{1 - e^{-2s}}{1 - e^{-4sN_e}} \quad (2)$$

where N_e is the effective population size and s is the selection coefficient $s = (P_{unbound}^{X,i} - P_{unbound}^{WT})/P_{unbound}^{WT}$. We also took in to account the mutational bias of *Plasmodium falciparum* using the nucleotide mutation matrix calculated in [7]. The probabilities of fixation for each mutation were normalised by the sum of the probabilities of fixation for all possible mutations at that step in the trajectory. A mutation is then chosen with a probability proportional to this normalised probability of fixation.

Once a single mutation is chosen, the binding free energy is set to ΔG_X^i of the chosen mutation, and a value is sampled from the distribution of each of the possible next steps, *X/Y* in the trajectory i.e. $\Delta\Delta G_{X/Y}^i$. This continues until the end of the trajectory is reached.

SNP data

Plasmodium falciparum and *vivax* SNP data were sourced from recent studies [45, 48]. In those studies, paired Illumina raw sequence data was mapped to the Pf3D7 (P. falciparum) or PVP01 (*P. vivax*) reference genome using bwa-mem software (default parameters). SNPs were called using the samtools and GATK software suites. Those SNPs occurring in non-unique, low quality or low coverage regions were discarded, and those in the candidate genes analysed here were extracted.

Acknowledgments

R.C.E and N.F. are funded by the Medical Research Council UK (Grant no. MR/T000171/1). T.G.C is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). S.C is funded by Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and MR/R020973/1) and

Bloomsbury SET. EM is funded by a Newton Institutional Links Grant (British Council, no. 261868591. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. R.C.E would like to thank Tanushree Tunstall for her helpful discussions.

References

1. Sun, D., et al., *Editorial: Horizontal Gene Transfer Mediated Bacterial Antibiotic Resistance.* Frontiers in microbiology, 2019. **10**: p. 1933-1933.
2. Davies, J. and D. Davies, *Origins and evolution of antibiotic resistance.* Microbiology and molecular biology reviews : MMBR, 2010. **74**(3): p. 417-433.
3. Levy, S.B. and B. Marshall, *Antibacterial resistance worldwide: causes, challenges and responses.* Nature Medicine, 2004. **10**(12): p. S122-S129.
4. Blair, J.M.A., et al., *Molecular mechanisms of antibiotic resistance.* Nature Reviews Microbiology, 2015. **13**(1): p. 42-51.
5. Weinreich, D.M., et al., *Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins.* Science, 2006. **312**(5770): p. 111.
6. Brown, K.M., et al., *Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase.* Molecular Biology and Evolution, 2010. **27**(12): p. 2682-2690.
7. Lozovsky, E.R., et al., *Stepwise acquisition of pyrimethamine resistance in the malaria parasite.* Proc Natl Acad Sci, U. S. A., 2009. **106**(29): p. 12025-12030.
8. Jiang, P.-P., et al., *Accessible Mutational Trajectories for the Evolution of Pyrimethamine Resistance in the Malaria Parasite Plasmodium vivax.* Journal of Molecular Evolution, 2013. **77**(3): p. 81-91.
9. Sirawaraporn, W., et al., *Antifolate-resistant mutants of Plasmodium falciparum dihydrofolate reductase.* Proc Natl Acad Sci U S A, 1997. **94**: p. 1124-1129.
10. Wellems, T.E., *Plasmodium Chloroquine Resistance and the Search for a Replacement Antimalarial Drug.* Science, 2002. **298**(5591): p. 124-126.
11. Wang, P., et al., *Sulfadoxine resistance in the human malaria parasite Plasmodium falciparum is determined by mutations in dihydropteroate synthetase and an additional factor associated with folate utilization.* Mol Microbiol., 1997. **23**(5): p. 979-986.
12. Organization, W.H., *World Malaria Report: 2020*. 2020, World Health Organization, Geneva, Switzerland.
13. Brooks, D.R., et al., *Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, Plasmodium falciparum, with differing resistance to sulfadoxine.* Eur, J. Biochem, 1994. **224**(2): p. 397-405.
14. Thaithong, S., et al., *Pyrimethamine resistant mutations in Plasmodium falciparum.* Mol Biochem, Parasitol, 1992. **52**(0166-6851 (Print)): p. 149-157.
15. Snounou, G. and N.J. White, *The co-existence of Plasmodium: sidelights from falciparum and vivax malaria in Thailand.* Trends in Parasitology, 2004. **20**(7): p. 333-339.
16. Korsinczky, M., et al., *Sulfadoxine resistance in Plasmodium vivax is associated with a specific amino acid in dihydropteroate synthase at the putative sulfadoxine-binding site.* Antimicrob Agents, Chemother, 2004. **48**(6): p. 2214-2222.

17. Imwong, M., et al., *Association of Genetic Mutations inPlasmodium vivax dhfr with Resistance to Sulfadoxine-Pyrimethamine: Geographical and Clinical Correlates.* Antimicrobial Agents and Chemotherapy, 2001. **45**(11): p. 3122-3127.
18. Leartsakulpanich, U., et al., *Molecular characterization of dihydrofolate reductase in relation to antifolate resistance in Plasmodium vivax.* Molecular and Biochemical Parasitology, 2002. **119**(1): p. 63-73.
19. Hastings, M.D., et al., *Dihydrofolate Reductase Mutations in Plasmodium vivax from Indonesia and Therapeutic Response to Sulfadoxine plus Pyrimethamine.* The Journal of Infectious Diseases, 2004. **189**(4): p. 744-750.
20. Kompis, I.M., R.L. Islam K Fau - Then, and R.L. Then, *DNA and RNA synthesis: antifolates.* Chem, Rev, 2005. **105**(2): p. 593-620.
21. Jiang, T., et al., *High prevalence of Pfdhfr-Pfdhps quadruple mutations associated with sulfadoxine-pyrimethamine resistance in Plasmodium falciparum isolates from Bioko Island, Equatorial Guinea.* Malar J., 2019. **18**(1): p. 101.
22. Berglez, J., et al., *Analysis in Escherichia coli of Plasmodium falciparum dihydropteroate synthase (DHPS) alleles implicated in resistance to sulfadoxine.* Int, J. Parasitol 2004. **34**(1): p. 95-100.
23. Ferlan, J.T., et al., *Mutagenesis of dihydrofolate reductase from Plasmodium falciparum: analysis in Saccharomyces cerevisiae of triple mutant alleles resistant to pyrimethamine or WR99210.* Mol Biochem, Parasitol, 2001. **113**(0166-6851 (Print)): p. 139-150.
24. Yuthavong, Y., et al., *Malarial (Plasmodium falciparum) dihydrofolate reductase-thymidylate synthase: structural basis for antifolate resistance and development of effective inhibitors.* Parasitology,, 2005. **130**(0031-1820 (Print)): p. 249-259.
25. Hawkins, V.N., et al., *Antifolates can have a role in the treatment of Plasmodium vivax.* Trends, Parasitol, 2007. **25**(5): p. 213-222.
26. Tamer, Y.T., et al., *High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a Rugged Fitness Landscape in the Presence of Trimethoprim Selection.* Molecular Biology and Evolution, 2019. **36**(7): p. 1533-1550.
27. Starr, T.N. and J.W. Thornton, *Epistasis in protein evolution.* Protein science : a publication of the Protein Society, 2016. **25**(7): p. 1204-1218.
28. Khan, A.I., et al., *Negative epistasis between beneficial mutations in an evolving bacterial population.* Science, 2011. **332**(6034): p. 1193-1196.
29. Gong, L.I., M.A. Suchard, and J.D. Bloom, *Stability-mediated epistasis constrains the evolution of an influenza protein.* eLife, 2013. **2**: p. e00631.
30. Sanjuán, R., et al., *Epistasis and the adaptability of an RNA virus.* Genetics, 2005. **170**: p. 1001-1008.
31. Rosenthal, P.J., *The interplay between drug resistance and fitness in malaria parasites.* Mol, Microbiol, 2013. **89**: p. 1025-1038.
32. Melnyk, A.H., A. Wong, and R. Kassen, *The fitness costs of antibiotic resistance mutations.* Evol, Appl, 2015. **8**(3): p. 273-283.
33. Vogwill, T. and R.C. MacLean, *The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach.* Evolutionary applications, 2015. **8**(3): p. 284-295.
34. Costelloe, C., et al., *Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis.* BMJ, 2010. **340**: p. c2096.

35. Enne, V.I., *Reducing antimicrobial resistance in the community by restricting prescribing: can it be done?* Journal of Antimicrobial Chemotherapy, 2010. **65**(2): p. 179-182.
36. Sundqvist, M., et al., *Little evidence for reversibility of trimethoprim resistance after a drastic reduction in trimethoprim use.* J. Antimicrob Chemother, 2010. **65**: p. 350-360.
37. Andersson, D.I., *Persistence of antibiotic resistant bacteria.* Curr Opin Microbiol., 2003. **6**: p. 452-456.
38. Barlow, K.A., et al., *Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation.* The Journal of Physical Chemistry B, 2018. **122**(21): p. 5389-5399.
39. Alford, R.A.-O., et al., *The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J. Chem Theory Comput, 2017. **13**: p. 3031-3048.
40. Aldeghi, M., V. Gapsys, and B.L. de Groot, *Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation.* ACS central science, 2018. **4**(12): p. 1708-1718.
41. Aldeghi, M., V. Gapsys, and B.L. de Groot, *Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches.* ACS Central Science, 2019. **5**(8): p. 1468-1474.
42. Eccleston, R.C., D.D. Pollock, and R.A. Goldstein, *Selection for cooperativity causes epistasis predominately between native contacts and enables epistasis-based structure reconstruction.* Proceedings of the National Academy of Sciences, 2021. **118**(16): p. e2010057118.
43. Pollock, D.D., G. Thiltgen, and R.A. Goldstein, *Amino acid coevolution induces an evolutionary Stokes shift.* Proceedings of the National Academy of Sciences, 2012. **109**(21): p. E1352.
44. Pollock, D.D., et al., *Mechanistic Models of Protein Evolution*, in *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*, P. Pontarotti, Editor. 2017, Springer International Publishing: Cham. p. 277-296.
45. Benavente, E.D., et al., *Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa.* Nat Commun, 2021. **12**(1): p. 3160.
46. Triglia, T., et al., *Mutations in dihydropteroate synthase are responsible for sulfone and sulfonamide resistance in Plasmodium falciparum.* Proc Natl Acad Sci, U. S. A., 1997(0027-8424 (Print)): p. 13944-13949.
47. Wang, P., et al., *Resistance to antifolates in Plasmodium falciparum monitored by sequence analysis of dihydropteroate synthetase and dihydrofolate reductase alleles in a large number of field samples of diverse origins.* Mol Biochem, Parasitol, 1997. **89**(0166-6851 (Print)): p. 161-177.
48. Turkiewicz, A., et al., *Genetic diversity of the Plasmodium falciparum GTP-cyclohydrolase 1, dihydrofolate reductase and dihydropteroate synthetase genes reveals new insights into sulfadoxine-pyrimethamine antimalarial drug resistance.* PLoS Genet, 2020. **16**(12): p. e1009268.
49. Hawkins, V.N., et al., *Assessment of the Origins and Spread of Putative Resistance-Conferring Mutations in Plasmodium vivax Dihydropteroate Synthase.* The American Journal of Tropical Medicine and Hygiene Am J Trop Med Hyg, 2009. **81**(2): p. 348-355.
50. Imwong, M., et al., *Limited polymorphism in the dihydropteroate synthetase gene (dhps) of Plasmodium vivax isolates from Thailand.* Antimicrob Agents Chemother, 2005. **49**(10): p. 4393-4395.

953 51. Pornthanakasem, W., et al., *Role of Plasmodium vivax Dihydropteroate Synthase*
954 *Polymorphisms in Sulfa Drug Resistance.* Antimicrob Agents, Chemother, 2016(1098-
955 6596 (Electronic)): p. 4453-4463.
956 52. Rungsihirunrat, K., et al., *Geographical Distribution of Amino Acid Mutations in*
957 *Plasmodium vivax DHFR and DHPS from Malaria Endemic Areas of Thailand.* The
958 American Journal of Tropical Medicine and Hygiene Am J Trop Med Hyg, 2008. **78**(3): p.
959 462-467.
960 53. Wang, X., G. Minasov, and B.K. Shoichet, *Evolution of an antibiotic resistance enzyme*
961 *constrained by stability and activity trade-offs.* J Mol Biol, 2002. **320**(1): p. 85-95.
962 54. Pires, D.E.V., D.B. Ascher, and T.L. Blundell, *mCSM: predicting the effects of mutations in*
963 *proteins using graph-based signatures.* Bioinformatics, 2014. **30**(3): p. 335-342.
964 55. Park, H., et al., *Simultaneous Optimization of Biomolecular Energy Functions on Features*
965 *from Small Molecules and Macromolecules.* Journal of Chemical Theory and
966 Computation, 2016. **12**(12): p. 6201-6212.
967 56. Guerois, R., J.E. Nielsen, and L. Serrano, *Predicting changes in the stability of proteins*
968 *and protein complexes: a study of more than 1000 mutations.* J Mol Biol, 2002. **320**(2): p.
969 369-87.
970 57. Savage, H.J., et al., *Lost hydrogen bonds and buried surface area: rationalising stability in*
971 *globular proteins.* Journal of the Chemical Society, Faraday Transactions, 1993. **89**(15):
972 p. 2609-2617.
973 58. Vogl, T., et al., *Thermodynamic Stability of Annexin V E17G: Equilibrium Parameters*
974 *from an Irreversible Unfolding Reaction.* Biochemistry, 1997. **36**(7): p. 1657-1668.
975 59. Ruvinov, S., et al., *Engineering the Independent Folding of the Subtilisin BPN′*
976 *Prodomain: Analysis of Two-State Folding versus Protein Stability.* Biochemistry, 1997.
977 **36**(34): p. 10414-10421.
978 60. Taverna, D.M. and R.A. Goldstein, *Why are proteins marginally stable?* Proteins:
979 Structure, Function, and Bioinformatics, 2002. **46**(1): p. 105-109.
980 61. Goldstein, R.A., *The evolution and evolutionary consequences of marginal*
981 *thermostability in proteins.* Proteins: Structure, Function, and Bioinformatics, 2011.
982 **79**(5): p. 1396-1407.
983 62. Gould, S.J., et al., *The spandrels of San Marco and the Panglossian paradigm: a critique*
984 *of the adaptationist programme.* Proceedings of the Royal Society of London. Series B.
985 Biological Sciences, 1979. **205**(1161): p. 581-598.
986 63. Závodszky, P., et al., *Adjustment of conformational flexibility is a key event in the*
987 *thermal adaptation of proteins.* Proceedings of the National Academy of Sciences, 1998.
988 **95**(13): p. 7406.
989 64. Tsou, C.L., *Active site flexibility in enzyme catalysis.* Ann N Y Acad Sci, 1998. **864**: p. 1-8.
990 65. Oguike, M.C., et al., *Molecular determinants of sulfadoxine-pyrimethamine resistance in*
991 *Plasmodium falciparum in Nigeria and the regional emergence of dhps 431V.* Int, J.
992 Parasitol Drugs Drug Resist, 2016.
993 66. Björkman, A. and P.A. Phillips-Howard, *The epidemiology of drug-resistant malaria.*
994 Transactions of The Royal Society of Tropical Medicine and Hygiene, 1990. **84**(2): p. 177-
995 180.
996 67. Nair, S., et al., *A selective sweep driven by pyrimethamine treatment in southeast asian*
997 *malaria parasites.* Mol Biol Evol, 2003. **20**(9): p. 1526-36.

998 68. Roper, C., et al., *Intercontinental spread of pyrimethamine-resistant malaria.* Science,
999 2004. **305**(5687): p. 1124.
1000 69. Mita, T., et al., *Independent Evolution of Pyrimethamine Resistance in Plasmodium*
1001 *falciparum Isolates in Melanesia.* Antimicrobial Agents and Chemotherapy, 2007. **51**(3):
1002 p. 1071-1077.
1003 70. Maïga, O., et al., *A Shared Asian Origin of the Triple-Mutant dhfr Allele in Plasmodium*
1004 *falciparum from Sites across Africa.* The Journal of Infectious Diseases, 2007. **196**(1): p.
1005 165-172.
1006 71. McCollum Andrea, M., et al., *Common Origin and Fixation of Plasmodium falciparum*
1007 *dhfr and dhps Mutations Associated with Sulfadoxine-Pyrimethamine Resistance in a*
1008 *Low-Transmission Area in South America.* Antimicrobial Agents and Chemotherapy,
1009 2007. **51**(6): p. 2085-2091.
1010 72. McCollum Andrea, M., et al., *Hitchhiking and Selective Sweeps of Plasmodium*
1011 *falciparum Sulfadoxine and Pyrimethamine Resistance Alleles in a Population from*
1012 *Central Africa.* Antimicrobial Agents and Chemotherapy, 2008. **52**(11): p. 4089-4097.
1013 73. Ravenhall, M., et al., *Characterizing the impact of sustained sulfadoxine/pyrimethamine*
1014 *use upon the Plasmodium falciparum population in Malawi.* Malar J, 2016. **15**(1): p. 575.
1015 74. Verdrager, J., *Epidemiology of the emergence and spread of drug-resistant falciparum*
1016 *malaria in South-East Asia and Australasia.* J Trop Med Hyg, 1986. **89**(6): p. 277-89.
1017 75. Espinal, T.C.A., et al., *Sensitivity of Plasmodium Falciparum to Antimalarial Drugs in*
1018 *Colombia.* The American Journal of Tropical Medicine and Hygiene, 1985. **34**(4): p. 675-
1019 680.
1020 76. Zhou, Z., et al., *Decline in sulfadoxine-pyrimethamine-resistant alleles after change in*
1021 *drug policy in the Amazon region of Peru.* Antimicrob Agents Chemother, 2008. **52**(2): p.
1022 739-41.
1023 77. Gregson, A. and C.V. Plowe, *Mechanisms of Resistance of Malaria Parasites to*
1024 *Antifolates.* Pharmacological Reviews, 2005. **57**(1): p. 117.
1025 78. Hawkins, V.N., et al., *Multiple origins of resistance-conferring mutations in Plasmodium*
1026 *vivax dihydrofolate reductase.* Malaria Journal, 2008. **7**(1): p. 72.
1027 79. Shaukat, A., et al., *Phylogenetic analysis suggests single and multiple origins of*
1028 *dihydrofolate reductase mutations in Plasmodium vivax.* Acta Trop, 2021. **215**: p.
1029 105821.
1030 80. Vinayak, S., et al., *Origin and Evolution of Sulfadoxine Resistant Plasmodium falciparum.*
1031 PLOS Pathogens, 2010. **6**(3): p. e1000830.
1032 81. Pearce, R.J., et al., *Multiple origins and regional dispersal of resistant dhps in African*
1033 *Plasmodium falciparum malaria.* PLoS Med, 2009. **6**(4): p. e1000055.
1034 82. Webb, B. and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr
1035 Protoc Bioinformatics, 2016. **54**: p. 5.6.1-5.6.37.

Figures and Tables

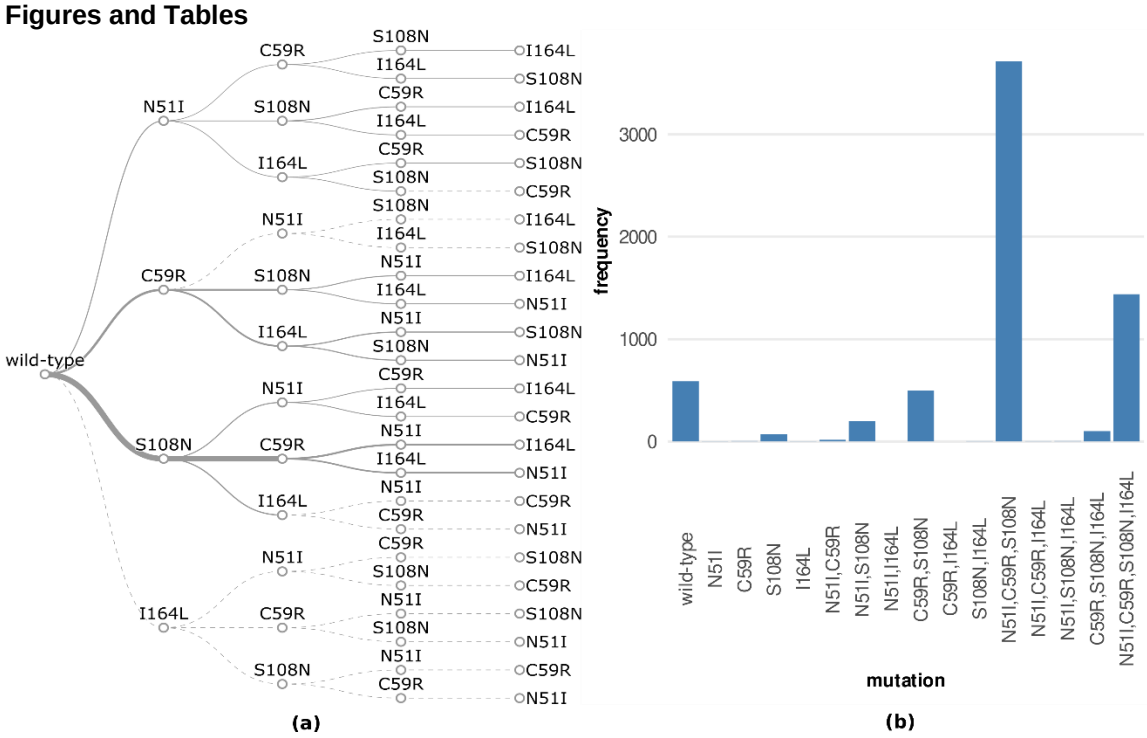


Figure 1. a) Simulated evolutionary pathways to highly-resistance quadruple *Pf*DHFR mutant N51I,C59R,S108N,I164L. Line thickness indicates the likelihood of a mutation at each step. Dotted lines indicate zero probability of a mutation at that step. The two most likely pathways are S108N/C59R/N51I/I164L and S108N/C59R/I164L/N51I respectively, **b)** The frequency of the 16 possible combinations of mutations N51I, C59R, S108N and I164L in *Pf*DHFR, including wild-type, observed in our isolate data.

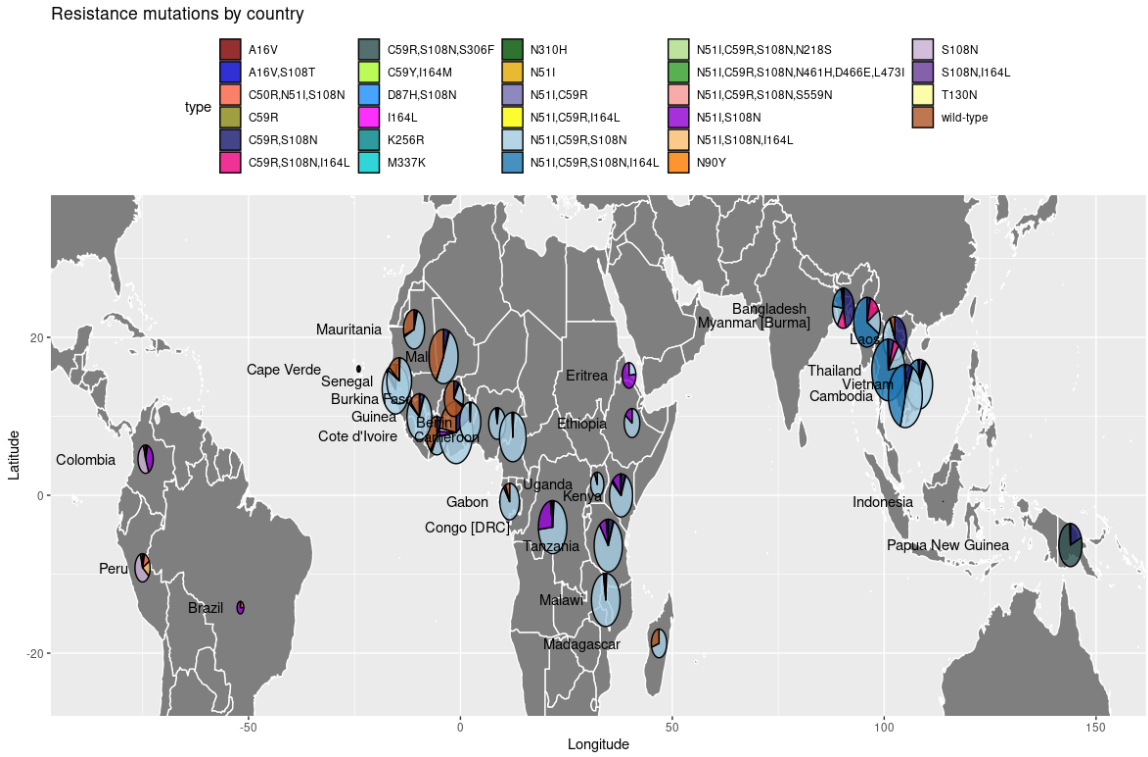
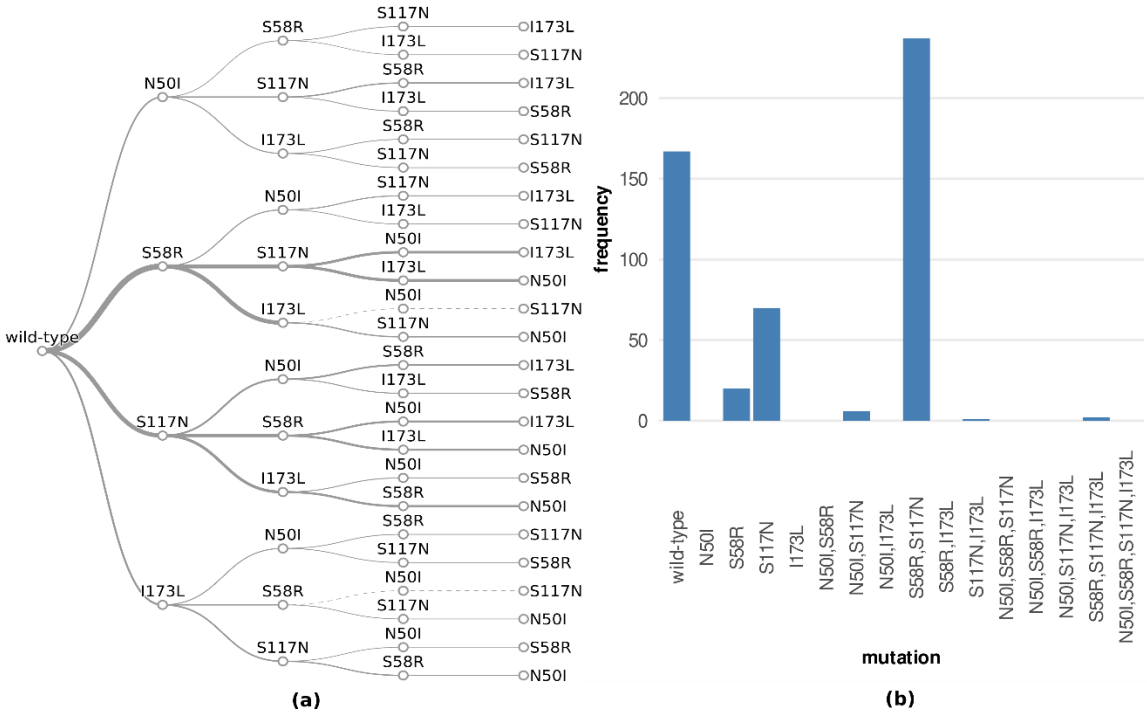


Figure 2. The geographical distribution of mutant *Pf*DHFR alleles found in our isolate data. The size of each pie chart is proportional to the number of isolates from that particular country.


wild-type
N50I
S58R
S117N
I173L
S58R
S117N
I173L
N50I
S117N
I173L
N50I
S58R
I173L
N50I
S58R
S117N
I173L
N50I
S117N
I173L
N50I
S58R
I173L
N50I
S117N
I173L
N50I
S58R
I173L
N50I
S58R
I173L
N50I
S58R

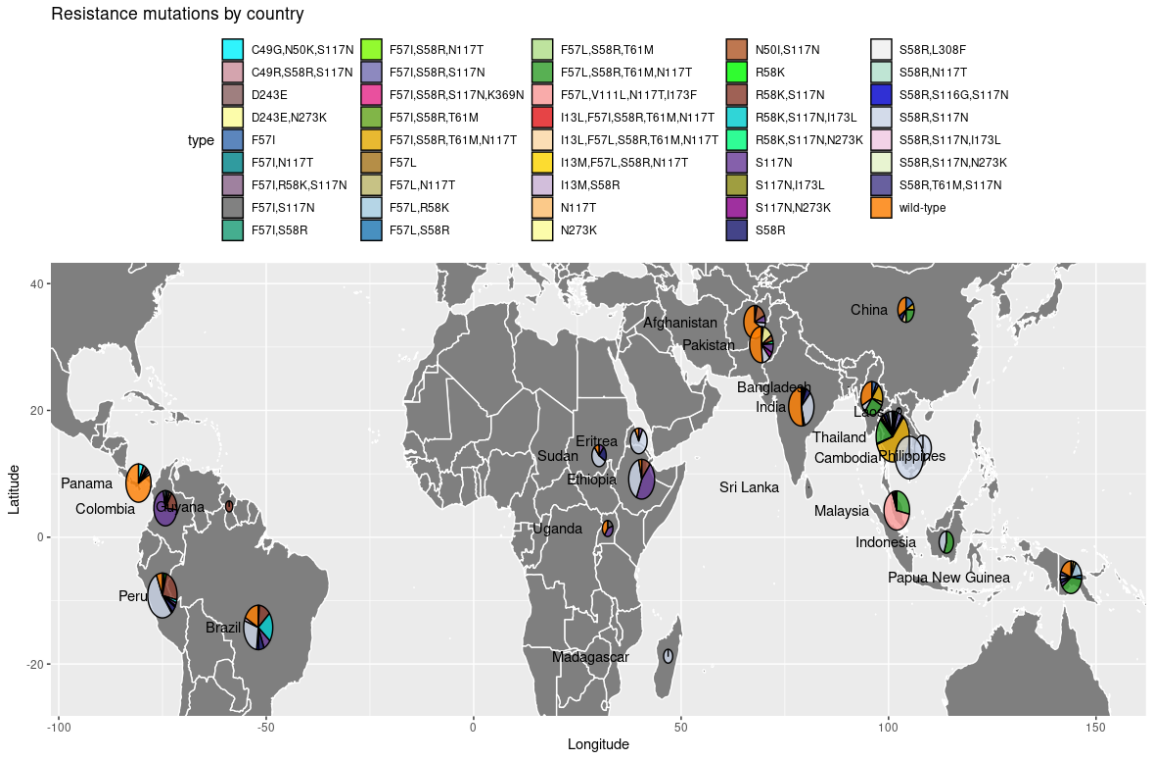
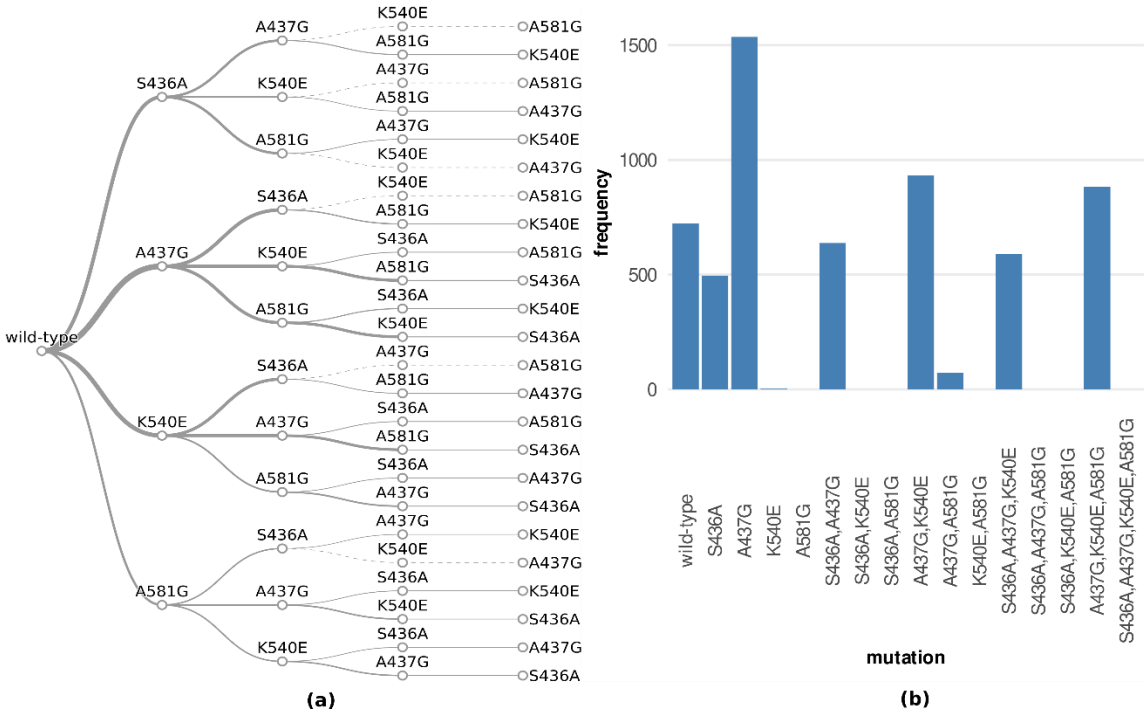


Figure 4. The geographical distribution of mutant *Pv*DHFR alleles found in our isolate data. The size of each pie chart is proportional to the number of isolates from that particular country.


wild-type
S436A
A437G
K540E
A581G
A437G
K540E
A581G
S436A
K540E
A581G
S436A
A437G
K540E
A581G
S436A
A437G
K540E
A581G
S436A
A437G

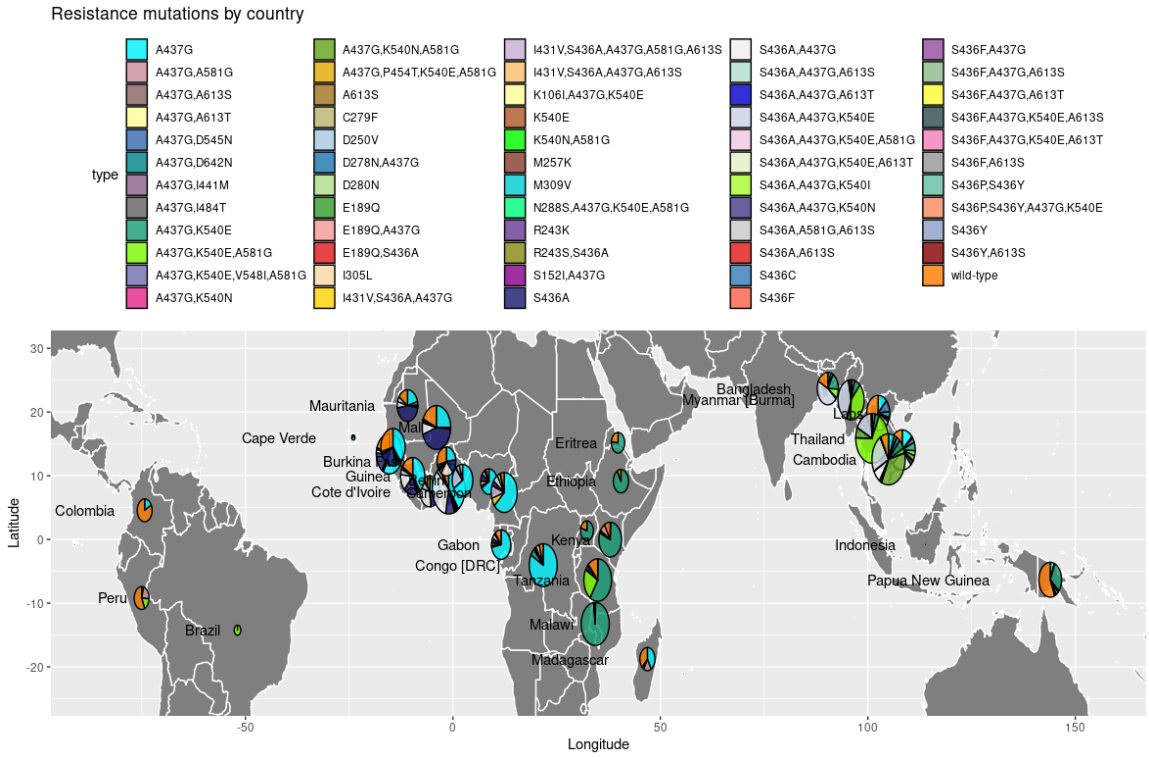
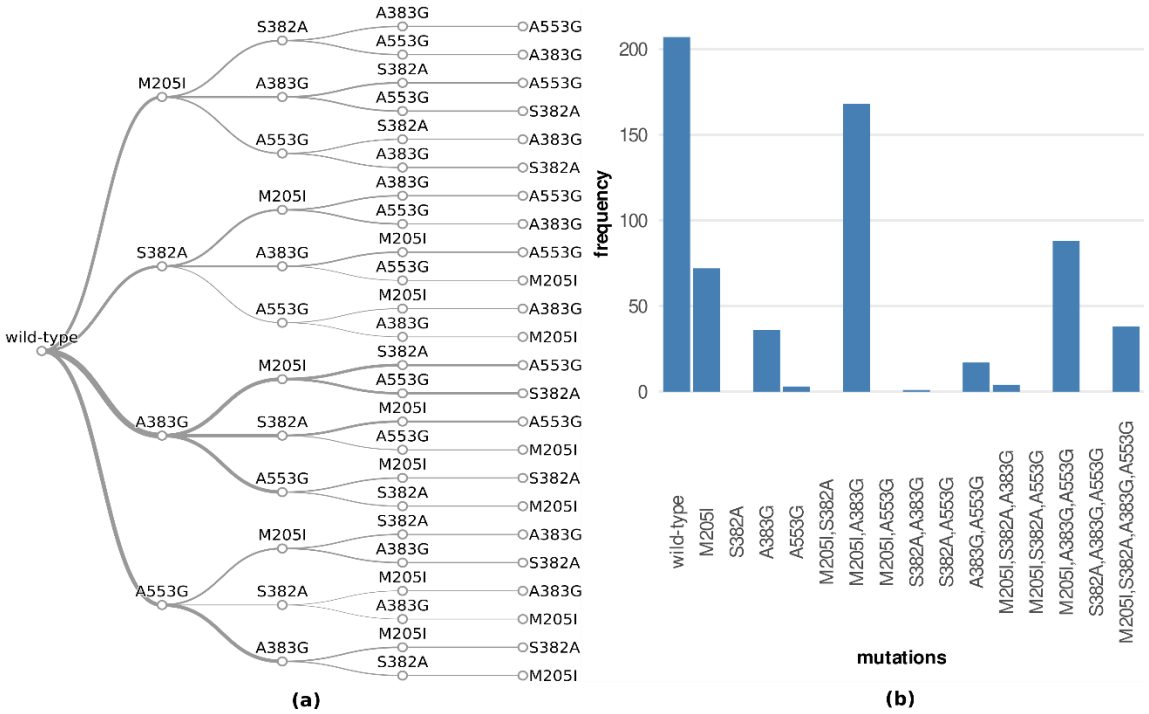


Figure 6. The geographical distribution of mutant *Pf*DHPS alleles found in our isolate data. The size of each pie chart is proportional to the number of isolates from that particular country.


wild-type
M205I
S382A
A383G
A553G
S382A
A383G
A553G
M205I
A383G
A553G
M205I
A553G
M205I
S382A
A553G
M205I
S382A
A553G
M205I
A553G
M205I
S382A
A553G
M205I
S382A
A553G
M205I
S382A

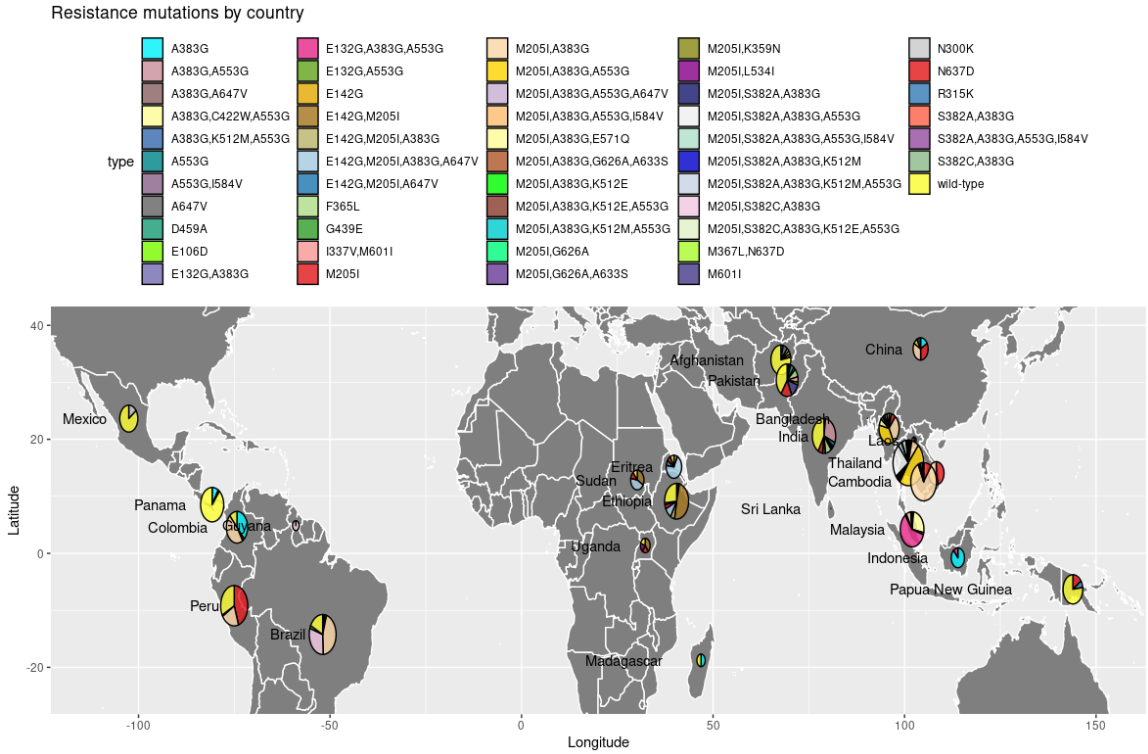


Figure 8. The geographical distribution of mutant *Pv*DHPS alleles found in our isolate data. The size of each pie chart is proportional to the number of isolates from that particular country.

Table 1. Correlation between Flex-ddG predictions and experimental data for *Pf*DHFR pyrimethamine resistance mutations

Mutation	$\Delta\Delta G_{exp}$*(kcal/mol)	Exp. Sum**	Exp. I.E.***	$\Delta\Delta G_{Flex\ ddG}$† (R.E.U)	Sum‡	I.E.§
N51I	-0.783			-0.065		
C59R	-0.184			-0.055		
S108N	1.297			0.275		
I164L	-0.351			-0.65		
N51I,S108N	1.89	0.514	1.376	-0.13	0.21	-0.34
C59R,S108N	2.29	1.113	1.177	0.575	0.22	0.355
N51I,C59R,S108N	2.595	0.33	2.265	0.225	0.155	0.07
C59R,S108N,I164L	3.283	0.762	2.521	0.025	-0.43	0.455
N51I,C59R,S108N,I164L	3.761	-0.021	3.782	0.425	-0.495	0.92
Pearson correlation				0.635	0.435	0.765
Correctly classified				8/9	4/5	4/5

*Experimentally measured *Pf*DHFR-pyrimethamine binding free energy change data from [9]
**Sum of experimental values of binding free energy change for independent mutations
***Interaction energy calculated as the difference between experimentally measured values of binding free energy change of multiple mutant compared to the sum of the independent mutations involved.
†Change in *Pf*DHFR-pyrimethamine binding free energy predicted by Flex ddG
‡Sum of Flex ddG predictions for independent mutations
§Interaction energy calculated as the difference between Flex ddG predicted binding free energy change of multiple mutant compared to the sum of the independent mutations.

Table 2. Correlation between Flex ddG predictions and experimental data for *Pf*DHPS sulfadoxine resistance mutations

Mutation	$\Delta\Delta G_{exp}$* (kcal/mol)	$\Delta\Delta G_{Flex\ ddG}$** (R.E.U)
A437G	1.36	0.025
A581G	0.811	0.11
S436A,A437G	2.939	0.275
A437G,A581G	2.817	0.125
A437G,K540E	3.856	0.125
Pearson correlation		0.469
Correctly classified		5/5

*Change in sulfadoxine binding free energy, calculated from affinity measurements (K_i) from [46] using the relation $\Delta G = -kT \ln K_i$.

**Change in *Pf*DHPS sulfadoxine binding free energy predicted by Flex-ddG

Table 3. Correlation between Flex ddG predictions and experimental data for *Pv*DHPS sulfadoxine resistance mutations

Mutation	$\Delta\Delta G_{exp}$* (kcal/mol)	$\Delta\Delta G_{Flex\ ddG}$** (R.E.U)
A383G	2.056	1.505
V585A	0.413	-0.05
A383G,A553G	2.826	1.875
S382A,A383G,A553G	3.072	1.525
Pearson correlation		0.94
Correctly classified		4/5

*Change in sulfadoxine binding free energy, calculated from affinity measurements (K_i) from [51] using the relation $\Delta G = -kT \ln K_i$.

**Change in *Pv*DHPS sulfadoxine binding free energy predicted by Flex-ddG