A computational method for predicting the most likely evolutionary trajectories in the step-wise accumulation of resistance mutations

R. Charlotte Eccleston[1*], Emilia Manko[1], Susana Campino[1], Taane G. Clark[1,2], Nicholas Furnham[1]

***For correspondence:** charlotte.eccleston@lshtm.ac.uk (RCE)

[1]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom; [2]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

Abstract Pathogen evolution of drug resistance often occurs in a stepwise manner via the accumulation of multiple mutations that in combination have a non-additive impact on fitness, a phenomenon known as epistasis. The evolution of resistance via the accumulation of point mutations in the DHFR genes of *Plasmodium falciparum* (*Pf*) and *Plasmodium vivax* (*Pv*) has been studied extensively and multiple studies have shown epistatic interactions between these mutations determine the accessible evolutionary trajectories to highly resistant multiple mutations. Here, we simulated these evolutionary trajectories using a model of molecular evolution, parameterized using Rosetta Flex ddG predictions, where selection acts to reduce the target-drug binding affinity. We observe strong agreement with pathways determined using experimentally measured IC50 values of pyrimethamine binding, which suggests binding affinity is strongly predictive of resistance and epistasis in binding affinity strongly influences the order of fixation of resistance mutations. We also infer pathways directly from the frequency of mutations found in isolate data, and observe remarkable agreement with the most likely pathways predicted by our mechanistic model, as well as those determined experimentally. This suggests mutation frequency data can be used to intuitively infer evolutionary pathways, provided sufficient sampling of the population.

Introduction

The development of new antimicrobial therapeutics and the design of 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 through a wide range of mechanisms (***Sun et al., 2019***; ***Davies and Davies, 2010***; ***Levy and Marshall, 2004***; ***Rodrigues et al., 2016***). One of the major routes to resistance, and focus of this work, 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 (**Blair et al., 2015**).

The evolution of resistance is affected by the interplay between selection for resistance, selection for protein function, drug concentration and mutational bias, and it is also influenced by a phenomenon known as epistasis (**Weinreich et al., 2006**; **Lozovsky et al., 2009**; **Jiang et al., 2013**).

Epistasis between mutations within the same protein arises due to energetic interactions between the amino acids, where the impact of a mutation depends upon the protein sequence (**Starr and Thornton, 2016**). 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. Epistasis determines the order of fixation of mutations and the accessibility of evolutionary trajectories to resistance phenotypes (**Weinreich et al., 2006**, **2005**) and has been observed in the evolution of many pathogens (**Khan et al., 2011**; **Gong et al., 2013**; **Sanjuán et al., 2005**), including the evolution of resistance in *Plasmodium falciparum* (**Lozovsky et al., 2009**; **Sirawaraporn et al., 1997**) and *Plasmodium vivax* (**Jiang et al., 2013**). It 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, which theoretically should result in reversion of resistance mutations, due to the fitness cost incurred in the absence of the drug (**Melnyk et al., 2015**; **Vogwill and MacLean, 2015**). However, the success of this strategy has been mixed, and in some cases bacterial populations remained resistant (**Costelloe et al., 2010**; **Enne, 2010**; **Sundqvist et al., 2010**), likely due to 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 (**Andersson and Hughes, 2011**).

Fragment-based drug discovery (FBDD) and AMR surveillance strategies require methods to predict evolutionary trajectories to resistance. For example, by identifying mutations involved in resistance trajectories that reduce the effectiveness of an antimicrobial drug, specific regions of a target molecule can be exploited or avoided, thus creating 'evolution proof' drugs. Therefore, understanding how epistasis arises and predicting which mutations will interact, is important for anticipating future mutations, designing new drugs and developing strategies to minimize resistance.

Evolution towards drug-resistant phenotypes in malaria species *P. falciparum* and *P. vivax* has been shown to occur in a stepwise manner, due to epistatic interactions between mutations, and the most likely trajectories to resistance phenotypes have been predicted using experimental measures of resistance (**Lozovsky et al., 2009**; **Jiang et al., 2013**; **Sirawaraporn et al., 1997**).

P. falciparum and *P. vivax* parasites cause the majority of malaria infections and have evolved strong resistance to many antimalarial drugs, including pyrimethamine (**Sirawaraporn et al., 1997**) and sulfadoxine (**Wang et al., 1997**). There were an estimated 241 million new cases of malaria world-wide in 2020, resulting in approximately 627,000 deaths predominately among children under 5 years of age (**WHO, 2021**). *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 (**Wang et al., 1997**; **Brooks et al., 1994**). 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 (**Snounou and White, 2004**). 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 (**Korsinczky et al., 2004**; **Hastings et al., 2004**).

The DHFR gene encodes an enzyme that uses NADPH to synthesize tetrahydrofolate, a co-factor in the synthesis of amino acids (**Kompis et al., 2005**) 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 (**Lozovsky et al., 2009**; **Sirawaraporn et al., 1997**) and *Pv*DHFR (**Jiang et al., 2013**).

Resistance in *Pf*DHFR has been studied extensively and 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 (***Ferlan et al., 2001***) by altering the binding pocket and reducing the affinity for the drug (***Yuthavong et al., 2005***). Epistasis 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 (***Lozovsky et al., 2009***; ***Sirawaraporn et al., 1997***). This means that mutations which on their own are not associated with a resistance phenotype, can be 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 (***Lozovsky et al., 2009***).

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 (***Jiang et al., 2013***), some combinations of which have been observed to result in pyrimethamine resistance both in vivo and in vitro (***Hastings et al., 2004***; ***Hawkins et al., 2007***). 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. The observed 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 (***Weinreich et al., 2006***; ***Tamer et al., 2019***), have highlighted the prevalence of epistasis among resistance mutations and the importance of considering epistatic interactions between mutations when predicting evolutionary trajectories to drug resistance.

The predictability of evolution is a central topic in biology of interest to experimentalists and theorists alike (***Achaz et al., 2014***; ***Lobkovsky and Koonin, 2012***; ***Szendro Ivan et al., 2013***) (for a review of the topic see ***de Visser and Krug*** (***2014***)). By using experimentally measured values to characterize the empirical fitness landscapes and simulate evolutionary trajectories, the work in ***Lozovsky et al.*** (***2009***) and ***Jiang et al.*** (***2013***) is determining the predictability of evolution in these landscapes by assessing which trajectories are accessible and the level of determinism associated with the evolution. Whilst such experimental methods have been successful in capturing epistasis, characterizing evolutionary landscapes and predicting evolutionary trajectories, they are expensive and time consuming.

The development of computational methods to predict resistance trajectories would enable fast and efficient predictions and would be more widely accessible than lab-based methods. Computational tools could help narrow down the pool of mutations to be studied experimentally and would also be applicable to difficult to study targets. Some target-specific computational tools to predict individual resistance mutations have been developed (***Karmakar et al., 2020***; ***Portelli et al., 2020***). However, such tools are target specific and so not generalizable. Furthermore, they only consider independent mutations on a single structure and so ignore epistasis between resistance mutations. Therefore, they are not suitable for predicting evolutionary trajectories to resistance.

To determine a generalizable computational method to predict evolutionary trajectories to resistance, we need to consider the main determinants of resistance. ***Rodrigues et al.*** (***2016***) investigated three mutations in *Escherichia coli* DHFR associated with trimethoprim resistance and considered activity, binding affinity, fold stability, and intracellular abundance. They found that whilst resistance is a trade-off between these factors, binding affinity is the single most predictive trait of resistance, especially at later points in evolution. Therefore, we decided to investigate if predictions of binding affinity change can be used to predict the order of fixation of resistance mutations involved in evolutionary trajectories to resistance.

Rosetta Flex ddG (***Barlow et al., 2018***) 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 (***Alford et al., 2017***). 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 (***Aldeghi et al., 2018***, ***2019***). However, its ability to capture epistasis has not yet been tested. Therefore, we investigated how well Flex ddG can capture epistasis between resistance mutations in *Pf* DHFR and observed a good agreement with experimental data.

Next, we used the Flex ddG predictions to parameterize a fitness function applied in an existing model of molecular evolution. We used this method to predict evolutionary trajectories to known resistant quadruple mutants in both *Pf* DHFR and *Pv*DHFR, where the evolutionary trajectories have been studied experimentally (***Lozovsky et al., 2009***; ***Jiang et al., 2013***). Good agreement was observed between the most likely trajectories to the quadruple mutations predicted by our model and those predicted experimentally. This suggests binding affinity is highly predictive of resistance, supporting the conclusions of ***Rodrigues et al.*** (***2016***).

The main advantage of this approach is that it does not require access to an experimental 'wet' lab and can be carried out by anyone with access to a high-performance computer. It is generalizable to any antimicrobial drug that acts by binding to its target and can be easily applied to any drug-target complex for which there is an available structure. Therefore, it can be used to study complexes and systems that might be problematic experimentally. It enables accurate assessment of the predictability of the evolutionary landscape and can predict whether we would expect to see constrained evolutionary trajectories on a fitness landscape as a result of epistatic interactions in drug binding free energy.

In addition, we analyzed if evolutionary pathways can be inferred from the frequency of mutations found in isolate data. We determined the frequency of mutations in *Pf* DHFR and *Pv*DHFR, and inferred the most likely evolutionary pathways under the assumption that the most likely mutation at each step corresponds to the most frequent mutation. We carried out this analysis first upon a combined set of global isolates and then upon isolates from individual regions. The most likely pathways inferred from the global isolate data agreed remarkably well with both the experimentally determined pathways and the pathways predicted by our computational method. This suggests evolutionary trajectories can be inferred from the frequency of mutations observed in isolate data, provided adequate sampling of the population. When considering geographical regions separately, the inferred pathways from several regions agreed well with the experimental pathways and our predicted pathways, however the most likely pathways inferred in some regions differed from the main pathways, highlighting the importance of considering the evolution in different regions separately.

Results

Rosetta Flex ddG captures general trends in binding free energy changes and epistasis

We investigated if Flex ddG predictions agree with experimentally measured binding free energy and if these predictions can be used to calculate the non-additivity in binding free energy (interaction energy), which for a double mutant defines the epistasis between the two mutations and, for a triple mutant or higher, captures the level of epistatic interactions. We calculated 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. 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.

The change in binding free energy was predicted using Flex ddG for the combinatorically complete set of the four *Pf* DHFR pyrimethamine resistance mutations N51I, C59R, S108N and I164L.

(Note on notation: lists of single mutations are written *X, Y, Z*, multiple mutations are written *X,Y,Z* (i.e. no space between the commas and the mutations) and pathways are written *X/Y/Z* to denote the order of fixation).

We compared the predictions to the data from ***Sirawaraporn et al.*** (***1997***) in which they determined binding free energy changes for a subset of the possible combinations of mutations, the sum of the independent mutations (calculated for multiple mutants to compare to the experimentally determined binding free energy changes of multiple mutants) and the interaction energy of the multiple mutants (Table 1). A positive $\Delta\Delta G$ value indicates a destabilising mutation and a negative $\Delta\Delta G$ 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 ***Sirawaraporn et al.*** (***1997***), a mutation that destabilized the binding corresponded to a negative $\Delta\Delta G$, whilst here we have reversed the signs of their data to enable comparison with our predictions).

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 (***Barlow et al., 2018***). We found the average of the distributions converges and the rank order of the mutations is constant at around 250 runs (Appendix 1-Figure 3 and Appendix 1-Figure 4). We compared the predictions for 250 runs and the data from ***Sirawaraporn et al.*** (***1997***)(1) and observed a correlation of 0.611 for the binding free energy data, 0.660 for the sum of the independent predictions for multiple mutants and 0.756 for the interaction energy. We found 8/9 binding free energy predictions were correctly classified, 4/ 5 of the sum of the independent predictions were correctly classified and 4/ 5 of the interaction energies were correctly classified. Comparing the predictions for 35 runs (Appendix 1-Figure 1, Appendix 1- Table 1) and 250 (Appendix 1-Figure 2, Table 1) runs, 250 runs provides the best trade-off between accuracy and efficiency (see Supplementary text for detailed discussion). Therefore, we will be discussing the predictions for n=250 going forward.

Mutation S108N was the only single mutation to destabilize pyrimethamine binding in both the experimental data and the Flex ddG predictions. However, in the experimental data the double mutation N51I,S108N is more destabilizing to binding than single mutation S108N, but the Flex ddG prediction was stabilizing. 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 destabilizing impact out of all combinations of single and multiple mutations, however, Flex ddG predicted it to be less destabilizing than the double mutation C59R,S108N and single mutation S108N.

Considering the interaction energy, the incorrectly classified mutation was again N51I,S108N which was predicted to be positive, but found experimentally to be negative, because the sum of the individual predictions was destabilizing but the double mutation itself was predicted to be stabilizing. Both the experimental data and our predictions found that the quadruple mutation had the largest magnitude interaction energy reflecting the greatest difference between the stabilizing impact of the sum of the individual mutations and the destabilizing impact of the quadruple mutation itself.

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 ***Lozovsky et al.*** (***2009***) However, whilst the interaction energy of the triple mutation N51I,C59R,S108N was positive for both the experimental 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 stabilizing – almost neutral - to pyrimethamine binding, whilst in the experimental data both mutations have a large stabilizing impact. Furthermore, the triple mutation was predicted to be only marginally more destabilizing than single mutation S108N, resulting in the small

Table 1. Correlation between Flex ddG predictions for 250 runs and experimental data (see table 4 of ***Sirawaraporn et al.*** (***1997***)) for *Pf* DHFR pyrimethamine resistance mutations

Mutation	$\Delta\Delta G^{*}_{exp}$ (kcal/mol)	Exp. Sum**	Exp I.E.***	$\Delta\Delta G^{\dagger}_{FlexddG}$ (kcal/mol)	Sum‡	I.E.§
N51I	-0.783			-0.124		
C59R	-0.184			-0.033		
S108N	1.297			0.312		
I164L	-0.351			-0.323		
N51I,S108N	1.89	0.514	1.376	-0.166	0.188	-0.354
C59R,S108N	2.29	1.113	1.177	0.399	0.279	0.119
N51I,C59R,S108N	2.595	0.33	2.265	0.162	0.155	0.007
C59R,S108N,I164L	3.283	0.762	2.521	0.018	-0.043	0.061
N51I,C59R,S108N,I164L	3.761	-0.021	3.782	0.301	-0.168	0.469
Pearson Correlation				0.611	0.660	0.756
Correctly Classified				8/9	4/5	4/5

*Experimentally measured *Pf* DHFR pyrimethamine binding free energy change data from ***Sirawaraporn et al.*** (***1997***)

**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 calculated as the average of the distribution of runs. Free energy predictions from Rosetta are in Rosetta Energy Units, however the authors of Flex ddG applied a generalized additive model to re-weight the predictions and make the output more comparable to units of kcal/mol (***Barlow et al., 2018***)

‡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.

negative interaction energy.

We conclude that although there are some disagreements between the predictions and the data, Flex ddG is able to capture the general trend of the data. However, if we use the average of the distributions as a summary metric of the predictions for the combinatorically complete set of the four mutations and try to infer a pathway through to the quadruple mutation, under the criteria that each subsequent mutation must destabilize pyrimethamine binding more than the last, then we are unable to find a pathway through. However, since the predictions capture the general trend observed in the data, and the summary metric does not fully characterize the entire distribution of predictions, we used the distributions to parameterize an evolutionary model to determine if we can predict a pathway through to the quadruple mutation and if the predicted evolutionary trajectories agree with experimentally determined evolutionary trajectories.

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

We simulated the evolutionary trajectories to the quadruple mutants described above for the genes *Pf* DHFR and *Pv* DHFR using an evolutionary model, adapted from previous studies (***Eccleston et al., 2021***; ***Pollock et al., 2012***, ***2017***). In this model, selection acts to reduce the binding affinity between target protein and the antimalarial drug with which the mutations have been associated with resistance. 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. If the set of sampled mutations at a step all have a fixation probability of zero, the algorithm terminates at that point in the pathway and begins the next run at the single mutation step. Therefore, it is not guaranteed that a run will reach the quadruple mutation. We carried out 50,000 runs and determined i) the number of runs that reached a single, double, triple or the quadruple mutation before the run ended (files ending `'_endpoint_numbers.csv'`), ii) the frequency of the observed trajectories up to the quadruple mutation, including trajectories that terminated before the quadruple mutation (files ending `'_pathway_endpoints.csv'`), iii) the frequency at which a mutational step was chosen in all runs (files ending `'_total_pathway_probabilities.csv'`) and iv) the most likely trajectories to the quadruple mutation predicted by our simulations (files ending `'_quadruple_pathways.csv'`).

To determine how well our simulations reflect the evolutionary process to the *Pf* DHFR quadruple mutation N51I,C59R,S108N,I164L, we compared our results to experimentally determined evolutionary trajectories as presented by ***Lozovsky et al.*** (***2009***). In our simulations, the quadruple mutation was reached in approximately 8% of runs (see Supplementary file `'PfDHFR_endpoint_numbers.csv'`). The majority of the runs (66%) terminated at a double mutation, with S108N/C59R the most likely trajectory over all. The algorithm was often unable to proceed passed S108N/C59R because the Flex ddG distribution for C59R,S108N is concentrated around large destabilizing values (Appendix 1 -Figure 2f) whilst the distributions of the two possible next steps, N51I,C59R,S108N and C59R,S108N,I164L, are concentrated around lower destabilizing values (Appendix 1- Figures 2g and h). Therefore, in many instances, the change in binding free energy caused by the next step in the pathway were predicted to be stabilizing, and thus were not be chosen by the algorithm. This demonstrates the dependence of the method upon the accuracy of the Flex ddG. In contrast, the majority of runs in the simulations based on IC50 measurements presented in ***Lozovsky et al.*** (***2009***) reached the quadruple mutation (see Figure 2 in ***Lozovsky et al.*** (***2009***).

However, since we are interested in how epistasis influences the order of fixation of mutations in an evolutionary trajectory to a high-resistance quadruple mutation, we compared the most likely trajectories to the quadruple mutation predicted by our simulations to the most likely trajectories to the quadruple mutation predicted in ***Lozovsky et al.*** (***2009***) and observed remarkable agreement. The top two most likely trajectories predicted by our model to the *Pf* DHFR quadruple mutation were S108N/C59R/N51I/I164L and S108N/C59R/I164L/N51I, respectively (Figure **??**) which correspond to the top two most likely pathways to the quadruple mutation determined in ***Lozovsky et al.*** (***2009***). The third most likely trajectory to the quadruple mutation in ***Lozovsky et al.*** (***2009***) was predicted to be S108N/N51I/C59R/I164L, however this pathway was predicted to be unlikely in our simulations, due to the fact that the distribution of Flex ddG predictions for double mutation N51I,S108N was mostly stabilizing to pyrimethamine binding (Appendix 1- Figure 2e), whereas all of the S108N distribution was destabilizing to pyrimethamine binding, so this step was unlikely to be chosen by the evolutionary algorithm.

Considering the frequency at which the single mutations were chosen as the first step in all simulated pathways (`'PfDHFR_total_pathway_probabilities.csv'`), S108N was the most likely single mutation and C59R was the second most likely single mutation, in agreement with the two most likely first steps in the pathways predicted in ***Lozovsky et al.*** (***2009***). The most likely pathway to a double mutation realized in all trajectories in both our simulations and the simulations in ***Lozovsky et al.*** (***2009***) is S108N/C59R. Similarly, the most likely pathway to a triple mutation realized in all our simulations and in ***Lozovsky et al.*** (***2009***) was S108N/C59R/N51I.

To simulate the evolutionary pathways for *Pv*DHFR, we also carried out predictions of binding free energy changes for the homologous set of four mutations in *Pv*DHFR, (N50I, S58R, S117N and I173L). Unfortunately, binding affinity data is not available for the mutations in *Pv*DHFR to compare to the Flex ddG predictions. However, ***Jiang et al.*** (***2013***) predicted pathways to the *Pv*DHFR quadruple mutation for four pyrimethamine concentrations using simulations informed by both

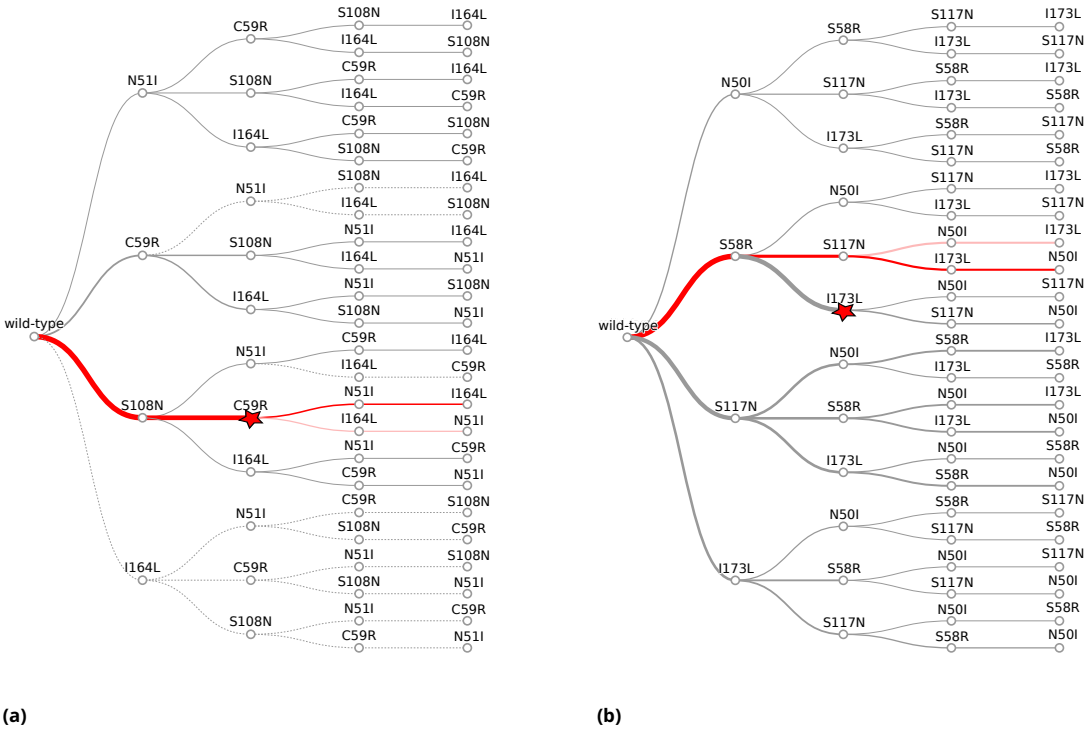


Figure 1. The probability of simulated evolutionary pathways to quadruple mutations **(a)** N51I,C59R,S108N,I164L in *Pf*DHFR and **(b)** N50I,S58R,S117N,I173L in *Pv*DHFR. Line thickness indicates the total probability of a mutation when considering all pathways it can occur in, determined from the frequency of that step in all realized mutational pathways from all runs. Dotted lines indicate zero probability of a mutation at that step. The most likely pathway in total is denoted by a red star. The most likely pathway to the quadruple mutation is highlighted in dark red and the second most likely pathway to the quadruple mutation is highlighted in lighter red. The probabilities corresponding to these plots can be found in Supplementary files 'PfDHFR_total_pathway_probabilities.csv' and 'PvDHFR_total_pathway_probabilities.csv' for a) and b), respectively.

drug resistance and catalytic activity, to which we can compare our simulations. In their simulations, the quadruple mutation fixed in 99.8% of runs for the highest pyrimethamine concentration, but it did not fix for the three lower concentrations. In our simulations, the quadruple mutation was reached in 39% of runs (`'PvDHFR_endpoint_numbers.csv'`), whilst 51% of runs terminated at a double mutation. The most likely endpoint overall in our simulations was S58R/I173L, which occurred in 32% of runs, however this path was not a frequent trajectory observed in the simulations in ***Jiang et al.*** (***2013***). All Flex ddG runs of double mutation S58R,I173L were predicted to be destabilizing and many predicted to have a medium to large impact. However, the triple mutation S58R,S117N,I173L was predicted to have a smaller destabilizing impact than S58R,S117N, making pathway S58R/I173L/N50I unlikely, whilst N50I,S58R,I173L was predicted to be stabilizing to pyrimethamine in all Flex ddG runs and therefore pathway S58R/I173L/N50I had a zero probability of occurring in the simulations. This resulted in many runs terminating at step S58R/I173L. Considering the order of fixation up to the quadruple mutation, we compared the most likely evolutionary trajectories to the quadruple mutation predicted by our simulations to the most likely evolutionary trajectories to the quadruple mutation presented in ***Jiang et al.*** (***2013***), and observed good agreement for the largest of the four pyrimethamine concentrations they considered. The most likely pathway to the quadruple mutation predicted by our simulations was S58R/S117N/I173L/N50I (Figure **??**) which corresponds to the second most likely pathway to the quadruple mutation predicted in ***Jiang et al.*** (***2013***) for the highest pyrimethamine concentration. Our second most likely pathway to the quadruple mutation (S58R/S117N/N50I/I173L) corresponds to the first most likely path-

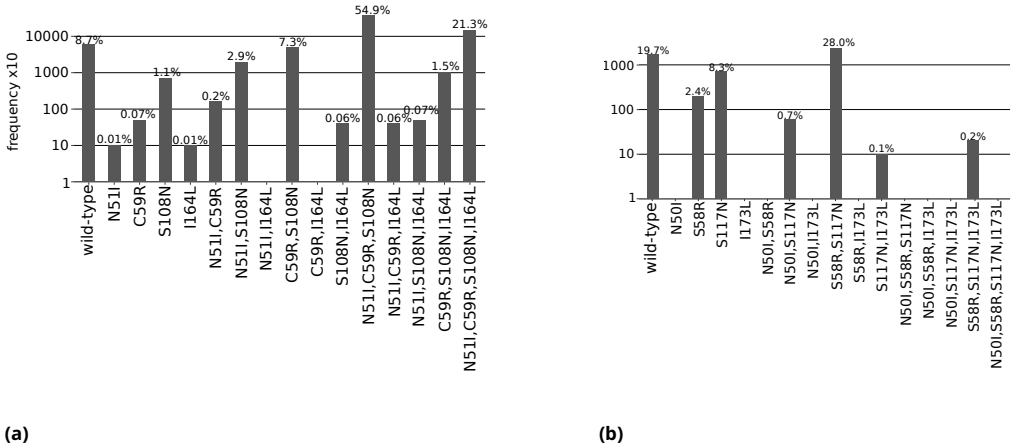
way predicted in ***Jiang et al.*** (***2013***) for the highest pyrimethamine concentration. There were two other possible pathways to the quadruple at the highest concentration, S117N/N50I/S58R/I173L and N50I/S117N/S58R/I173L, which correspond to the fourth and twelfth most likely pathways to the quadruple in our simulations.

The first step in the evolutionary trajectories determined in ***Jiang et al.*** (***2013***) for the highest concentration was S58R whereas for the three lower concentrations it was S117N. The most likely first step in all pathways predicted by our simulations was S58R (Figure **??**), whilst S117N was the second most likely first step (`'PvDHFR_total_pathway_probabilities.csv'`). Analyzing the Flex ddG distributions, the S58R predictions are mostly destabilizing to pyrimethamine and it is the only single mutation to reduce the binding affinity when considering both the average and the peak of the distribution (Figure S1). The S117N distribution peaks around zero, meaning the majority of the runs predict this mutation has a neutral impact on pyrimethamine binding, whilst there is a smaller peak in the distribution for mildly destabilizing values (Figure S2).

To quantify the predictability of an evolutionary landscape, previous studies have calculated the Gibbs-Shannon entropy distribution of the path weights (***Szendro Ivan et al., 2013***; ***de Visser and Krug, 2014***), namely $S = -\sum P_i \ln P_i$, where P_i is the probability of the i^{th} pathway and the value of S ranges from 0 to $\ln n$ for n equally likely pathways. The lower the value of S the higher the predictability of the evolution i.e. most of the probability is concentrated around a small number of pathways, suggesting epistasis is influential in constraining the accessible trajectories. The value of S when considering the probability distribution of all realized evolutionary trajectories in the simulations was 1.19 for *Pf*DHFR and 2.82 for *Pv*DHFR (both simulations have an equal number of possible pathways because they have an equal number of mutations, so the values of S are comparable and the maximum value of S for both simulations is 4.16). This means the evolutionary trajectories were more constrained in the *Pf*DHFR simulations than in the *Pv*DHFR simulations and suggests that epistasis between the mutations plays a greater role in constraining the trajectories in the evolution of *Pf*DHFR resistance. Unfortunately, the probabilities of all possible pathways determined in ***Lozovsky et al.*** (***2009***) and ***Jiang et al.*** (***2013***) are not made available (the data is represented in pathway diagrams, the probabilities of a step are indicated by line thickness and only the probabilities of the most likely pathways annotated), therefore we cannot calculate the corresponding values of S for these distributions for comparison.

The frequency of mutations in isolate data can be used to infer evolutionary trajectories to multiple resistance mutations

It was noted in ***Lozovsky et al.*** (***2009***) that their most likely pathways to the *Pf*DHFR quadruple mutation were consistent with combinations of these four mutations observed in high frequencies in worldwide surveys of *P. falciparum* polymorphisms. To expand on this idea, we analyzed the frequency of the combinations of mutations in *Pf*DHFR and *Pv*DHFR found in our isolate data to identify if there is agreement between these frequencies, the experimentally determined trajectories and our predicted trajectories and if, therefore, isolate frequency data may be used to infer evolutionary trajectories. We inferred evolutionary trajectories from the frequency data by assuming if a specific mutation was found in high frequency (and is part of the combinatorically complete set of four mutations found in the four genes) then it is likely to be part of the evolutionary trajectory towards the quadruple mutation. To infer the first step in the most likely trajectory, we considered the frequency of single mutations of the set of four mutations considered for each gene and selected the most frequent mutation. To infer subsequent steps in the trajectory, we considered the frequency of only those mutations that contain the previous mutation and another of the set of four mutations in some combination and chose the most frequent mutation at each step. We also inferred alternative pathways which from the frequency data are less likely than the main pathway, but still a possibility due to the occurrence of intermediate mutations in the isolate data. To do this, we considered each step in the most likely trajectory and identified any other high frequency mutations that would enable alternative pathways from the double mutation onwards.



(a) **(b)**

Figure 2. The total frequency of the combinations of mutations found in our isolate data for sets of four mutations **(a)** N51I, C59R, S108N and I164L in *Pf*DHFR, and **(b)** N50I, S58R, S117N and I173L in *Pv*DHFR. All frequencies have been multiplied by a factor of 10 to enable clear identification of those mutations occurring in one isolate only. The frequencies are also given as the percentage of the total number of isolates, which for *Pf*DHFR is 6762 and *Pv*DHFR is 847.

If there were no alternative pathways, we began the process again but chose the second most frequent single mutation (if applicable) and built the pathway from there. In the event of multiple alternative pathways, we are unable to quantify their relative likelihoods, only that they are less likely than the most likely pathway. It is sometimes not clear which pathway is most likely. For example, for the set of mutation frequencies *A*:9, *D*:10, *AB*:20, *CD*:2, *ABC*:50, *BCD*:1, *ABCD*:75, the most likely pathway from the method stated above would be *D/C/B/A* and the alternative pathway would be *A/B/C/D*, purely because mutation *D* is more abundant than mutation *A*. However, the frequencies of the intermediate mutations in the most likely pathway are low compared to the alternative pathway. Therefore, in these situations we will not refer to any one pathway as the most likely pathway and will refer to all pathways as possible trajectories.

Considering the total frequency of each mutation in the set of four *Pf*DHFR mutations (N51I, C59R, S108N and I164L) in the isolate data (Figure **??**), S108N was the most frequent single mutation (72/6762 isolates), C59R,S108N the most frequent double mutation (496/6762 isolates) and N51I,C59R,S108N the most frequent triple mutation (3714/6762 isolates). The quadruple mutation N51I,C59R,S108N,I164L was found in 1439/6762 isolates. This suggests the pathway proceeds in the order S108N/C59R/N51I/I164L, in agreement with the most likely pathway to the quadruple mutation from both our evolutionary simulations and those using experimental data (***Lozovsky et al., 2009***).

Triple mutation C59R,S108N,I164L was found in 101/6762 isolates, suggesting that the second most likely pathway to the quadruple from our simulations and experimental data, S108N/C59R/I164L/N51I, is a possible alternative trajectory to the quadruple mutation. Double mutation N51I,S108N was the second most frequent double mutation in the isolate data (198/6762 isolates), allowing for another alternative pathway S108N/N51I/C59R/I164L. This agrees with the third most likely pathway presented by ***Lozovsky et al.*** (***2009***), however this pathway was unlikely in our simulations.

Single mutations C59R and N51I were the second and third most prevalent single mutations in our isolate data, found in 5/6762 and 1/6762 of isolates, respectively. They were also the second and third most likely first step in our pathway predictions ('PfDHFR_total_pathway_probabilities.csv'). Single mutation I164L was absent from the isolate data and had zero probability of being selected as the first step of our evolutionary trajectories.

A Chi-squared analysis revealed the worldwide distribution of mutations is significantly dif-

ferent than would be expected if there was no preferred evolutionary pathway, and the mutations which were over-represented were those involved in the most likely pathway inferred above S108N/C59R/N51I/I164L (see Appendix 2 and Appendix 2 - Figure 1). This provides further support that the epistatic interactions between the mutations determine the order of fixation resulting in preferred pathways to the quadruple mutation.

Considering the set of four *Pv*DHFR mutations (N50I, S58R, S117N and I173L), in our isolate data, the mutations S58R and S117N are fixed at these locations and the wild-type alleles are now considered to have an Arginine at codon 58 and Asparagine at codon 117. However in ***Jiang et al.*** (***2013***) they consider the wild-type allele to have a Serine at codons 58 and 117 and therefore we have changed our definition of the wild-type allele to agree with ***Jiang et al.*** (***2013***) for ease of comparison with their evolutionary pathways and our own.

The most frequent single mutation was S117N (70/847 isolates), the most frequent double mutation was S58R,S117N (237/847) and the only observed triple mutation was S58R,S117N,I173L (2/847) (Figure **??**). The quadruple mutation was not observed in our isolate data, and has not been reported in the literature either. By considering the frequency of the possible combinations of mutations, we inferred the evolution towards triple mutation S58R,S117N,I173L most likely occurs via pathway S117N/S58R/I173L. This corresponds to the fifth most likely pathway to a triple mutation when considering all pathways observed in our simulations, however this pathway is not observed in any of the most frequent pathways at any of the four concentrations studied in ***Jiang et al.*** (***2013***).

Single mutation S58R was the second most frequent single mutation in our isolate data (20/847). This supports the predicted first evolutionary steps in ***Jiang et al.*** (***2013***) which were predicted to be S58R for the highest pyrimethamine concentration and S117N for three lower concentrations of pyrimethamine, which suggests both single mutations are possible, but S117N is more likely for a lower pyrimethamine concentrations. An alternative pathway to the triple mutation could therefore be S58R/S117N/I173L, which may be more likely under higher pyrimethamine concentrations. This corresponds to the most likely pathway to a triple mutation in our simulations and is part of the second most likely pathway to the quadruple mutation at the highest pyrimethamine concentration considered in ***Jiang et al.*** (***2013***).

A Chi-squared test on the frequency distributions of the single and double *Pv*DHFR mutations (the triple mutations were too infrequent to include in the analysis, see Appendix 2 and Appendix 2 -Figure 1 for more details) revealed the worldwide distribution of *Pv*DHFR mutations is significantly different than would be expected if there were no preferred order of fixation of this set of mutations, with S117N and S58R,S117N being over-represented in the single and double distributions, respectively. This supports our inference that pathway S117N/S58R/I173L is the most likely pathway in the worldwide data.

Analysis of geographical distribution of mutations found in our isolate data reveals alternative pathways to resistance

We next considered the evolutionary trajectories by geographical location to determine if there are any differences in the inferred trajectories compared to the global trajectories, and which areas agree with the trajectories predicted by our simulations. The *P. falciparum* isolates were grouped in to seven geographical regions, as defined by the United Nations Statistics Division: 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), Middle Africa (Congo [DRC]), Eastern Africa (Eritrea, Ethiopia, Kenya, Madagascar, Malawi, Tanzania, Uganda), Southern Asia (Bangladesh), Southeastern Asia (Cambodia, Indonesia, Laos, Myanmar, Thailand and Vietnam) and Melanesia (Papua New Guinea). The *P. vivax* isolates were grouped into seven broad geographical regions, as defined by the United Nations Statistics Division: Central America (Mexico), South America (Brazil, Colombia, Guyana, Panama, Peru), Eastern Africa (Ethiopia, Eritrea, Madagascar, Sudan, Uganda), Southern Asia (Afghanistan, Bangladesh, India, Pakistan, Sri

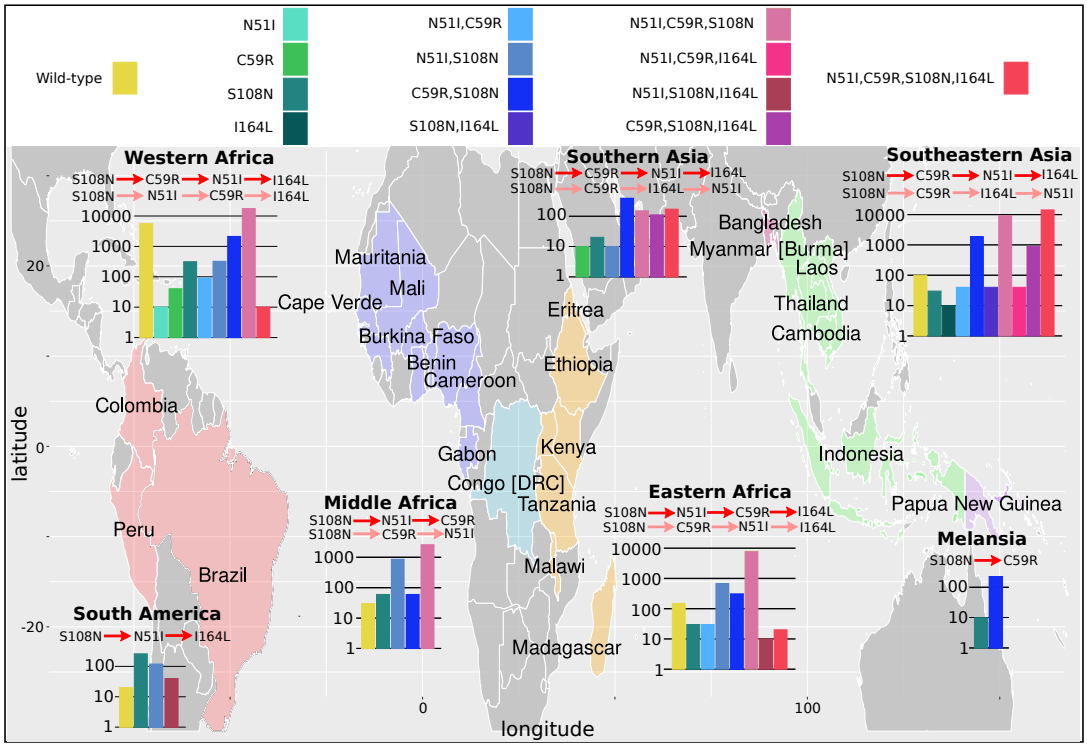


Figure 3. The *Pf*DHFR isolate data was grouped into seven geographical areas: South America, West Africa, Middle Africa, Eastern Africa, Southern Asia, Southeastern Asia and Melanesia. The bar charts display the frequency (log scale) of the combinations of the four mutations N51I, C59R, S108N and I164L. The frequency data has been multiplied by a factor of 10 to enable clear identification of those mutations occurring in one isolate only. The most likely evolutionary trajectory inferred from the frequency of combinations are included above the corresponding frequency chart from which the pathways were inferred indicated by mutations separated by dark red arrows. Alternative pathways are indicated by mutations separated by light red arrows. Where only single mutations are present a pathway is not inferred. (See Supplementary data folder 'PfDHFR/IsolateMutationFrequency' for the frequency of all mutations found in the isolate data from these regions).

Lanka), Southeastern Asia (Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Thailand and Vietnam), Eastern Asia (China) and Melanesia (Papua New Guinea). The mutation frequency data discussed in this section for each country can be found in Supplementary data folders 'PfDHFR/IsolateMutationFrequency' and 'PvDHFR/IsolateMutationFrequency'. For an analysis of the frequencies and inferred pathways per country per region, as well as information on additional mutations found in the data, see Supplementary text.

As in the previous section, we inferred the most likely pathway by assuming the most frequent mutation at each step corresponds to the most likely evolutionary trajectory. The inferred most likely pathway to the quadruple mutation agreed with the main pathway, S108N/C59R/N51I/I164L, predicted by our evolutionary model and the data presented in ***Lozovsky et al.*** (***2009***), as well as the most likely pathway inferred by considering the frequency of the worldwide *Pf*DHFR isolate data in Western Africa (S108N: 31/2594; C59R,S108N: 211/2594; N51I,C59R,S108N: 1739/2594; N51I,C59R,S108N,I164L: 1/2594), Southern Asia (S108N: 2/86; C59R,S108N: 39/86; N51I,C59R,S108N: 15/86; N51I,C59R,S108N,I164L: 17/86) and Southeastern Asia (S108N: 3/2650; C59R,S108N: 186/2650; N51I,C59R,S108N: 920/2650; N51I,C59R,S108N,I164L: 1419/2650). Additionally, the alternative worldwide pathway S108N/C59R/I164L/N51I was inferred to be an alternative in Southern Asia (C59R,S108N,I164L: 11/86) and Southeastern Asia (C59R,S108N,I164L: 90/2650), corresponding to the second most likely pathway to the quadruple predicted by our simulations, the data in ***Lozovsky et al.*** (***2009***) and worldwide frequency data.

This main pathway to the quadruple mutation was also inferred to be a possible alternative pathway in Eastern Africa (S108N: 3/904; C59R,S108N: 31/904; N51I,C59R,S108N: 782/904, N51I,C59R,S108N,I164L: 2/904). However, in Eastern Africa S108N/N51I/C59R/I164L was the most likely pathway to the quadruple mutation (S108N,N51I: 67/904), which corresponds to the third most likely pathway presented in ***Lozovsky et al.*** (***2009***) and inferred from the total frequency data, but which was unlikely in our simulations. This was also an alternative pathway in Western Africa (S108N,N51I: 32/2594).

Furthermore, the quadruple mutation was not observed in the isolate data from Middle Africa, South America, and Melanesia. In Middle Africa, evolution up to the triple mutation N51I,C59R,S108N was observed and the most likely inferred pathway to this mutation was S108N/N51I/C59R (S108N: 6/359; N51I,S108N: 86/359; N51I,C59R,S108N: 258/359), with an alternative less likely pathway of S108N/C59R/N51I (C59R,S108N: 6/359), corresponding to the thirteenth and first most likely trajectories to a triple mutation in our simulations, respectively. In South America, evolution up to the triple mutation N51I,S108N,I164L was observed in the isolate data and was inferred to follow the pathway S108N/N51I/I164L (S108N: 26/50, N51I,S108N: 12/50; N51I,S108N,I164L: 4/50), however this pathway does not occur in our simulations. In Melanesia, evolution up to the double mutation C59R,S108N was observed in the isolate data and was inferred using the frequency data to have followed the pathway S108N/C59R (S108N: 1/119; C59R,S108N: 23/119), which corresponds to the most likely pathway of all our evolutionary runs.

We performed an analysis of the significance of the regional frequency distributions (see Appendix 3 and Appendix 3 - Figure 2). All regions had mutations which were significantly over- or underrepresented compared to what would be expected from the worldwide distribution. The overrepresented mutations were always part of the inferred most likely evolutionary pathway for each region. This suggests any differences in the inferred pathways between the regions and the worldwide data are significant. For example, in South America (Appendix 3 Figure 2d), the three mutations involved in the most likely inferred pathway (single mutation S108N, double mutation N51I,S108N and triple mutation N51I,S108N,I164L) are all overrepresented. These mutations are involved in the most likely inferred pathway in that region (S108N/N51I/I164L), suggesting this region is indeed following a different evolutionary trajectory to what we would expect from the worldwide data (see Supplementary Text for detailed analysis of all regions).

Next, we considered the frequency of combinations of the four *Pv*DHFR mutations in different geographical regions and used these frequencies to infer evolutionary trajectories (Figure 4). 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 where possible, and compare to the most likely pathways to triple mutations in our simulations and in the experimental data. Triple mutation S58R,S117N,I173L was the only triple mutant combination observed in the isolate data and was only found in South America. Analyzing the frequency of the constituent mutations (S117N: 34/257; S58R,S117N: 74/257; S58R,S117N,I173L: 2/257), the most likely inferred pathway to this mutation is S117N/S58R/I173L. This was the fifth most likely triple mutation pathway in our simulations, however it was not an observed pathway in ***Jiang et al.*** (***2013***). There are two other possible pathways inferred from the frequency data to this triple mutation: S117N/I173L/S58R (S117N,I173L: 1/257), and S58R/S117N/I173L (S58R: 9/257). Pathway S117N/I173L/S58R was the third most likely pathway to a triple mutation in our simulations but was not observed in the simulations in ***Jiang et al.*** (***2013***) and S58R/S117N/I173L the most likely pathway to a triple mutation in our simulations and was observed as part of the second most likely pathway to the quadruple mutation at the highest pyrimethamine concentration considered in ***Jiang et al.*** (***2013***).

Evolution only up to double mutation S58R,S117N was found in Southeastern Asia, Eastern Africa and Melanesia and S117N/S58R was the most likely fixation order in Eastern Africa, whilst S58R/S117N was the most likely fixation order in Southeastern Asia and Melanesia. Pathways S58R/S117N and S117N/S58R were the second and fourth most likely double mutation pathways in our simulations. In Southern Asia, evolution up to double mutations S58R,S117N and N50I,S117N

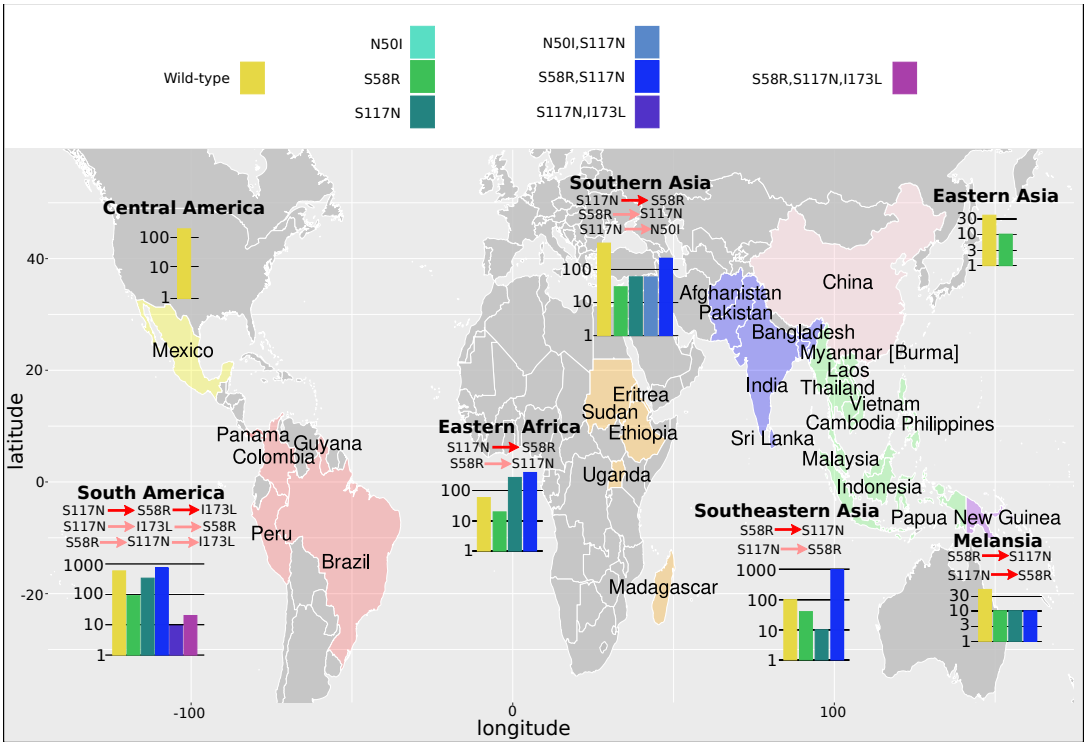


Figure 4. The *Pv*DHFR isolate data was grouped into seven geographical areas: Central America, South America, Eastern Africa, Southern Asia, Eastern Asia, Southeastern Asia and Melanesia. The bar charts display the frequency (log scale) of the combinations of the four mutations N50I, S58R, S117N and I173L. The frequency data has been multiplied by a factor of 10 to enable clear identification of those mutations occurring in one isolate only. The most likely evolutionary trajectory inferred from the frequency of combinations are included above the corresponding frequency chart from which the pathways were inferred indicated by mutations separated by dark red arrows. Alternative pathways are indicated by mutations separated by light red arrows. Where only single mutations are present a pathway is not inferred. (See Supplementary Text for further explanation and see Supplementary data folder 'PvDHFR/IsolateMutationFrequency' for the frequency of all mutations found in the isolate data from these regions).

was observed, following pathways S117N/S58R and S117N/N50I, respectively (S58R: 3/37, S117N: 6/37, N50I,S117N: 6/37, S58R,S117N:22/37), with the pathway S117N/S58R appearing to be more prevalent. In Eastern Asia, evolution only up to single mutation S58R was observed and in Central America no steps in the evolutionary pathway including combinations of these four mutations were found.

We performed an analysis of the significance of the regional distributions, similar to described above for *Pf*DHFR, however the frequencies of the four *Pv*DHFR in many of the regions was too small to definitively draw conclusions (see Appendix 3 and Appendix 3 - Figure 2 for more details). However, from this analysis it does appear that the distribution of mutations in South America is very similar to the worldwide distribution (Appendix 3 -Figure 2d) and this region is following the same inferred evolutionary pathway as the pathway inferred from the worldwide data (S117N/S58R/I173L). The distribution of mutations in Eastern Africa is similar to the worldwide distribution (Appendix 3 - Figure 2a), however this region is enriched for single mutation S117N and appears to have not evolved to the double mutant step in the most likely worldwide pathway (S58R,S117N) as frequently as would be expected, suggesting it is at an earlier stage of evolution compared to the worldwide distribution. Finally, double mutation N50I,S117N is overrepresented in Southern Asia (Appendix 3 - Figure 2e), suggesting an alternative evolutionary pathway may be occurring in this region.

Discussion

We have presented a method for predicting the most likely evolutionary trajectories to multiple mutants by parameterizing thermodynamic evolutionary model using Flex ddG predictions. The most likely pathways predicted by our model to the pyrimethamine-resistant quadruple *Pf* DHFR mutant correspond well to those predicted in ***Lozovsky et al.*** (***2009***), generated using experimentally determined IC50 values of *Pf* DHFR pyrimethamine binding. The two most likely pathways based on experimental IC50 values were found in the top two most likely pathways to the quadruple mutation based on our simulations using predictions of binding free energy. Whilst our simulations disagreed with the simulations in ***Lozovsky et al.*** (***2009***) in terms of which were the most frequently realized pathways out of the total number of runs, where a realized pathway does not necessarily have to reach the quadruple mutation, our model is able to capture the most likely order of fixation of mutations leading to a particular multiple mutant in general agreement with the simulations in ***Lozovsky et al.*** (***2009***).

We also simulated the most likely evolutionary trajectories to the *Pv*DHFR quadruple mutation N50I,S58R,S117N,I173L and compared our results to those predicted in ***Jiang et al.*** (***2013***) . 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 to the quadruple 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 mutations that significantly reduce binding affinity will be selectively favoured even if there is a slight reduction in catalytic activity. Indeed, ***Rodrigues et al.*** (***2016***) observed a clear tradeoff between catalytic activity and binding affinity for increased drug resistance and found that whilst many molecular features affect drug resistance, drug binding affinity was the key determinant of drug resistance in later stages of evolution. 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.

A further limitation of the evolutionary model is that it operates in the weak mutation regime, in which the mutation rate is so low that mutations appear and fix in isolation. However, this assumption breaks down when considering large microbial populations where clonal interference means that mutations can arise simultaneously and compete for fixation (***Gerrish and Lenski, 1998***). This can lead to a process known as 'greedy adaptation', in which the mutation of larger beneficial effect is fixed with certainty (***Jain et al., 2011***). Clonal interference has been shown to emerge rapidly in laboratory cultivated *P. falciparum*, where the parasite cycles through only asexual stages, suggesting it may influence the dynamics of the emergence of resistance (***Jett et al., 2020***). The evolutionary model used here may therefore overestimate the fixation probability of mutations with milder beneficial effects and underestimate the fixation probability of mutations with larger beneficial effects (***de Visser and Krug, 2014***). Future iterations of this work could be improved by making using a fixation probability that models of clonal interference such as the work of ***Gerrish and Lenski*** (***1998***) and ***Campos et al.*** (***2004***). However, such models are more difficult to implement as they can require species-specific derivations for certain functions.

Mutations occurring at a drug-binding site may also reduce the protein's thermodynamic stabil-

ity (**Wang et al., 2002**) 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. However, it must also be noted that most proteins are marginally stable (**Vogl et al., 1997**; **Ruvinov et al., 1997**), a property which may have evolved either as an evolutionary spandrel (**Taverna and Goldstein, 2002**; **Goldstein, 2011**) (a characteristic that arises as a result of non-adaptive processes which is then used for adaptive purposes (**Gould et al., 1979**) or due to selection for increased flexibility to improve certain functionalities (**Závodszky et al., 1998**; **Tsou, 1998**)). 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.

Despite the limitations of our computational method to predict evolutionary trajectories by only considering the impact on drug binding, it is able to accurately predict the most likely order of fixation of mutations in a trajectory in general agreement with trajectories determined using experimental values such as IC50 or more complex fitness landscapes informed by multiple parameters including drug concentration and growth rates. This supports the findings in **Rodrigues et al.** (**2016**), that drug binding is a major determinant of resistance, especially at later stages of evolution. It also suggests evolution in such landscapes is more predictable than might be expected, since trajectories can be predicted considering only the impact on binding affinity.

We also inferred evolutionary pathways from the total frequency in worldwide clinical isolate data as well as from different geographical regions. This analysis suggests evolutionary pathways may be inferred from the frequency of mutations found in isolate data, however it requires a large number of isolates to properly sample the mutations in the population. Furthermore, this method can only be used to predict trajectories once resistance has emerged in a population, whereas the computational method presented here can predict evolutionary trajectories before introduction of a new drug.

This analysis also suggested that different regions often follow different evolutionary trajectories and that the most likely evolutionary trajectories predicted by our model, and experimental trajectories, 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. Combination drug SP was first used in 1967 to treat *P. falciparum* in Southeastern Asia, and resistance was first noted that same year on the Thai-Cambodia and Thai-Myanmar borders (**Björkman and Phillips-Howard, 1990**). 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 Southeastern Asia and Africa suggest a single resistant lineage that appeared in Southeastern Asia accumulated multiple mutations, including the triple N51I,C59R,S108N (**Roper et al., 2004**; **Mita et al., 2007**), migrated to Africa and spread throughout the continent (**Maïga et al., 2007**; **McCollum Andrea et al., 2007**, **2008**). Variation in the frequency of *Pf* DHFR 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 (**Turkiewicz et al., 2020**; **Ravenhall et al., 2016**).

Pyrimethamine resistance increased in West Papua in the early 1960s following the introduction of mass drug administration (**Verdrager, 1986**). Microsatellite haplotype analysis suggests C59R,S108N in Melanesia has two lineages, one of which originated in Southeastern Asia whilst the other evolved indigenously **Roper et al.** (**2004**).

Pyrimethamine resistance in South America looks surprisingly different from the distributions in Africa and Southeastern Asia. SP was introduced in South America and low-level resistance was first noted in Colombia in 1981 (**Espinal et al., 1985**). 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 (**?**). 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 (***Zhou et al., 2008***).

In general, the *Pv*DHFR gene is much more polymorphic than *Pf* DHFR gene, with over 20 alleles observed in a limited geographical sampling (***Hawkins et al., 2007***), whereas fewer *Pf* DHFR alleles have been observed despite much more extensive surveillance with non-synonymous changes and insertions/deletions occurring rarely (***Gregson and Plowe, 2005***). It also appears that the origin of *Pv*DHFR pyrimethamine resistance mutation is much more diverse than *Pf* DHFR. ***Hawkins et al.*** (***2008***) investigated isolates from Colombia, India, Indonesia, Papua New Guinea, Sri Lanka, Thailand and Vanuatu and found multiple origins of the double *Pv*DHFR mutant 58R,117N in Thailand, Indonesia and Papua New Guinea/Vanuatu. ***Shaukat et al.*** (***2021***) assessed the resistance mutations in Punjab, Pakistan and found multiple origins of single mutation S117N and a common origin of double mutant 58R,S117N and triple mutant 58R,117N,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.

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 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 than laboratory-based methods. 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.

Methods and Materials

Homology Modelling

Homology modelling was carried out in Modeller (***Webb and Sali, 2016***) 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 structural model. 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 structural model 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.

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 examples of a Rosetta script, resfile and 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 250 times per mutation to produce a distribution of predictions of the change in the free energy of binding. We then found the mean 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 $\epsilon_{X,Y} = \Delta\Delta G_{X,Y} - (\Delta\Delta G_X + \Delta\Delta G_Y)$. 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 $\Delta\Delta G_{X/Y} = \Delta\Delta G_Y + \epsilon_{X,Y}$.

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

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 $\Delta\Delta G^i_X$ for the first mutation X, from the predictions for the double mutation X,Y, $\Delta\Delta G^i_{X,Y}$, to create a set of 250 'predictions' for the change in binding free energy when Y occurs in the background of X, $\Delta\Delta G^i_{X/Y}$ i.e. $\Delta\Delta G^i_{X/Y} = \Delta\Delta G^i_{X,Y} - \Delta\Delta G^i_X$ for $i = \{1, \ldots, 150\}$. To estimate the change in binding free energy when mutation *Z* occurs in the background of mutations *X* and *Y* we calculated $\Delta\Delta G^i_{X,Y/Z} = \Delta\Delta G^i_{X,Y,Z} - \Delta\Delta G^i_{X,Y}$. 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

The Rosetta energy function is a mix of a combination of physic-based and statistics-based potentials and so raw predictions using this function don't up match up with physical energy units (e.g. kcal/mol or kJ/mol). However, the authors of Flex ddG applied a generalized additive model (GAM)-like approach to the Rosetta energy function to reweight its terms and to fit experimentally known values (in kcal/mol). The resulting nonlinear reweighting model reduced the absolute error between the predictions and experimental values and so improved the agreement with experimentally determined interface $\Delta\Delta G$ values. They found that by doing this the Flex ddG predictions of binding free energy changes were in a similar range as experimental binding free energy changes and observed improved correlation and classification of mutations as stabilising or destabilising (***Barlow et al., 2018***). Therefore, we assume Flex ddG can provide approximate predictions of binding free energy changes comparable to experimental changes in kcal/mol and can therefore be used to parameterize a thermodynamic model.

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 1M), 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 $P^{WT}_{unbound}$, we extract one sample *i* from the 250 values of the predicted binding affinity changes for the single mutations to determine the binding free energy after mutation *X*, $\Delta G^i_X = \Delta G_{WT} + \Delta\Delta G^i_X$, and calculate the fitness of each single mutant protein $P^{X(i)}_{unbound}$. We can calculate the probability the mutation will fix in the population using the Kimura fixation probability for a haploid organism

$$p_{fix} = \frac{1 - e^{-2s}}{1 - e^{2sN_e}} \tag{2}$$

where N_e is the effective population size (set to 10^6 as previous models in **Eccleston et al. (2021)**; **Pollock et al. (2012)**) and *s* is the selection coefficient $s = (P^{X,i}_{unbound} - P^{WT}_{unbound})/P^{WT}_{unbound}$. We also took in to account the mutational bias of *Plasmodium falciparum* using the nucleotide mutation matrix calculated in **Lozovsky et al. (2009)**. 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^i_X 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. $\Delta\Delta G^i_{X/Y}$. This continues until the end of the trajectory is reached. If the fixation probabilities of all mutations sampled at a step are effectively zero, no mutation is chosen at that step and the algorithm begins again by choosing a single mutation. Therefore, not all of the runs produce a complete trajectory and some will terminate before reaching the quadruple mutation. The algorithm was written in *R*.

We calculate the probabilities, P_i, of each realized pathway (even those which don't reach the quadruple mutation) by dividing the total number of times that specific pathway occurs by the number of runs. We calculate the probability of a particular step by dividing the number of times that step occurs in all realized pathways by the total number of runs.

SNP data

P. falciparum and *P. vivax* data was obtained from publicly available raw sequence data from European Nucleotide Archive. These data include Illumina raw sequences from the MalariaGEN Community Project for *P. falciparum* (**Ahouidi et al., 2021**) and *P. vivax* (**Adam et al., 2022**). *P. vivax* data additionally includes the Public Health England Malaria Reference Laboratory isolates from returning travelers to UK from regions where malaria is endemic (study accession number ERP128476) (**Benavente et al., 2021**). Data was filtered and processed to SNP data with the methodology described in the recent publications **Turkiewicz et al. (2020)** and **Benavente et al. (2021)** respectively. In this study, we analysed genotype data for 6,762 high-quality isolates from 32 countries across regions of Africa, South Eastern Asia, Oceania and South America to identify the genetic diversity in the *Pf* DHFR gene. A similar analysis was carried out on 847 *P. vivax* isolates spanning 25 countries across Eastern Africa, Southern Asia, Southeastern Asia, Eastern Asia and South America to identify genetic diversity in *Pv*DHFR gene. SNPs occurring in non-unique, low quality or low coverage regions were discarded, and those with a missense effect in the candidate genes were analysed. Functional annotation was done with *SnpEff* (version 5.0) (**Cingolani et al., 2012**) with the following options: *-no-downstream -no-upstream*.

Data availability

Supplementary data can be found at 10.5281/zenodo.7082168

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, MR/R020973/1 and MR/T000171/1) 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. E.M 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 (London School of Hygiene and Tropical Medicine) for the helpful discussions. R.C.E would also like to thank Professor Richard Goldstein (University College London) for introducing her to the evolutionary algorithm used in this paper.

References

Achaz G, Rodriguez-Verdugo A, Gaut BS, Tenaillon O. The reproducibility of adaptation in the light of experimental evolution with whole genome sequencing. Adv Exp Med Biol. 2014; 781:211–31. doi: 10.1007/978-94-007-7347-9_11.

Adam I, Alam MS, Alemu S, Amaratunga C, Amato R, Andrianaranjaka V, Anstey NM, Aseffa A, Ashley E, Assefa A, Auburn S, Barber BE, Barry A, Batista Pereira D, Cao J, Chau NH, Chotivanich K, Chu C, Dondorp AM, Drury E, et al. An open dataset of Plasmodium vivax genome variation in 1,895 worldwide samples. Wellcome Open Res. 2022; 7:136. doi: 10.12688/wellcomeopenres.17795.1.

Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C, Amato R, Amenga-Etego L, Andagalu B, Anderson TJC, Andrianaranjaka V, Apinjoh T, Ariani C, Ashley EA, Auburn S, Awandare GA, Ba H, Baraka V, Barry AE, Bejon P, Bertin GI, et al. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. Wellcome Open Res. 2021; 6:42. doi: 10.12688/wellcomeopenres.16168.2.

Aldeghi M, Gapsys V, de Groot BL. Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation. ACS central science. 2018; 4(12):1708–1718. https://pubmed.ncbi.nlm.nih.gov/30648154https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6311686/, doi: 10.1021/acscentsci.8b00717.

Aldeghi M, Gapsys V, de Groot BL. Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. ACS Central Science. 2019; 5(8):1468–1474. https://doi.org/10.1021/acscentsci.9b00590, doi: 10.1021/acscentsci.9b00590.

Alford RAO, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park HAO, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MAOX, Bonneau R, Bradley P, Dunbrack J R L, Das R, Baker D, Kuhlman B, Kortemme T, Gray JAO. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017; 13:3031–3048. doi: 10.1021/acs.jctc.7b00125.

Andersson DI, Hughes D. Persistence of antibiotic resistance in bacterial populations. FEMS Microbiol Reviews. 2011; 35:901–911. doi: 10.1111/j.1574-6976.2011.00289.x.

Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, Kortemme T. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. The Journal of Physical Chemistry B. 2018; 122(21):5389–5399. https://doi.org/10.1021/acs.jpcb.7b11367, doi: 10.1021/acs.jpcb.7b11367.

Benavente ED, Manko E, Phelan J, Campos M, Nolder D, Fernandez D, Velez-Tobon G, Castaño AT, Dombrowski JG, Marinho CRF, Aguiar ACC, Pereira DB, Sriprawat K, Nosten F, Moon R, Sutherland CJ, Campino S, Clark TG. Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. Nat Commun. 2021; 12(1):3160. doi: 10.1038/s41467-021-23422-3.

Björkman A, Phillips-Howard PA. The epidemiology of drug-resistant malaria. Transactions of The Royal Society of Tropical Medicine and Hygiene. 1990; 84(2):177–180. https://doi.org/10.1016/0035-9203(90)90246-B, doi: 10.1016/0035-9203(90)90246-B.

Blair JMA, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJV. Molecular mechanisms of antibiotic resistance. Nature Reviews Microbiology. 2015; 13(1):42–51. https://doi.org/10.1038/nrmicro3380, doi: 10.1038/nrmicro3380.

Brooks DR, Wang P Fau Read M, Read M Fau Watkins WM, Watkins Wm Fau Sims PF, Sims Pf Fau Hyde JE, Hyde JE. 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):397–405. doi: 10.1111/j.1432-1033.1994.00397.x.

Campos PRA, Adami C, Wilke CO. In: Ciobanu G, Rozenberg G, editors. Modelling Stochastic Clonal Interference Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 21–38. https://doi.org/10.1007/978-3-642-18734-6_2, doi: 10.1007/978-3-642-18734-6_2.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; 6(2):80–92. doi: 10.4161/fly.19695.

Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. BMJ. 2010; 340:c2096. http://www.bmj.com/content/340/bmj.c2096.abstract, doi: 10.1136/bmj.c2096.

Davies J, Davies D. Origins and evolution of antibiotic resistance. Microbiology and molecular biology reviews : MMBR. 2010; 74(3):417–433. https://pubmed.ncbi.nlm.nih.gov/20805405https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2937522/, doi: 10.1128/MMBR.00016-10.

Eccleston RC, Pollock DD, Goldstein RA. 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):e2010057118. http://www.pnas.org/content/118/16/e2010057118.abstract, doi: 10.1073/pnas.2010057118.

Enne VI. Reducing antimicrobial resistance in the community by restricting prescribing: can it be done? Journal of Antimicrobial Chemotherapy. 2010; 65(2):179–182. https://doi.org/10.1093/jac/dkp443, doi: 10.1093/jac/dkp443.

Espinal TCA, Cortes CGT, Guerra P, Arias R AE. Sensitivity of Plasmodium Falciparum to Antimalarial Drugs in Colombia. The American Journal of Tropical Medicine and Hygiene. 1985; 34(4):675–680. https://www.ajtmh.org/view/journals/tpmd/34/4/article-p675.xml, doi: 10.4269/ajtmh.1985.34.675.

Ferlan JT, Mookherjee S Fau Okezie IN, Okezie In Fau Fulgence L, Fulgence L Fau Sibley CH, Sibley CH. 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)):139–150. doi: 10.1016/s0166-6851(01)00207-9.

Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. Genetica. 1998; 102(0):127. https://doi.org/10.1023/A:1017067816551, doi: 10.1023/A:1017067816551.

Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins: Structure, Function, and Bioinformatics. 2011; 79(5):1396–1407. https://doi.org/10.1002/prot.22964, doi: https://doi.org/10.1002/prot.22964.

Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. eLife. 2013; 2:e00631. https://doi.org/10.7554/eLife.00631, doi: 10.7554/eLife.00631.

Gould SJ, Lewontin RC, Maynard Smith J, Holliday R. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London Series B Biological Sciences. 1979; 205(1161):581–598. https://doi.org/10.1098/rspb.1979.0086, doi: 10.1098/rspb.1979.0086.

Gregson A, Plowe CV. Mechanisms of Resistance of Malaria Parasites to Antifolates. Pharmacological Reviews. 2005; 57(1):117. http://pharmrev.aspetjournals.org/content/57/1/117.abstract, doi: 10.1124/pr.57.1.4.

Hastings MD, Porter KM, Maguire JD, Susanti I, Kania W, Bangs MJ, Sibley CH, Baird JK. Dihydrofolate Reductase Mutations in Plasmodium vivax from Indonesia and Therapeutic Response to Sulfadoxine plus Pyrimethamine. The Journal of Infectious Diseases. 2004; 189(4):744–750. https://doi.org/10.1086/381397, doi: 10.1086/381397.

Hawkins VN, Joshi H Fau Rungsihirunrat K, Rungsihirunrat K Fau Na-Bangchang K, Na-Bangchang K Fau Sibley CH, Sibley CH. Antifolates can have a role in the treatment of Plasmodium vivax. Trends, Parasitol. 2007; 25(5):213–222.

Hawkins VN, Auliff A, Prajapati SK, Rungsihirunrat K, Hapuarachchi HC, Maestre A, O'Neil MT, Cheng Q, Joshi H, Na-Bangchang K, Sibley CH. Multiple origins of resistance-conferring mutations in Plasmodium vivax dihydrofolate reductase. Malaria Journal. 2008; 7(1):72. https://doi.org/10.1186/1475-2875-7-72, doi: 10.1186/1475-2875-7-72.

Jain K, Krug J, Park SC. EVOLUTIONARY ADVANTAGE OF SMALL POPULATIONS ON COMPLEX FITNESS LANDSCAPES. Evolution. 2011; 65(7):1945–1955. https://doi.org/10.1111/j.1558-5646.2011.01280.x, doi: https://doi.org/10.1111/j.1558-5646.2011.01280.x.

Jett C, Dia A, Cheeseman IH. Rapid emergence of clonal interference during malaria parasite cultivation. bioRxiv. 2020; p. 2020.03.04.977165. http://biorxiv.org/content/early/2020/03/05/2020.03.04.977165.abstract, doi: 10.1101/2020.03.04.977165.

Jiang PP, Corbett-Detig RB, Hartl DL, Lozovsky ER. Accessible Mutational Trajectories for the Evolution of Pyrimethamine Resistance in the Malaria Parasite Plasmodium vivax. Journal of Molecular Evolution. 2013; 77(3):81–91. https://doi.org/10.1007/s00239-013-9582-z, doi: 10.1007/s00239-013-9582-z.

Karmakar M, Rodrigues CHM, Horan K, Denholm JT, Ascher DB. Structure guided prediction of Pyrazinamide resistance mutations in pncA. Scientific Reports. 2020; 10(1):1875. https://doi.org/10.1038/s41598-020-58635-x, doi: 10.1038/s41598-020-58635-x.

Khan AI, Dinh Dm Fau Schneider D, Schneider D Fau Lenski RE, Lenski Re Fau Cooper TF, Cooper TF. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 2011; 332(6034):1193–1196.

Kompis IM, Islam K Fau Then RL, Then RL. DNA and RNA synthesis: antifolates. Chem, Rev. 2005; 105(2):593–620. doi: 10.1021/cr0301144.

Korsinczky M, Fischer K Fau Chen N, Chen N Fau Baker J, Baker J Fau Rieckmann K, Rieckmann K Fau Cheng Q, Cheng Q. 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):2214–2222. doi: 10.1128/AAC.48.6.2214-2222.2004.

Levy SB, Marshall B. Antibacterial resistance worldwide: causes, challenges and responses. Nature Medicine. 2004; 10(12):S122–S129. https://doi.org/10.1038/nm1145, doi: 10.1038/nm1145.

Lobkovsky A, Koonin E. Replaying the Tape of Life: Quantification of the Predictability of Evolution. Frontiers in Genetics. 2012; 3. https://www.frontiersin.org/article/10.3389/fgene.2012.00246.

Lozovsky ER, Chookajorn T Fau Brown KM, Brown Km Fau Imwong M, Imwong M Fau Shaw PJ, Shaw Pj Fau Kamchonwongpaisan S, Kamchonwongpaisan S Fau Neafsey DE, Neafsey De Fau Weinreich DM, Weinreich Dm Fau Hartl DL, Hartl DL. Stepwise acquisition of pyrimethamine resistance in the malaria parasite. Proc Natl Acad Sci, U S A. 2009; 106(29):12025–12030. doi: 10.1073/pnas.0905922106.

Maïga O, Djimdé AA, Hubert V, Renard E, Aubouy A, Kironde F, Nsimba B, Koram K, Doumbo OK, Bras JL, Clain J. A Shared Asian Origin of the Triple-Mutant dhfr Allele in Plasmodium falciparum from Sites across Africa. The Journal of Infectious Diseases. 2007; 196(1):165–172. https://doi.org/10.1086/518512, doi: 10.1086/518512.

McCollum Andrea M, Basco Leonardo K, Tahar R, Udhayakumar V, Escalante Ananias A. Hitchhiking and Selective Sweeps of Plasmodium falciparum Sulfadoxine and Pyrimethamine Resistance Alleles in a Population from Central Africa. Antimicrobial Agents and Chemotherapy. 2008; 52(11):4089–4097. https://doi.org/10.1128/AAC.00623-08, doi: 10.1128/AAC.00623-08.

McCollum Andrea M, Mueller K, Villegas L, Udhayakumar V, Escalante Ananias A. Common Origin and Fixation of Plasmodium falciparum dhfr and dhps Mutations Associated with Sulfadoxine-Pyrimethamine Resistance in a Low-Transmission Area in South America. Antimicrobial Agents and Chemotherapy. 2007; 51(6):2085–2091. https://doi.org/10.1128/AAC.01228-06, doi: 10.1128/AAC.01228-06.

Melnyk AH, Wong A, Kassen R. The fitness costs of antibiotic resistance mutations. Evol, Appl. 2015; 8(3):273–283. doi: 10.1111/eva.12196.

Mita T, Tanabe K, Takahashi N, Tsukahara T, Eto H, Dysoley L, Ohmae H, Kita K, Krudsood S, Looareesuwan S, Kaneko A, Björkman A, Kobayakawa T. Independent Evolution of Pyrimethamine Resistance in Plasmodium falciparum Isolates in Melanesia. Antimicrobial Agents and Chemotherapy. 2007; 51(3):1071–1077. https://doi.org/10.1128/AAC.01186-06, doi: 10.1128/AAC.01186-06.

Palmer ME, Moudgil A, Feldman MW. Long-term evolution is surprisingly predictable in lattice proteins. Journal of The Royal Society Interface. 2013; 10(82):20130026. https://doi.org/10.1098/rsif.2013.0026, doi: 10.1098/rsif.2013.0026.

Pollock DD, Pollard ST, Shortt JA, Goldstein RA. In: Pontarotti P, editor. Mechanistic Models of Protein Evolution Cham: Springer International Publishing; 2017. p. 277–296. https://doi.org/10.1007/978-3-319-61569-1_15, doi: 10.1007/978-3-319-61569-1_15.

Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proceedings of the National Academy of Sciences. 2012; 109(21):E1352. http://www.pnas.org/content/109/21/E1352.abstract, doi: 10.1073/pnas.1120084109.

Portelli S, Myung Y, Furnham N, Vedithi SC, Pires DEV, Ascher DB. Prediction of rifampicin resistance beyond the RRDR using structure-based machine learning approaches. Scientific Reports. 2020; 10(1):18120. https://doi.org/10.1038/s41598-020-74648-y, doi: 10.1038/s41598-020-74648-y.

Ravenhall M, Benavente ED, Mipando M, Jensen AT, Sutherland CJ, Roper C, Sepúlveda N, Kwiatkowski DP, Montgomery J, Phiri KS, Terlouw A, Craig A, Campino S, Ocholla H, Clark TG. Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the Plasmodium falciparum population in Malawi. Malar J. 2016; 15(1):575. doi: 10.1186/s12936-016-1634-6.

Rodrigues JV, Bershtein S, Li A, Lozovsky ER, Hartl DL, Shakhnovich EI. Biophysical principles predict fitness landscapes of drug resistance. Proceedings of the National Academy of Sciences. 2016; 113(11):E1470. http://www.pnas.org/content/113/11/E1470.abstract, doi: 10.1073/pnas.1601441113.

Roper C, Pearce R, Nair S, Sharp B, Nosten F, Anderson T. Intercontinental spread of pyrimethamine-resistant malaria. Science. 2004; 305(5687):1124. doi: 10.1126/science.1098876.

Ruvinov S, Wang L, Ruan B, Almog O, Gilliland GL, Eisenstein E, Bryan PN. Engineering the Independent Folding of the Subtilisin BPN Prodomain : Analysis of Two-State Folding versus Protein Stability. Biochemistry. 1997; 36(34):10414–10421. https://doi.org/10.1021/bi9703958, doi: 10.1021/bi9703958.

Sanjuán R, Cuevas Jm Fau Moya A, Moya A Fau Elena SF, Elena SF. Epistasis and the adaptability of an RNA virus. Genetics. 2005; 170:1001–1008. doi: 10.1534/genetics.105.040741.

Shaukat A, Ali Q, Raud L, Wahab A, Khan TA, Rashid I, Rashid M, Hussain M, Saleem MA, Sargison ND, Chaudhry U. Phylogenetic analysis suggests single and multiple origins of dihydrofolate reductase mutations in Plasmodium vivax. Acta Trop. 2021; 215:105821. doi: 10.1016/j.actatropica.2020.105821.

Sirawaraporn W, Sathitkul T, Sirawaraporn R, Yuthavong Y, Santi DV. Antifolate-resistant mutants of Plasmodium falciparum dihydrofolate reductase. Proc Natl Acad Sci U S A. 1997; 94:1124–1129. doi: 10.1073/pnas.94.4.1124.

Snounou G, White NJ. The co-existence of Plasmodium: sidelights from falciparum and vivax malaria in Thailand. Trends in Parasitology. 2004; 20(7):333–339. https://www.sciencedirect.com/science/article/pii/S147149220400128X, doi: https://doi.org/10.1016/j.pt.2004.05.004.

Starr TN, Thornton JW. Epistasis in protein evolution. Protein science : a publication of the Protein Society. 2016; 25(7):1204–1218. https://pubmed.ncbi.nlm.nih.gov/26833806https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4918427/, doi: 10.1002/pro.2897.

Sun D, Jeannot K, Xiao Y, Knapp CW. Editorial: Horizontal Gene Transfer Mediated Bacterial Antibiotic Resistance. Frontiers in microbiology. 2019; 10:1933–1933. https://pubmed.ncbi.nlm.nih.gov/31507555https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6718914/, doi: 10.3389/fmicb.2019.01933.

Sundqvist M, Geli P Fau Andersson DI, Andersson Di Fau Sjölund-Karlsson M, Sjölund-Karlsson M Fau Runehagen A, Runehagen A Fau Cars H, Cars H Fau Abelson-Storby K, Abelson-Storby K Fau Cars O, Cars O Fau Kahlmeter G, Kahlmeter G. Little evidence for reversibility of trimethoprim resistance after a drastic reduction in trimethoprim use. J Antimicrob Chemother. 2010; 65:350–360. doi: 10.1093/jac/dkp387.

Szendro Ivan G, Franke J, de Visser JAGM, Krug J. Predictability of evolution depends nonmonotonically on population size. Proceedings of the National Academy of Sciences. 2013; 110(2):571–576. https://doi.org/10.1073/pnas.1213613110, doi: 10.1073/pnas.1213613110.

Tamer YT, Gaszek IK, Abdizadeh H, Batur TA, Reynolds KA, Atilgan AR, Atilgan C, Toprak E. 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):1533–1550. https://doi.org/10.1093/molbev/msz086, doi: 10.1093/molbev/msz086.

Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins: Structure, Function, and Bioinformatics. 2002; 46(1):105–109. https://doi.org/10.1002/prot.10016, doi: https://doi.org/10.1002/prot.10016.

Tsou CL. Active site flexibility in enzyme catalysis. Ann N Y Acad Sci. 1998; 864:1–8. doi: 10.1111/j.1749-6632.1998.tb10282.x.

Turkiewicz A, Manko E, Sutherland CJ, Diez Benavente E, Campino S, Clark TG. 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):e1009268. doi: 10.1371/journal.pgen.1009268.

Verdrager J. Epidemiology of the emergence and spread of drug-resistant falciparum malaria in South-East Asia and Australasia. J Trop Med Hyg. 1986; 89(6):277–89.

de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics. 2014; 15(7):480–490. https://doi.org/10.1038/nrg3744, doi: 10.1038/nrg3744.

Vogl T, Jatzke C, Hinz HJ, Benz J, Huber R. Thermodynamic Stability of Annexin V E17G: Equilibrium Parameters from an Irreversible Unfolding Reaction. Biochemistry. 1997; 36(7):1657–1668. https://doi.org/10.1021/bi962163z, doi: 10.1021/bi962163z.

Vogwill T, MacLean RC. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. Evolutionary applications. 2015; 8(3):284–295. https://pubmed.ncbi.nlm.nih.gov/25861386https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380922/, doi: 10.1111/eva.12202.

Wang P, Read M, Sims PFG, Hyde JE. 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):979–986.

Wang X, Minasov G, Shoichet BK. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. J Mol Biol. 2002; 320(1):85–95. doi: 10.1016/s0022-2836(02)00400-x.

Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016; 54:5.6.1–5.6.37. doi: 10.1002/cpbi.3.

Weinreich DM, Watson RA, Chao L. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution. 2005; 59(6):1165–74.

Weinreich DM, Delaney NF, DePristo MA, Hartl DL. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science. 2006; 312(5770):111. http://science.sciencemag.org/content/312/5770/111.abstract, doi: 10.1126/science.1123539.

WHO. World Malaria Report: 2021. World Health Organization, Geneva, Switzerland; 2021.

Yuthavong Y, Yuvaniyama J Fau Chitnumsub P, Chitnumsub P Fau Vanichtanankul J, Vanichtanankul J Fau Chusacultanachai S, Chusacultanachai S Fau Tarnchompoo B, Tarnchompoo B Fau Vilaivan T, Vilaivan T Fau Kamchonwongpaisan S, Kamchonwongpaisan S. Malarial (Plasmodium falciparum) dihydrofolate reductase-thymidylate synthase: structural basis for antifolate resistance and development of effective inhibitors. Parasitology,. 2005; 130(0031-1820 (Print)):249–259. doi: 10.1017/s003118200400664x.

Závodszky P, Kardos J, Svingor A, Petsko GA. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. Proceedings of the National Academy of Sciences. 1998; 95(13):7406. http://www.pnas.org/content/95/13/7406.abstract, doi: 10.1073/pnas.95.13.7406.

Zhou Z, Griffing SM, de Oliveira AM, McCollum AM, Quezada WM, Arrospide N, Escalante AA, Udhayakumar V. Decline in sulfadoxine-pyrimethamine-resistant alleles after change in drug policy in the Amazon region of Peru. Antimicrob Agents Chemother. 2008; 52(2):739–41. doi: 10.1128/aac.00975-07.

Appendix 1

Comparison of Flex ddG predictions for 35 to 250 runs

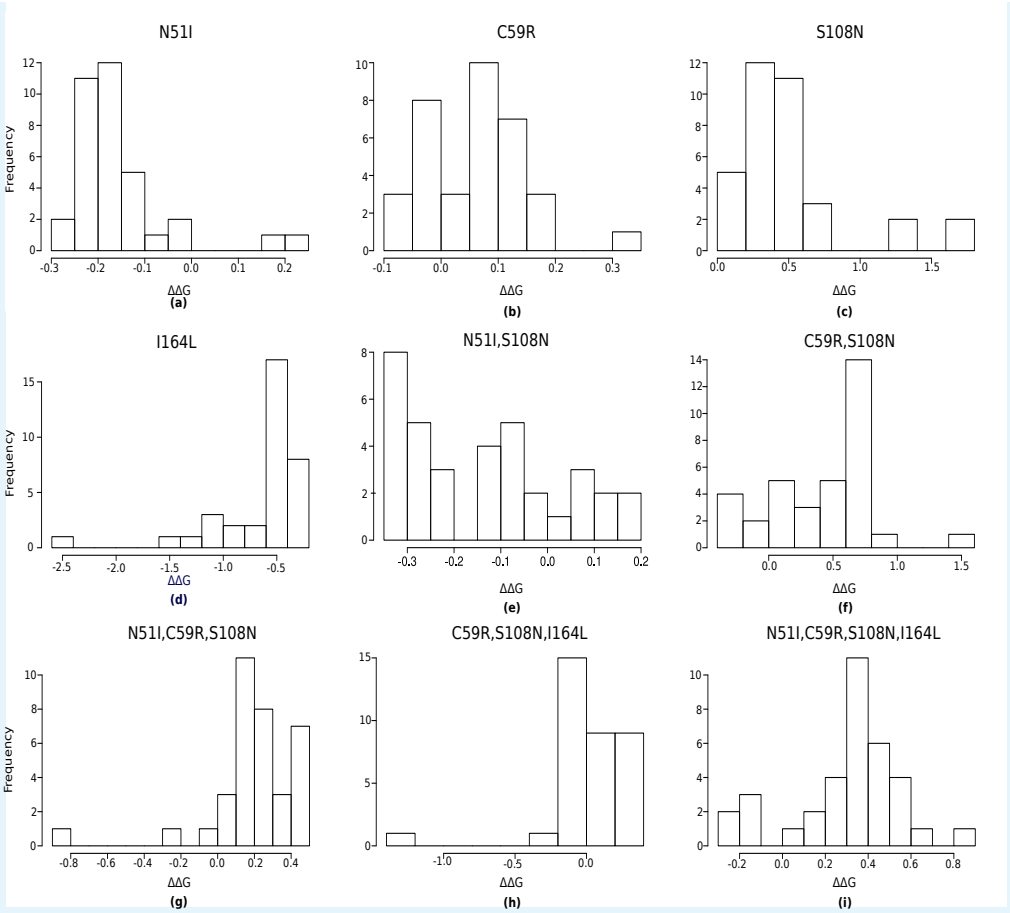
For the set of four *Pf*DHFR mutations (N51I, C59R, S108N and I164L) and combinations thereof studied in ***Lozovsky et al.*** (***2009***), we ran 35 runs per mutation (as suggested by the Flex ddG authors (***Barlow et al., 2018***)) and found the average for each distribution and used these predictions to calculate the sum and the interaction energies for the multiple mutants. We compared the predictions of the binding free energy change, the sum of the independent changes for multiple mutants and the interaction energy to the experimental data from ***Lozovsky et al.*** (***2009***) and observed Pearson correlations of 0.536, 0.580 and 0.900, respectively (Appendix 1 - Table 1). We also determined the number of correctly classified predictions. For the binding free energy predictions, 5/9 predictions were correctly classified as stabilizing or destabilizing, 4/5 of the sum of the independent impacts were correctly classified as stabilizing or destabilizing and 2/5 interaction energy predictions were correctly classified as either positive or negative. Therefore, whilst the predictions for 35 runs achieved a good correlation with the data, the predictions of the interaction energy (and so the epistasis) using this data were correctly classified for less than half of the data set.

Examining the distributions of the predicted change in binding free energy for 35 runs for each of the mutations considered in ***Lozovsky et al.*** (***2009***) (Appendix 1 - Figure 1) we can see that the distributions are not well characterized. We therefore decided to carry out a larger number of runs per prediction and determine the number of runs required for the rank order of the mutations to converge. We found the rank order of the average of the distributions sufficiently converged by 250 runs (Appendix 1 - Figure 3), as demonstrated in Appendix 1 - Figure 4 where the gradient of the average for each mutation is close to zero. Whilst more runs may have achieved better convergence, because of the time it takes to run Flex ddG it is important to achieve a balance between efficiency and accuracy. Whilst these new distributions are still not Gaussian, the distributions are better explored (Appendix 1 - Figure 2). We compared the predictions for 250 runs and the data from ***Lozovsky et al.*** (***2009***) (Table 1) and observed a correlation of 0.611 for the binding free energy data, 0.660 for the sum of the independent predictions for multiple mutants and 0.756 for the interaction energy. We found 8/9 binding free energy predictions were correctly classified, 4/5 of the sum of the independent predictions were correctly classified and 4/5 of the interaction energies were correctly classified. We therefore conclude that the predictions for n=250 runs present a better agreement with the data presented in ***Lozovsky et al.*** (***2009***) in terms of compromising between correlation and correct classification, both of which are important here.

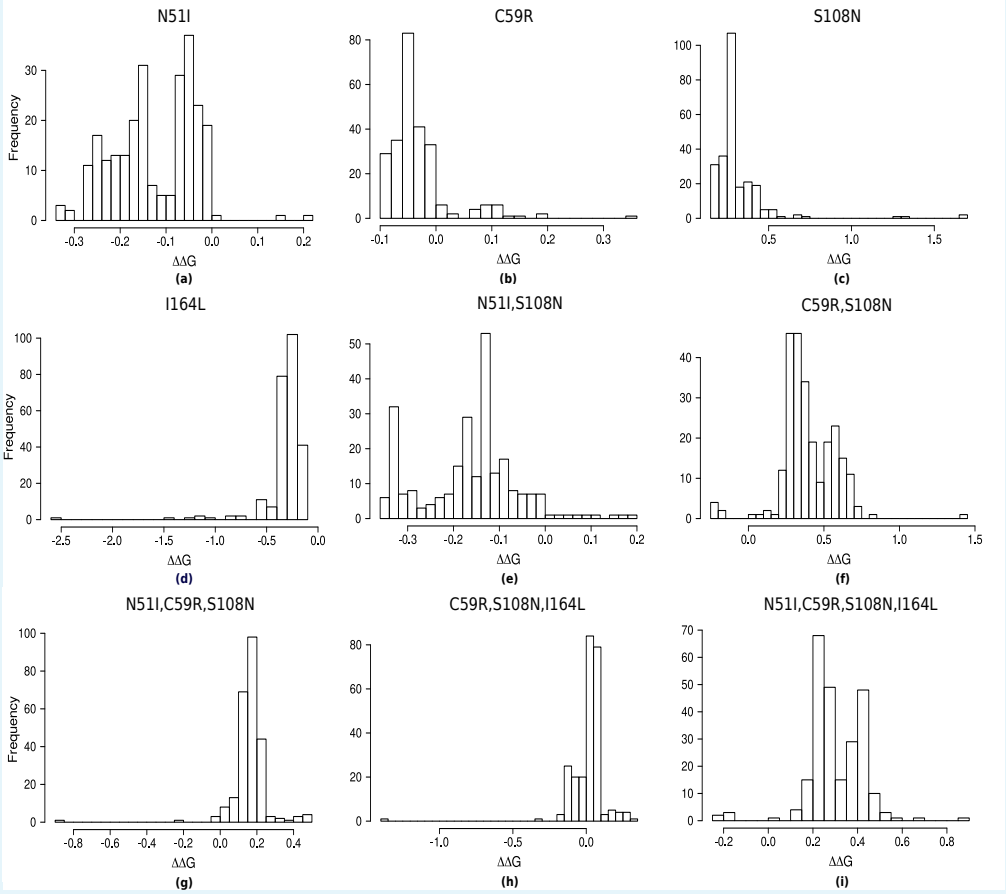
Mutation	$\Delta\Delta G^{*}_{exp}$ (kcal/mol)	Exp. Sum**	Exp I.E.***	$\Delta\Delta G^{\dagger}_{FlexddG}$ (kcal/mol)	Sum‡	I.E.§
N51I	-0.783			-0.156		
C59R	-0.184			0.059		
S108N	1.297			0.521		
I164L	-0.351			-0.661		
N51I,S108N	1.89	0.514	1.376	-0.132	0.365	-0.497
C59R,S108N	2.29	1.113	1.177	0.399	0.580	-0.183
N51I,C59R,S108N	2.595	0.33	2.265	0.196	0.425	-0.228
C59R,S108N,I164L	3.283	0.762	2.521	-0.004	-0.081	0.077
N51I,C59R,S108N,I164L	3.761	-0.021	3.782	0.306	-0.237	0.542
Pearson Correlation				0.536	0.580	0.900
Correctly Classified				5/9	4/5	2/5

Appendix 1—table 1. Correlation between Flex ddG predictions for 35 runs and experimental data (see table 4 of ***Sirawaraporn et al.*** (***1997***)) for *Pf* DHFR pyrimethamine resistance mutations.

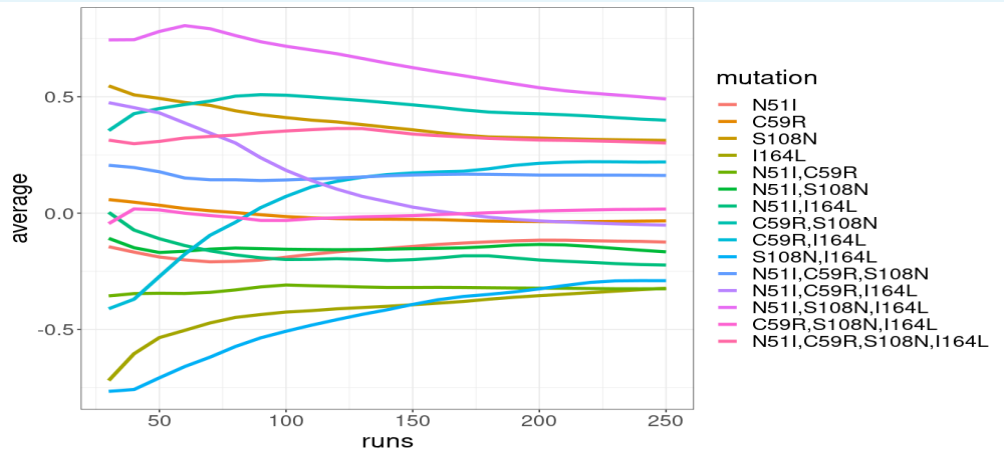
*Experimentally measured *Pf* DHFR pyrimethamine binding free energy change data from ***Sirawaraporn et al.*** (***1997***)
**Sum of experimental values of binding free energy change for independent mutations
***Interaction energy calculated as the difference between experimentally measured values of 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 calculated as the average of the distribution of runs. Free energy predictions from Rosetta are in Rosetta Energy Units, however the authors of Flex ddG applied a generalized additive model to re-weight the predictions and make the output more comparable to units of kcal/mol (***Barlow et al., 2018***)
‡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.



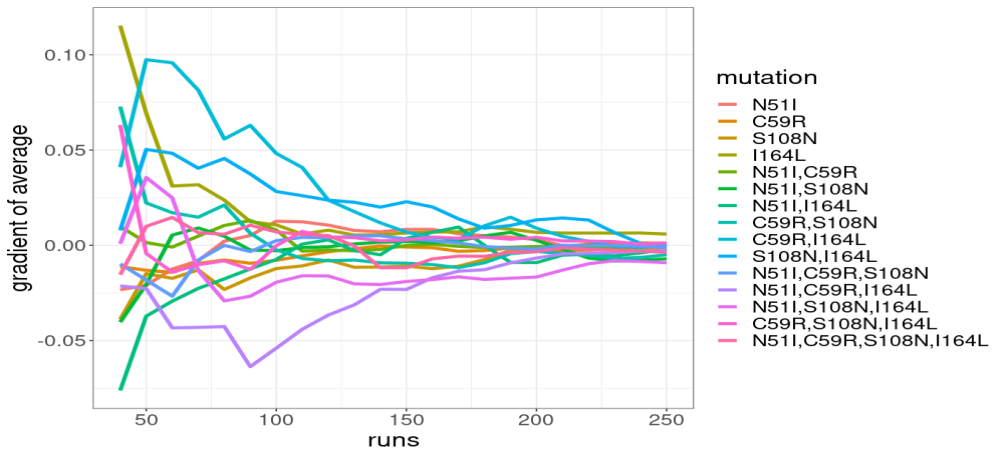
Appendix 1—figure 1. The distribution of pyrimethamine-*Pf* DHFR binding free energy changes predicted by Flex ddG for 35 runs for subset of mutations considered in ***Sirawaraporn et al.*** (***1997***) namely a) N51I, b) C59R, c) S108N, d) I164L, e) N51I,S108N, f) C59R,S108N, g) N51I,C59R,S108N, h) C59R,S108N,I164L and i) N51I,C59R,S108N,I164L



Appendix 1—figure 2. The distribution of pyrimethamine-*Pf* DHFR binding free energy changes predicted by Flex ddG for 250 runs for subset of mutations considered in ***Sirawaraporn et al.*** (***1997***) namely a) N51I, b) C59R, c) S108N, d) I164L, e) N51I,S108N, f) C59R,S108N, g) N51I,C59R,S108N, h) C59R,S108N,I164L and i) N51I,C59R,S108N,I164L



Appendix 1—figure 3. The average of the Flex ddG prediction distributions for n=30,. . . ,250 runs for the *Pf* DHFR mutation combinations.



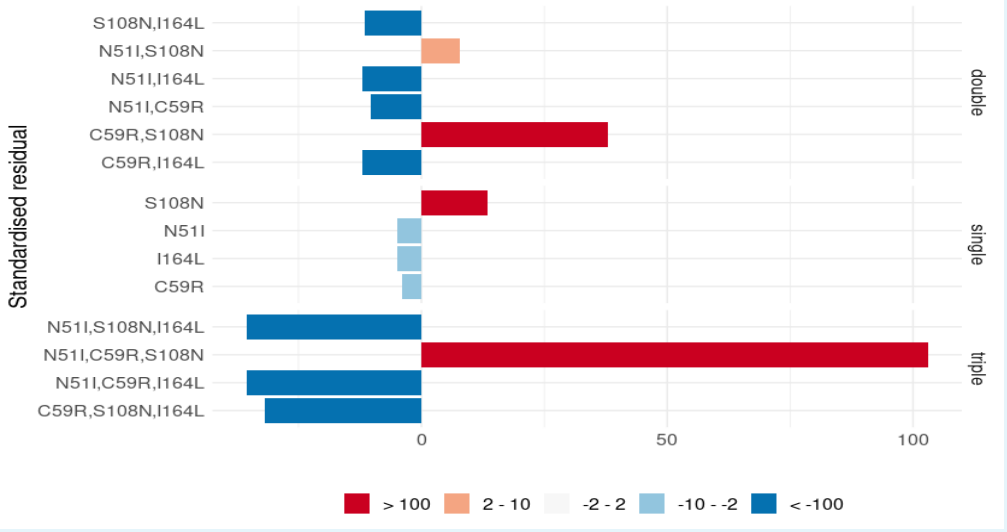
Appendix 1—figure 4. The gradient of the average of the prediction distributions for n=30,…,250 runs for the *Pf*DHFR mutation combinations

Appendix 2

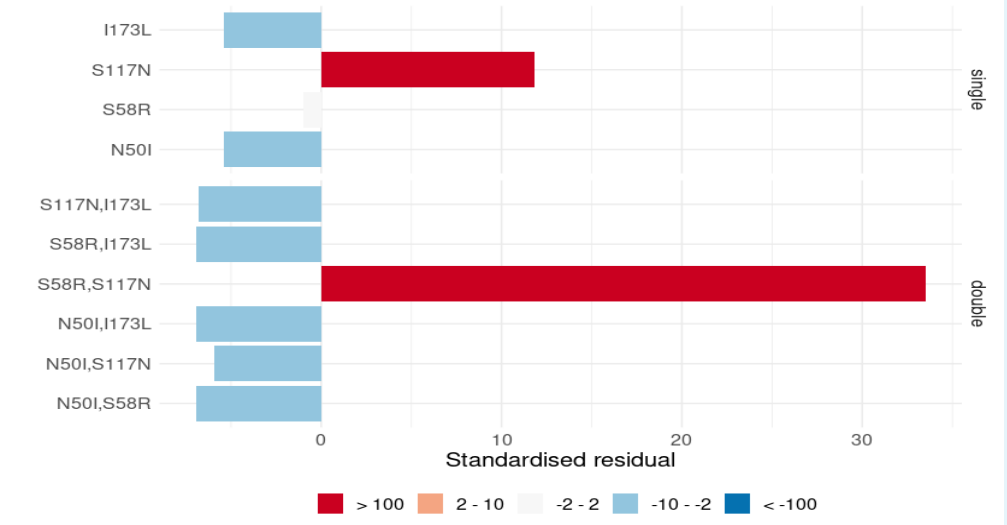
Assessing the significance of the frequency distribution of mutations in the worldwide isolate data

We performed a Chi-squared test on the worldwide distribution of the four *Pf*DHFR mutations considered here, to determine if the mutations involved in the most likely inferred pathway were overrepresented compared to what we would expect under a null hypothesis in which there is no preferred pathway. Under this null hypothesis, if there was no preferred pathway, the single mutations would be observed at the same frequency, the double mutations would be observed at the same frequency and the triple mutations would be observed at the same frequency. Therefore, we carried out three separate Chi-squared tests on the distributions of the single, double and triple mutations. (It would be difficult to carry out a Chi-squared test on the combined distribution of these mutations because we would not expect the single, double and triple mutations to have the same frequency and it would be difficult to determine what their appropriate relative frequencies would be). The Chi-squared tests determined all three distributions were significantly different from the null hypothesis ($p_{singles} < 0.01$, $p_{doubles} = 0$, $p_{triples} = 0$). Analysing the residuals of each distribution (Appendix 2 - Figure 1), S108N was found to be overrepresented in the distribution of single mutations, C59R,S108N was overrepresented in the distribution of double mutations and N51I,C59R,S108N was overrepresented in the triple mutation distribution. These mutations are the single, double and triple mutations involved in the most likely inferred pathway for the worldwide data, supporting our assertion this is the most likely stepwise trajectory to the quadruple mutation.

We performed a Chi-squared test for the distribution of the *Pv*DHFR mutations in a similar way to *Pf*DHFR, however the frequency of triple mutants in the *Pv*DHFR worldwide dataset was not large enough to accurately carry out the test on the distribution of triple mutations. Therefore, we only carried out the test for the distribution of single and double mutations (Appendix 2 - Figure 1), with the null hypothesis that if there were no preferred order of fixation, the frequency of the single mutations would be equal and the frequency of the double mutations would be equal. The Chi-squared test revealed the distribution of both the single and double mutants is significantly different from what we would expect from the null hypothesis ($p_{singles} < 0.01$ and $p_{doubles} < 0.01$ for the single and double distributions, respectively). Analysing the residuals of each distribution, S117N was found to be overrepresented among the single mutations and S58R,S117N was found to be overrepresented among the double mutations. This provides support for the idea that epistasis determines the order of fixation and suggests that the most likely pathway to the only triple mutation observed (S58R,S117N,I173L) occurs via pathway S117N/S58R/I173L. This corresponds to the fifth most likely pathway to a triple mutation when considering all pathways observed in our simulations, however this pathway is not observed in any of the most frequent pathways at any of the four concentrations studied in ***Jiang et al.*** (***2013***).



Appendix 2—figure 1. The standardized residuals of the individual *Pf*DHFR mutations from the Chi-squared tests applied to single, double and triple mutants.



Appendix 2—figure 2. The standardized residuals of the individual *Pv*DHFR mutations from the Chi-squared tests applied to single and double mutants.

Appendix 3

Assessing the significance of the frequency distribution of mutations per region

The frequency of the mutations in the separate regions is often too small to reliably carry out a Chi-squared test to determine if the regional distributions are significantly different from the worldwide distribution. Therefore, for each region, a sample of size N (where N is the size of the dataset from that region) was drawn with replacement from the worldwide distribution, and this was repeated 50,000 per region to create a dataset of 50,000 bootstrap samples per region. Comparing the frequency of mutations in the samples to the regional data highlights those mutations whose frequency differs significantly from what would be expected from the worldwide distribution.

*Pf*DHFR

The distribution of mutations in Western Africa is similar to the worldwide distribution with the exception of mutations N51I,C59R,S108N and N51I,C59R,S108N,I164L which are over-represented and underrepresented in the region, respectively (Appendix 3 - Figure 1g). This suggests the region is following the same pathway as the worldwide distribution but that the evolution is at an earlier stage. Conversely, in Southeastern Asia (Appendix 3 - Figure 1f), N51I,C59R,S108N was underrepresented in the region, whilst N51I,C59R,S108N,I164L was overrepresented, suggesting the region is following the same pathway as the worldwide distribution but evolution to the quadruple mutation had occurred more often than expected. Analysis of the distribution of mutations in Southern Asia (Appendix 3 - Figure 1e) suggests double mutation C59R,S108N is overrepresented in this region and the evolution in this region is more concentrated around the double mutant step in the pathway than would be expected. Triple mutations N51I,C59R,S108N and C59R,S108N,I164L are under- and overrepresented in this region, respectively, suggesting the alternative pathway S108N/C59R/I164L/N51I is more prevalent in this region than expected.

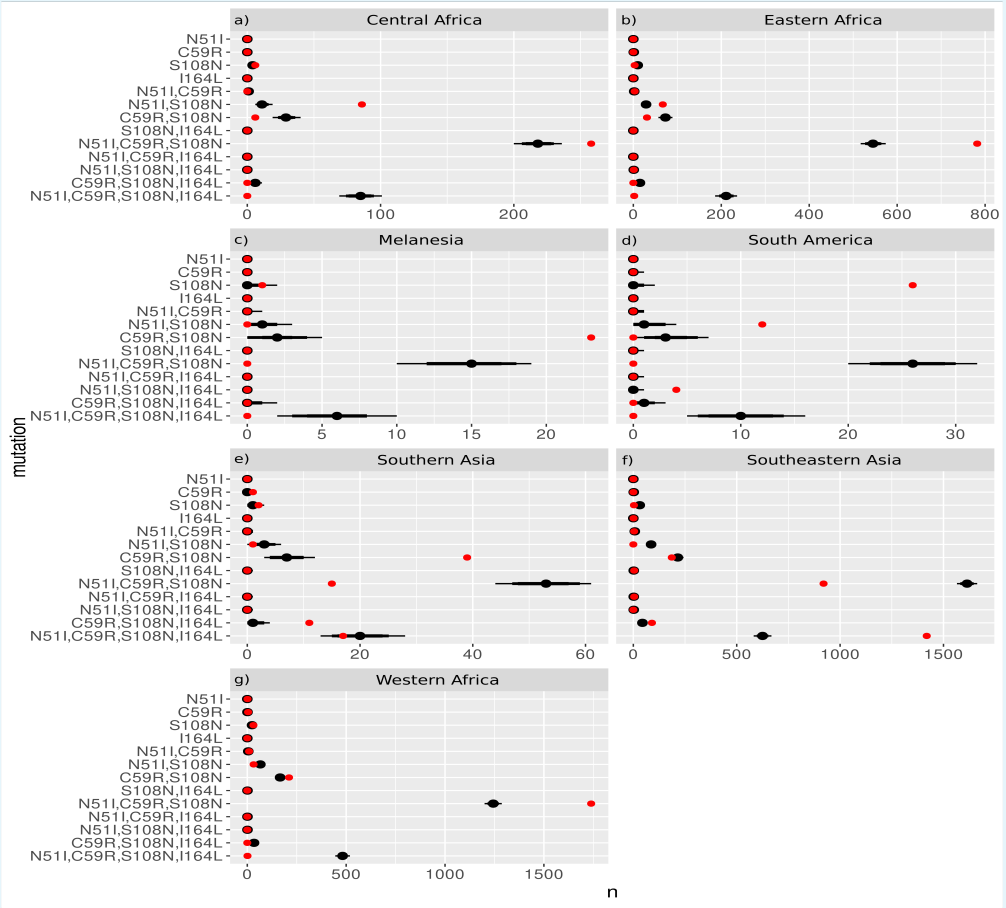
The distribution of mutations in Eastern Africa (Appendix 3 - Figure 1b) is similar to the worldwide distribution with the exception of double mutations N51I,S108N and C59R,S108N which are slightly over- and underrepresented in the region, respectively. Furthermore, triple mutation N51I,C59R,S108N and the quadruple mutation are over-and underrepresented, respectively. This suggests a true preference for the double mutant step S108N/N51I over S108N/C59R in the pathway in this region compared to the worldwide distribution. Furthermore, similar to Western Africa, the overrepresentation of the triple mutant step in the most likely inferred pathway and the underrepresentation of the quadruple mutation suggests this region is at an earlier stage in evolution compared to what would be expected from the worldwide distribution.

Similar to Eastern Africa, the distribution of mutations in Middle Africa (Appendix 3 - Figure 1a) showed significant differences in the frequency of double mutations N51I,S108N and C59R,S108N which were over- and underrepresented, respectively. Triple mutation N51I,C59R,S108N was overrepresented in this region and the quadruple mutation was underrepresented. This suggests that like Eastern Africa, the evolutionary pathway in Middle Africa shows a significant preference for the double mutant step S108N/N51I over S108N/C59R and that the evolution is at an earlier stage in this region than would be expected from the worldwide distribution i.e. evolution to the quadruple mutation has not occurred as frequently as would be expected.

The distribution of mutations in South America is markedly different from the worldwide distribution, with single mutation S108N, double mutation N51I,S108N and triple mu-

tation N51I,S108N,I164L all overrepresented, whilst triple mutation N51I,C59R,S108N and the quadruple mutation are underrepresented. This suggests this region has a significant preference for the mutations involved in the most likely inferred pathway in this region S108N/N51I/I164L and the evolution is following a significantly different trajectory than the worldwide distribution.

Finally, we analysed the distribution of mutations in Melanesia and found the double mutation C59R,S108N is significantly overrepresented, whilst N51I,C59R,S108N and the quadruple mutation, which were both absent from this region, were underrepresented. This suggests the evolution in this region is at a much earlier stage than would be expected from the worldwide data.



Appendix 3—figure 1. The frequency distributions of the *Pf*DHFR mutations from the 50,000 samples taken from the worldwide distribution with replacement for sample sizes equal to the regional datasets from a) Middle Africa, b) Eastern Africa, c) Melanesia, d) South America, e) Southern Asia, f) Southeastern Asia, and g) Western Africa. The red dots show the frequency of each mutation from the regional datasets and the black distributions show the 69%, 80% and 90% quantile intervals of frequency distributions from the samples.

*Pv*DHFR

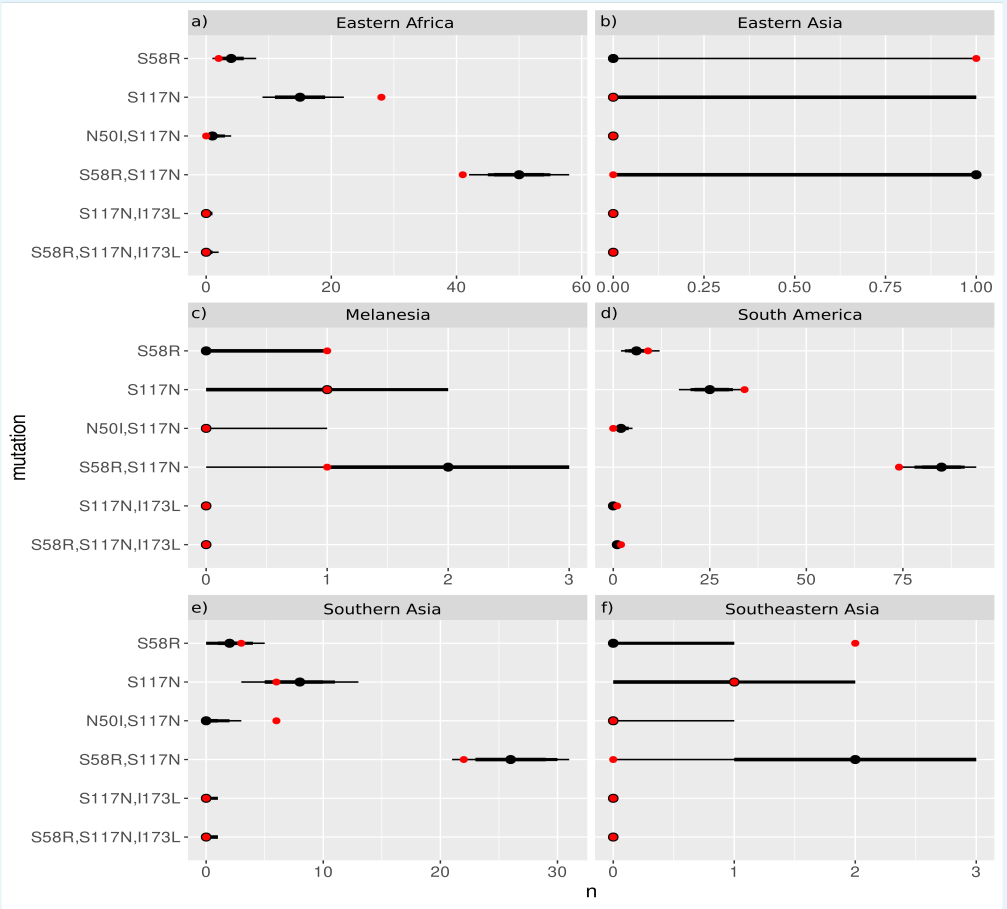
The distribution of the four *Pv*DHFR mutations in South America is similar to the worldwide distribution (Appendix 3 - Figure 2d) with all of the observed mutations occurring at frequencies within or just outside the expected range. This supports our inference that the evolution in South America is following the same most likely pathway as the worldwide data.

In Eastern Africa, S117N was found to be overrepresented and S58R,S117N was found

to be underrepresented compared to the worldwide distribution (Appendix 3 - Figure 2a). This suggests evolution to the double mutation has not occurred as frequently in this region as would be expected from the worldwide distribution.

In Southern Asia, the distribution of mutations was as expected from the worldwide distribution, with the exception of N50I,S117N, which was overrepresented in this region (Appendix 3 - Figure 2e). This suggests the alternative pathway S117N/N50I inferred from the frequency data from this region is more prevalent than would be expected.

The frequency of the four *Pv*DHFR mutations in Eastern Asia, Southeastern Asia and Melanesia is very low, therefore it is difficult to draw many conclusions about the frequency of mutations in these regions. From the distribution plots (Appendix 3 - Figures 2b, 2f and 2c for Eastern Asia, Southeastern Asia and Melanesia, respectively), the mutations appear to be found at similar frequencies to what would be expected from the worldwide distribution, however due to the low frequencies it is difficult to conclude that definitively. More data is required from these areas to draw conclusions regarding the distribution of their mutations and the evolutionary pathways they appear to be following.



Appendix 3—figure 2. The frequency distributions of the *Pv*DHFR mutations from the 50,000 samples taken from the worldwide distribution with replacement for sample sizes equal to the regional datasets from a) Eastern Africa, b) Eastern Asia, c) Melanesia, d) South America, e) Southern Asia and f) Southeastern Asia. The distribution from Central America was not analysed because it did not contain any combinations of the four *Pv*DHFR mutations being studied. The red dots show the frequency of each mutation from the regional datasets and the black distributions show the 69%, 80% and 90% quantile intervals of frequency distributions from the samples.