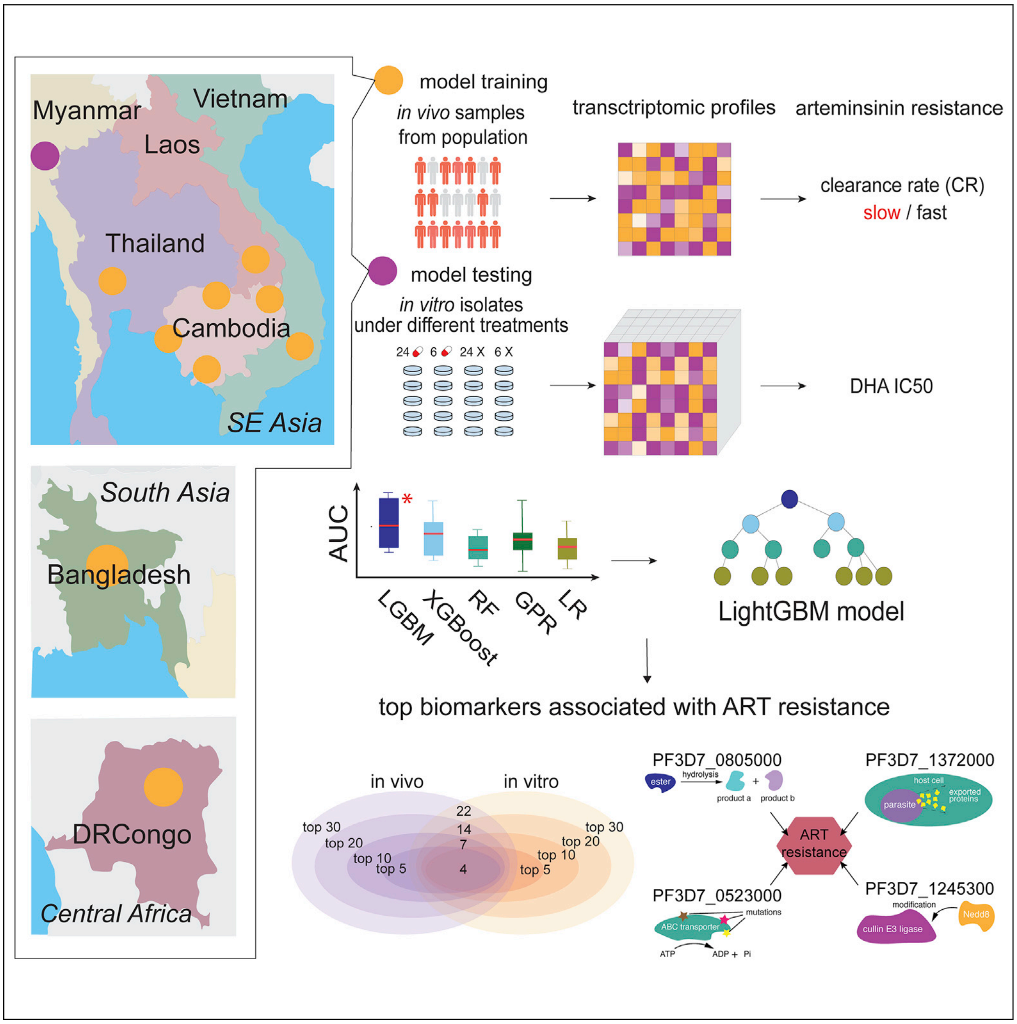
iScience

CellPress
OPEN ACCESS

Article

Machine learning for artemisinin resistance in malaria treatment across *in vivo-in vitro* platforms



Hanrui Zhang,
Jiantao Guo,
Hongyang Li,
Yuanfang Guan

gyuanfan@umich.edu

Highlights

Artemisinin resistance can be predicted from transcriptomes by machine learning

Our model can be transferred between *in vivo* and *in vitro* and different platforms

We identified top transcription biomarkers of artemisinin resistance

Zhang et al., iScience *25*, 103910
March 18, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.103910

iScience

CellPress
OPEN ACCESS

Article

Machine learning for artemisinin resistance in malaria treatment across *in vivo-in vitro* platforms

Hanrui Zhang,[1] Jiantao Guo,[1] Hongyang Li,[1] and Yuanfang Guan[1,2,3,*]

SUMMARY

Drug resistance has been rapidly evolving with regard to the first-line malaria treatment, artemisinin-based combination therapies. It has been an open question whether predictive models for this drug resistance status can be generalized across *in vivo-in vitro* transcriptomic measurements. In this study, we present a model that predicts artemisinin treatment resistance developed with transcriptomic information of *Plasmodium falciparum*. We demonstrated the robustness of this model across *in vivo* clearance rate and *in vitro* IC50 measurement and based on different microarray and data processing modalities. The validity of the algorithm is further supported by its first placement in the DREAM Malaria challenge. We identified transcription biomarkers to artemisinin treatment resistance that can predict artemisinin resistance and are conserved in their expression modules. This is a critical step in the research of malaria treatment, as it demonstrated the potential of a platform-robust, personalized model for artemisinin resistance using molecular biomarkers.

INTRODUCTION

Malaria raises major public health concerns in southeastern Asia and Africa (Asenso-Okyere et al., 2011; Conn et al., 2018; Dhiman, 2019; Mbacham et al., 2019; Organization and Others, 2020; Sachs and Malaney, 2002; Tabbabi et al., 2020; WHO, 2020). *Plasmodium falciparum*, one of the five *Plasmodium* species leading to malaria, is the main cause of mortality, resulting in 400,000 deaths each year (Fact Sheet about Malaria, n.d., Cowman et al., 2016; Talapko et al., 2019). The most effective treatment is artemisinin (ART)-based combination therapies, which has been used as the first-line treatment for malaria since late 1990s (Miller and Su, 2011). Today, malaria remains to be a global health threat, and drug resistance is a major contributor (Dhiman, 2019; Dondorp et al., 2009; Mok et al., 2015). After being transmitted from mosquitoes into the human body, *P. falciparum* experiences the rest of its life cycle in peripheral bloodstream and liver. In the blood stage, they propagate asexually in red blood cells in the form of ring, trophozoite and schizont developmental stages in 48 h, resulting in daughter cells released in the peripheral bloodstream. The ART resistance of *P. falciparum* happens specifically at the ring stage, when the parasites lose their apical complex and de-differentiate into round immature trophozoites, pushing their nuclei to one side of the cell, making the cell morphologically resemble rings under the microscope (Dondorp et al., 2009; Suresh and Halder, 2018).

In the past years, the research field has been tirelessly searching for the genomic and transcriptomic traits associated with ART resistance (Ariey et al., 2014; Ashley et al., 2014; Cheeseman et al., 2012; Hunt et al., 2010; Mok et al., 2015; Takala-Harrison et al., 2013). For instance, it has been reported that a point mutation in the gene *ubp1* confers ART resistance in a *Plasmodium chabaudi* mouse malaria model (Hunt et al., 2010). This gene encodes a de-ubiquitinating enzyme, and the missense mutation reduces de-ubiquitinating activity and alters the associated protein degradation pathways (Hunt et al., 2007). In addition, multiple loci on chromosomes 10, 13, and 14 have been identified to be associated with the heritable trait of ART resistance (Cheeseman et al., 2012; Takala-Harrison et al., 2013). Particularly, mutations in the gene *kelch* PF3D7_1343700 ("K13-propeller") on chromosome 13 have been reported to be a significant molecular marker associated with ART resistance (Ariey et al., 2014; Ashley et al., 2014; Zhu et al., 2018). Beyond mutations, changes in expression of genes involved in the unfolded protein response (UPR) pathways have been linked to human ART resistance (Mok et al., 2015).

[1] Department of Computational Medicine and Bioinformatics, The University of Michigan Medical School, Ann Arbor, MI 48109, USA

[2] Department of Internal Medicine, The University of Michigan Medical School, Ann Arbor, MI 48109, USA

[3] Lead contact

*Correspondence: gyuanfan@umich.edu

https://doi.org/10.1016/j.isci.2022.103910

Although many studies have focused on the relationship between individual gene mutation and expression and drug resistance in malaria, a systematic evaluation of the value of these biomarkers in clinical or pre-clinical applications remains in need. The recent Malaria DREAM challenge, which blindly evaluated algorithms for predicting ART resistance, addressed this need (Bionetworks, n.d.a). The Malaria DREAM challenge leveraged an important dataset previously published (Mok et al., 2015), in which transcriptome profiles of *P. falciparum* isolates from 1,043 patients were measured *in vivo* without treatment and the resistance status was reported. The participants of the challenge were asked to predict the *in vitro* drug response of independent isolates with expression data obtained before and after perturbations with dihydroartemisinin (DHA).

We are presenting here the top performing algorithm ranked by accuracy to the above-described question, a machine learning model for predicting ART resistance based on the transcriptomic profile of the parasite. This model addresses several key challenges in malaria genomics and drug research: how to build models that can deliver across *in vivo* and *in vitro* datasets? Most of the *P. falciparum* experiments are cultured with human blood and carried out *in vitro*, whereas clinical applications require the model to be robust for *in vivo* dataset. How to make models deliverable from one measurement platform to another and thus allow wide application and generalization of the models? Of note, the training dataset of the DREAM challenge comes from a customized, two-color expression panel, whereas the test dataset came from one-color Agilent HD Exon Array with much more probes for each gene. How to identify the biomarkers and create the minimal panel of genes that both reveal the biological insights/pathways related to ART resistance and are capable of making good predictions? We address the above challenges by developing a cross-platform, *in vivo-in vitro* generalizable model for ART resistance prediction and analyzing independent contributions of gene expression signatures. We identified four molecular signatures important to the model: PF3D7_0523000 (pfmdr1), PF3D7_1245300, PF3D7_1372000, and PF3D7_0805000, creating a panel that almost matched the entire transcriptome in performance when predicting the cross-*in vivo-in vitro* drug resistance. Examination of co-expression modules reveals stable co-regulation modules of the top molecular features related to ART resistance.

RESULTS

Study design to investigate the transferability of models for *in vivo-in vitro* and cross-platform generalization

The overall study design intends to construct a model that is transferable across microarray platforms and across *in vivo-in vitro* conditions. The training dataset comes from Mok et al., which is a large cohort (1,043 isolates) of transcriptomic data of *P. falciparum* collected from southeast Asia during 2012–2014 (Mok et al., 2015). The parasite samples were directly taken from the peripheral blood of patients with acute falciparum malaria. The customized, printed expression panel measured 4,978 genes out of ~5,591 genes of the *P. falciparum* genome. ART-resistance phenotype was identified by the rate of clearance of parasites in the patient's peripheral blood, which is quantified by the clearance half-life upon ACT treatment. In this study, the samples with clearance half-life >5 h are considered as ART-resistant and labeled with "Slow" clearance rate. One the other hand, the samples with ≤5 h of clearance half-life are labeled as "Fast" in terms of clearance rate and considered as non-ART-resistant samples (Figure 1).

This study, as shown below, starts with cross-validation with the above-described dataset. In addition, the design of the test set differs from the training set in its sampling geographic site and timing of sample collection, synchronization status, microarray platform, and measurement target, introducing new challenges to the prediction models. The *in vitro* test set consisted of unpublished data of 32 isolates collected from the Thai-Mayanmar border (Figure 1A). The isolates are synchronized *in vitro*. Each isolate was examined twice, once without treatment and once with ART (DHA) treatment. The expression level was taken separately at 6 and 24 h postinfection (hpi). The test data were measured using Agilent HD Exon Array with much more probes (on average 12/gene) than the printed array in the training data (on average 2/gene) (Figures 1B and 1C). This test set was the test set for sub-challenge 1 of the Malaria DREAM challenge in which the task was to predict ART IC50, given a training set consisting of transcriptomes of parasites with known IC50. In addition, the training data used a two-color array and the test set used the Bozdech one-color array, which is expected to introduce challenges in data analytics (Patterson et al., 2006). Due to the differences in the array platforms, the methods used to preprocess the arrays also differ (see STAR Methods). The test set panel included noncoding RNAs, which are excluded in the training set. This results in a total of 5,540 genes in the test set data. For the test set, a continuous value of IC50 upon ART treatment is given as the testing target. The direct test data on this challenge remain a hidden

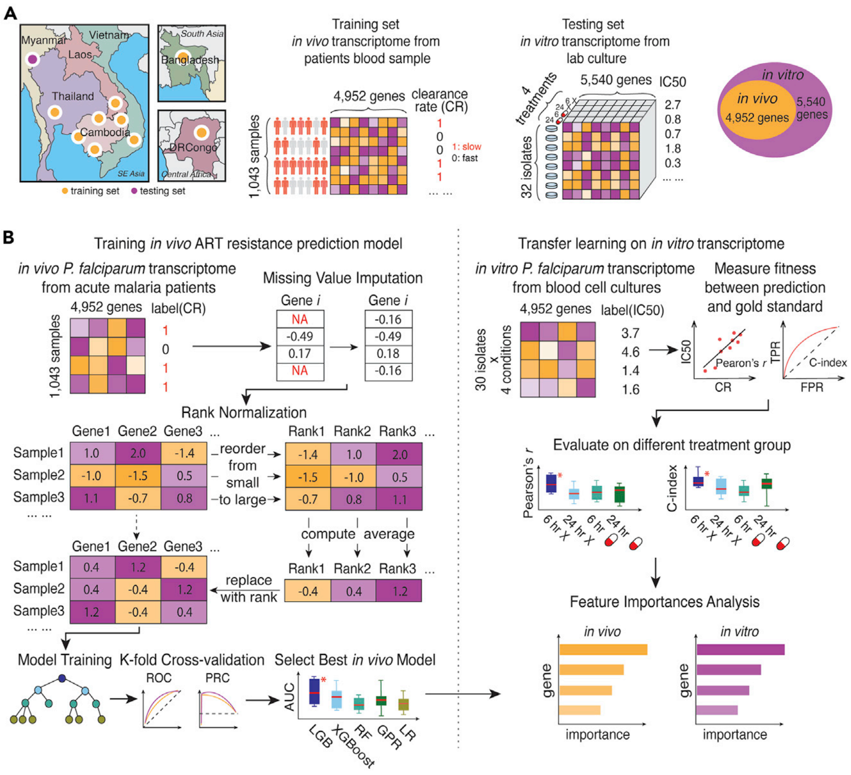


Figure 1. Study design

(A) Demonstration of the training data given by the DREAM Challenge.

(B) Strategy of training *in vivo* malaria ART prediction models and transferring the model to *in vitro* malaria transcriptome datasets. First, we imputed missing values and rank-normalized the expression data. Second, we cross-validated models of different base learners. We then selected the base learner and sample conditions with best performance by cross-validations and reverse test. Lastly, important predictive biomarkers are prioritized by SHAP analysis.

set for future model refinement by the scientific community. However, an independent test set of 30 isolates collected in the exact manner and cohort was available through sub-challenge 1 of this challenge (Bionetworks, n.d.a), which are used as the test set to evaluate model transferability in this study. Besides the DREAM challenge dataset, we also collect four independent public *P. falciparum* transcriptome datasets, of which two were sampled *ex vivo* and two *in vitro*, to further validate the robustness of transferability of the cross-platform model in this study. All transcriptomes used in this study were analyzed by t-SNE to show the differences between ART-resistance/sensitive samples, sampled condition (*in vivo*, *in vitro* or *ex vivo*), independent studies, and treatment type (Figure S4).

Excellent performance for within cohort prediction of ART clearance rate

The large collection of the Mok et al. data allows us to evaluate the models by two approaches. First, we can evaluate the model performance by cross-validation within the 1,043 isolates. Cross-validation is a commonly used scheme to evaluate model performance by holding out part of the data as the testing set and using the other part as the training set. Second, we can evaluate the model performance by training a model on the Mok et al. data and test on the *in vitro* data as described earlier. In this section, we describe the behavior of the model in the within-cohort cross-validation using the Mok et al. data. Clearance half-life was labeled ''fast'' or ''slow'' according to whether the parasite clearance half-life is longer than 5 h. We labeled ''slow'' as 1 and ''fast'' as 0 in the following experiments.

We carried out 10-fold cross validation by including all genes as features (Figure 2). Specifically, in each round, 10% of the isolates were held out as the test set, and 90% were used as the training set. We tested a selection of base learners, including LightGBM, xgboost, random forest, Gaussian Process Regression (GPR), and linear regression (see STAR Methods). Because an important goal of this study is to develop

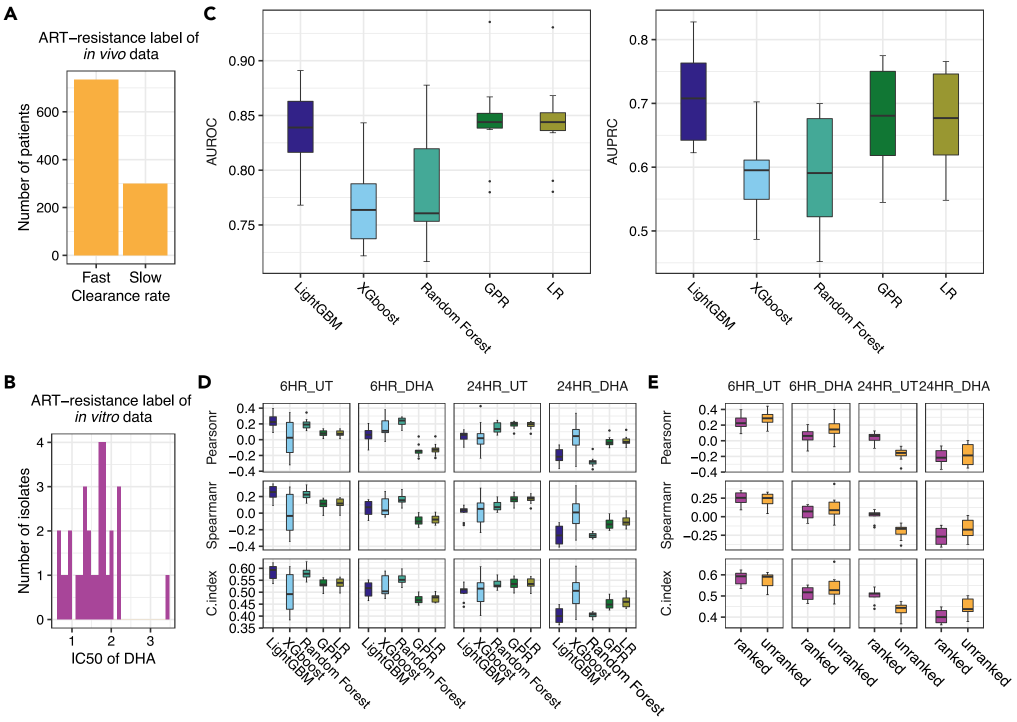


Figure 2. Model performances across platforms

(A and B) Distribution of ART resistance measurement labels in the *in vivo* and the *in vitro* datasets.

(C) Cross-validation performance in the *in vivo* dataset.

(D) Performance of transferring the model trained on *in vivo* dataset to the *in vitro* dataset, presented as the correlation between prediction and gold standard (IC50) under four conditions.

(E) Performance of transfer learning with/without rank normalization.

a model transferable to transcriptome data collected using different platforms, which can be of drastically different distribution, we also tested if rank normalization of the expression data changed performance.

LightGBM, a tree-based gradient boosting method, marginally excelled in performance for both Area Under the Receiver Operating Curve AUROC and AUPRC measurements (Figure 2C) compared with other alternatives. It achieved a mean AUROC [95% CI] of 0.8384 [0.8121, 0.8705], compared with XGboost (0.7669 [0.7262, 0.7910]), random forest (0.7782 [0.7441, 0.8099]), GPR (0.8456 [0.8212, 0.8673]), and linear regression (0.8448 [0.8206, 0.8668]). For AUPRC [95% CI], LightGBM performed at 0.6983 [0.6438, 0.7522], compared with XGboost (0.6613 [0.5994, 0.7234]), random forest (0.5752 [0.5049, 0.6387]), GPR (0.6742 [0.6198, 0.7280]), and linear regression (0.6717 [0.6176, 0.7252]). Rank normalization does not present substantial changes in performance (Figure 2E and Table S1); we chose to maintain this operation to support cross-platform robustness.

Transferring models across platforms

The test data differ from the above-examined *in vivo* data in that it was collected from laboratory cultured *P. falciparum* strains. This allows synchronization, and thus the gene expression levels were sampled under four different conditions: (1) 6 h postinvasion (hpi), (2) 24 hpi, (3) 6 hpi and treated with dihydroartemisinin (DHA) (6 hpi-p), and (4) 24 hpi and treated with DHA (24 hpi-p). We evaluated the models based on different base learners as described earlier for each of the expression data. Because the test target is IC50, we labeled ''slow'' as 1 and ''fast'' as 0 in our training.

As expected, 6 hpi without treatment demonstrated the strongest performance, as the original training data were pretreatment as well (Figure 3D). In addition, LightGBM maintains to be the strongest base learner. In this case, rank normalization does not change the performance substantially, so we retained it in the preprocessing steps (Figure 3E and Table S2). This combination achieved a Pearson correlation [95% CI] of 0.2318

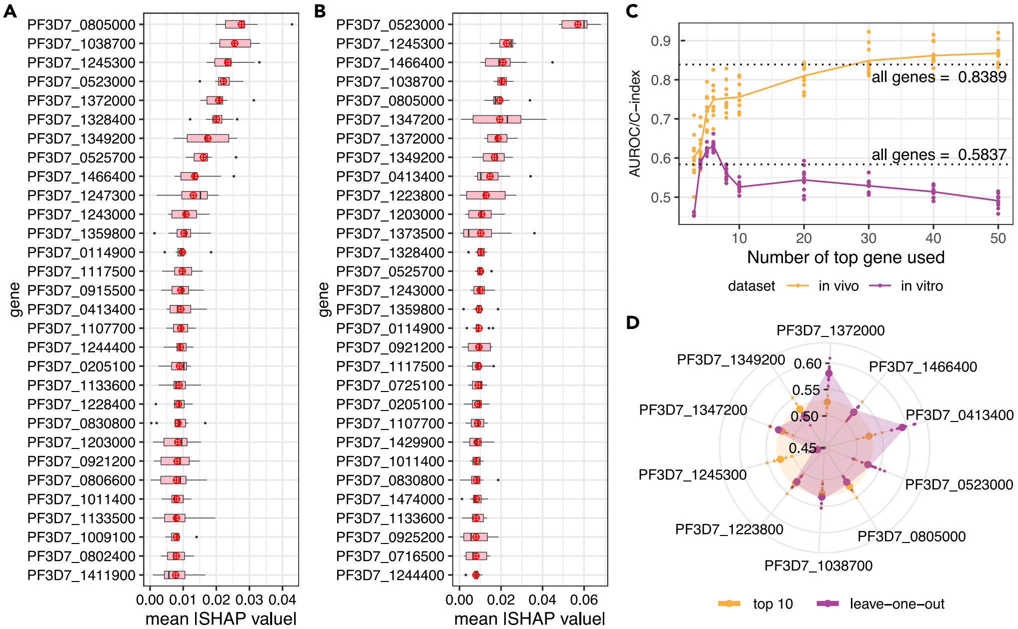


Figure 3. Top genes related to malaria ART resistance as identified by SHAP feature importance analysis and performances of machine learning model after feature selection using the top ranked genes

(A) Top 30 genes of ART resistance prediction model visualized by SHAP analysis based on *in vivo P. falciparum* transcriptome.

(B) Top 30 genes of ART resistance prediction model visualized by SHAP analysis based on *in vitro P. falciparum* transcriptome. Mean SHAP values in both A and B were denoted by red dots. Genes were ordered by mean SHAP contributions across all test examples in a 10-fold cross validation. 25% and 75% percentile of SHAP values were denoted by low and upper limits of boxplots.

(C) Model performances for *in vivo* and *in vitro* predictions when including only top genes selected by SHAP analysis, as evaluated by AUROC (for binary labels) and C-index (for continuous labels). "All genes" shows prediction performance without feature selection.

(D) Comparison of *in vitro* prediction performances between using all top ten genes as features and leaving one gene out at a time. Each dot denotes the performce of each of the ten models in cross validation, and the average perfornences of all ten models were denoted by larger dots.

[0.1379, 0.5306], Spearman's correlation of 0.2467 [0.1457, 0.3548], and a C-index of 0.5837 [0.5474, 0.6216] between the predicted clearance rate and IC50. Of note, the gold standard used in training is nongranular values but rather a binary value of "fast" and "slow." Yet, we still received meaningful predictions using a different microarray platform and data collection status ($p < 1e-6$) compared with random prediction.

We further evaluate the best-performing *in vivo* LightGBM model to four other public datasets for ART-resistance prediction, where the ART resistance for each sample were available (Table S3) (Mok et al., 2011, 2015, 2021; Shaw et al., 2015; Zhu et al., 2018), and results were shown in Figure S5. We noticed that on *ev vivo* data, the model achieved better cross-platform accuracy than *in vitro* data overall. The *in vivo* model achieved 0.75 [0.6431, 0.9773] and 0.6894 [0.6065,0.8060] AUROC [95% CI)] on the GEO: GSE25878 and GEO: GSE59098 dataset, respectively. While on the *in vitro* dataset GEO: GSE151189, the model only achieved 0.5355 [0.4530, 0.6416] overall AUROC [95% CI]. One possible reason could be the *ex vivo* transcriptomes show more similarity to the *in vivo* data the model was trained on (Figures S4B and S4C). Interestingly, we also noticed the model prediction heavily relies on the *ex vivo* cultured time, treatment by DHA, and developmental stages (hpi), indicating these factors may change the expression levels of effector genes related to ART resistance.

Robustness in molecular features across *in vivo* and *in vitro* environment

It was very encouraging that a model can be developed and carried across such different *in vivo* and *in vitro* scenarios, and across experimental platforms, which prompted us to examine the top molecular features that contributed to this prediction. We first used SHapley Additive exPlanations analysis (SHAP) to find out which genes played important roles in the *in vivo* ART-resistance prediction (Lundberg and Lee, 2017). SHAP analysis

Table 1. Twenty-two shared features (among top 30) between the *in vivo* and the *in vitro* datasets

Gene id	*In vivo* SHAP importance	*In vitro* SHAP importance	Annotations
PF3D7_0805000	0.027577556	0.019392157	Alpha/beta hydrolase, putative
PF3D7_1038700	0.025549987	0.020552675	Plasmodium exported protein, unknown function
PF3D7_1245300	0.023449244	0.022728070	NEDD8-conjugating enzyme UBC12, putative
PF3D7_0523000	0.022131747	0.056972396	Multidrug resistance protein 1
PF3D7_1372000	0.020715635	0.018587311	Plasmodium exported protein (PHISTa), unknown function
PF3D7_1328400	0.020094142	0.010121155	Conserved protein, unknown function
PF3D7_1349200	0.017402912	0.016821882	Glutamate–tRNA ligase, putative
PF3D7_0525700	0.016087702	0.010102856	Conserved protein, unknown function
PF3D7_1466400	0.013496112	0.020929573	AP2 domain transcription factor AP2-EXP
PF3D7_1243000	0.01083543	0.009874289	Syntaxin-16, putative
PF3D7_1359800	0.010260771	0.009438231	ADP-ribosylation factor, putative
PF3D7_0114900	0.00976088	0.009389257	Plasmodium exported protein, unknown function, pseudogene
PF3D7_1117500	0.009725304	0.009184379	Tyrosine–tRNA ligase
PF3D7_0413400	0.009276499	0.014652890	Erythrocyte membrane protein 1 (PfEMP1), exon 1, pseudogene
PF3D7_1107700	0.009240973	0.008681170	Pescadillo homolog
PF3D7_1244400	0.009191554	0.007895322	RNA-binding protein, putative
PF3D7_0205100	0.008908881	0.008790011	Conserved Plasmodium protein, unknown function
PF3D7_1133600	0.0087618	0.008062268	Conserved Plasmodium protein, unknown function
PF3D7_0830800	0.00846947	0.008136531	Surface-associated interspersed protein 8.2 (SURFIN 8.2)
PF3D7_1203000	0.008352643	0.010673660	Origin recognition complex subunit 1
PF3D7_0921200	0.008144911	0.009305749	Conserved Plasmodium membrane protein, unknown function
PF3D7_1011400	0.007971128	0.008214462	Proteasome subunit beta type-5

is a feature importance analysis method that recently gained popularity, in which the importance of one feature is considered in the context of all other features. This approach has the advantage to delineate gene features that are important for predicting ART resistance versus the ones that happened to be correlated with an important feature. Table 1 shows the top genes during the 10-fold cross-validation. Among them, there were five genes recognized by all ten models, showing consistent importance (Figure S2). The SHAP analysis is test set dependent. This unique feature allows us to test the robustness of these features further in the *in vitro* data. We found the same set of top genes still showed significant contribution in *in vitro* prediction (Figures 3A, 2B, and S3). Of note, about ~70% top genes (4 out of top 5, 7 out of top 10, 14 out of top 20, and 22 out of top 30) were found to be shared by both *in vivo* and *in vitro* datasets, showing coherence in top-ranked features across platforms (Figure 5A). Pfmdr1 is among the most significant contributors in both *in vitro* prediction and *in vivo* prediction. This result supports the robustness of the identified molecular features.

We further investigated the functions of top contributing genes considering both *in vivo* and *in vitro* predictions of ART resistance in malaria (Figure 4A; Table 1). Among them, pfmdr1 (PF3D7_0523000), *P. falciparum* multidrug drug resistance gene 1, has been reported to play an essential role in response to a broad range of ACT antimalarials (Gil and Krishna, 2017; Koenderink et al., 2010; Sidhu et al., 2006). Mutants and polymorphisms of this protein have been widely reported to be associated with antimalarial drug resistance, and the increase of pfmdr1's expression will increase susceptibility to ART (Chavchich et al., 2010; Dahlström et al., 2009; Eastman et al., 2016; Gupta et al., 2014; Holmgren et al., 2006, 2007; Imwong et al., 2010; Ngalah et al., 2015; Ould Ahmedou Salem et al., 2017; Sidhu et al., 2006; Sisowath et al., 2007; Ursing et al., 2006). The identification of this gene among the top of the list and its positive contribution to both IC50 and clearance rate corroborates the validity of the approach (Figures S2 and S3).

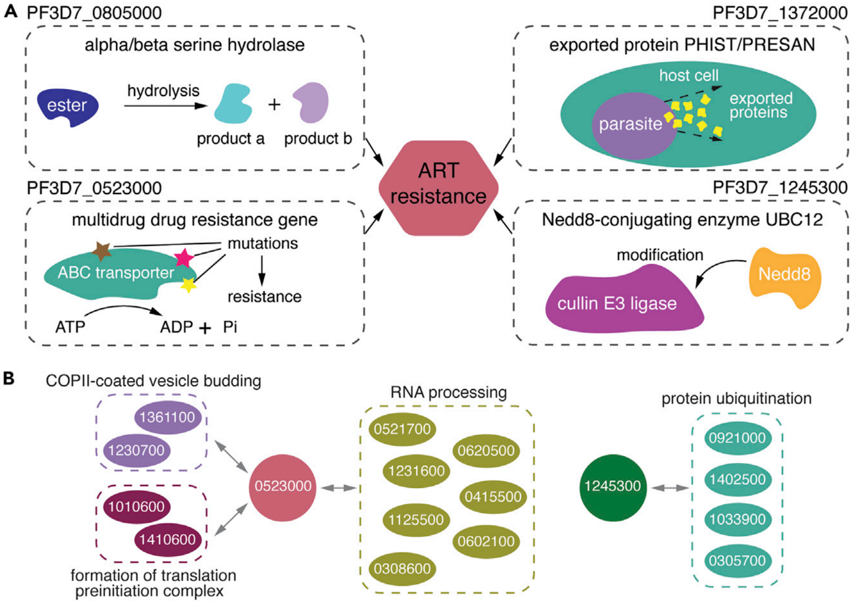


Figure 4. Cellular functions of top contributing genes in predicting ART resistance

(A) The functions of top contributing genes and their relationship with ART resistance.

(B) The functionally enriched protein clusters that interact with PF3D7_052300 and PF3D7_1245300. The prefix "PF3D7_" of these gene ids is omitted, and only the numbers are shown for simplicity.

We found other interesting genes in this list. First, PF3D7_1372000 is a *Plasmodium* exported protein of the Poly-Helical Interspersed Sub-Telomeric (PHIST) protein family (Tarr et al., 2014; Warncke et al., 2016), also known as the PRESAN family (Oakley et al., 2007; Sargeant et al., 2006). Although detailed functions of most Plasmodium exported proteins are yet to be revealed, in general, the parasite exported proteins are pivotal for parasite survival by interacting and interfering activities of the infected cells (Maier et al., 2008). A recent study has suggested that the expression level of PF3D7_1372000 is associated with mutations of *kelch* PF3D7_1343700 ("K13-propeller") (Siddiqui et al., 2020), whose mutations have been reported to be a significant molecular marker associated with ART resistance (Ariey et al., 2014; Zhu et al., 2018). Second, PF3D7_1245300 is a Nedd8-conjugating enzyme UBC12, which has a central role in cell cycle and DNA damage repair (Karpiyevich et al., 2019). Because the malaria parasite has a unique and unusual life cycle, the molecular machines in cell replication processes are specially designed for its survival. As Plasmodium responds to ART-induced stress by delaying their cell-cycle progression and inducing a state of dormancy during early ring-stage development (van Biljon et al., 2018), it is likely UBC12 presents as an important feature through this mechanism. Leave-one-out feature selection strategy based on the top ten genes shows that taking PF3D7_1245300 away will undermine *in vitro* prediction performance (Figure 3D), indicating this gene is crucial for *P. falciparum*'s survival in both laboratory environments and in the human body. Two other genes, PF3D7_0805000, a putative member of the alpha/beta serine hydrolase superfamily that mediates a variety of metabolic reactions of ester hydrolysis, and PF3D7_1038700, another Plasmodium exported protein with unknown function, appeared in the top list. The association between these two genes with ART resistance is currently unknown.

We further investigated other proteins related to these top contributing genes based on the protein-protein interactome generated from blue native-polyacrylamide electrophoresis with quantitative mass spectrometry (Hillier et al., 2019). We first extracted interacting proteins with pfmdr1 and PF3D7_1245300 and found 20 and 37 interacting proteins, respectively. The other three proteins of the top genes were not observed in the interactome. Then we performed GO functional enrichment analysis of these proteins and identified the significantly enriched protein clusters with FDR p value cutoff of 0.05 (Figure 4B). For the multidrug resistance gene pfmdr1, the interacting proteins are associated with RNA processing, COPII-coated vesicle budding, and formation of translation preinitiation complex. For the Nedd8-conjugating enzyme UBC12 (PF3D7_1245300), as expected, the interacting proteins are associated with protein

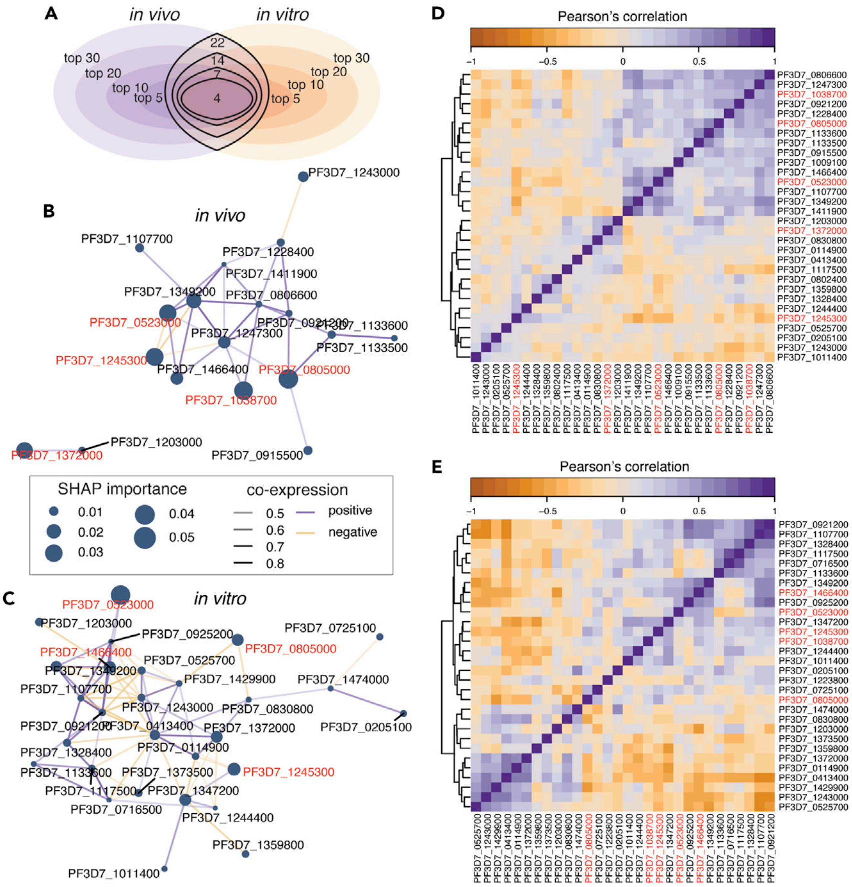


Figure 5. Co-expression networks of top genes in *in vivo* and *in vitro* dataset

(A) Sharing of top genes across *in vivo*/*in vitro* datasets.

(B and D) The co-expression network and the co-expression matrix of the top 30 genes in the *in vivo* dataset.

(C and E) The co-expression network and co-expression matrix of the top 30 genes in the *in vitro* dataset. We retained all co-expression relationships with an absolute correlation value >0.4 in the plot. The five most important genes in either *in vivo* or *in vitro* dataset were marked as red (except PF3D4_1038700, which was not shown in (C), as there were no other genes that shared significant correlation (|r| > 0.4) with this gene).

ubiquitination, a process previously found to be important for treatment resistance in malaria (Dogovski et al., 2015; Tilley et al., 2016).

We went on to construct models only based on the top genes identified by SHAP analysis (Figure 3C). We found that for within *in vivo* cross-validation, 30 genes can completely recover the performance of the model using the entire transcriptome. In addition, the top genes identified in the above analysis successfully reaches the performance of the entire gene panel when delivering the model to the *in vitro* test set. We acknowledge the existence of fluctuation in performance after the sixth top genes. The likely reason is that SHAP identifies independent features, and as we increase the number of features beyond six, the ones that are comparably weaker yet orthogonal to the top features are included. Despite this limitation, this result supports the validity of the top features we identified in this study as potential biomarkers for ART resistance.

Although kelch13 genetic mutations has been found to be significantly correlated with ART resistance phenotype in the *in vivo* population study (pearson'|'s r = 0.6143, p < 1e-6), no significant correlation of kelch13 transcription with ATR-resistance phenotype has been found (Mok et al., 2015). This result is concordant with our SHAP analysis results, as kelch13 transcription level turned out with no contribution to ART resistance prediction. Machine learning models with feature sets excluding kelch13 transcription level still maintained similar performances (Tables S1 and S3). We also evaluated the ART-resistance model performances in different genetic variation cohorts, including K13 KP/BTB mutations, crt-N326S, crt-I356T, fd-D193Y, and mdr2-T484I (Table S4 and Figure S1). The ART-resistance model is still quite predictive within K13 subgroups, with mutations (group 2) and heterozygous alleles (group 3) (Figure S1).

Conserved co-expression patterns of top-ranking features

We next examined if the top ranking features in the *in vivo* test and in the *in vitro* test share similar expression patterns or regulatory modules. We took the top 30 features for each and calculated the Pearson correlation of expression values across all samples separately for the *in vivo* and *in vitro* datasets. This step created co-expression networks (Figure 5). Among the top 30 genes, 22 are shared between *in vivo* and *in vitro* tests, a piece of supporting evidence to the robustness of the features (Table 1 and Figure 5A).

We then examined if the co-expression networks of the top features share similarity between the *in vivo* and *in vitro* datasets. We identified many co-expression relationships maintained across the *in vivo* and *in vitro* datasets. For example, the correlation between PF3D7_0523000 and PF3D7_1466400 is 0.46 ($p < 2.2e-16$, the smallest value storable in the computer) in *in vivo* dataset and 0.42 ($p < 2.2e-16$) in *in vitro* dataset. Therefore, we calculated the correlation values of the network weights (*i.e.*, correlation between genes) for the 22 shared genes. The correlation is 0.55 ($p < 2.2e-16$), indicating strong and conserved co-expression modules involved in ART resistance.

DISCUSSION

In this study, we presented a model that is transferable between *in vivo* measured clearance rate and *in vitro* measured IC50 for ART in malaria treatment and across expression measurement platforms. This is a meaningful step in the research of malaria treatment, as the work demonstrated the potential and robustness of a personalized model for ART resistance, which has not been achieved before. Some studies addressed the prediction on either *in vivo* or *in vitro* study but did not generalize the model across different conditions (Ford and Janies, 2020; Li et al., 2021; Sastry et al., 2021). In fact, previous studies reported that generating predictive models for ART resistance has been challenging, as the *in vitro* IC50 of *P. falciparum* in standard drug susceptibility assay correlates poorly with its clearance rate *in vivo* (Chotivanich et al., 2014; Fairhurst and Dondorp, 2016). Thus, the ability of this model to deliver across drastically different scenarios makes this model favorable.

Delivering models between platforms and *in vivo-in vitro* environments has always been a challenge for many medical problems. Several techniques developed in this study may be instructive to other problems. For example, rank normalization of the shared genes in the transcriptomic profiles can potentially help to match two different sets of data and address batch effects. Tree-based algorithms may help interrogate the interactions and overlaps between genes and construct robust models.

We discovered important biomarkers that can be used to create a simplified model for predicting ART resistance. Among them, interesting molecular biomarkers were identified. Pfmdr1 (PF3D7_0523000), *P. falciparum* multidrug drug resistance gene 1, was identified among the shared top genes by both *in vivo* and *in vitro* datasets, consistent with previous reports stating that it plays an essential role in the response processes of a broad range of ACT antimalarials (Chavchich et al., 2010; Dahlström et al., 2009; Eastman et al., 2016; Gupta et al., 2014; Holmgren et al., 2006, 2007; Imwong et al., 2010; Ngalah et al., 2015; Ould Ahmedou Salem et al., 2017; Sidhu et al., 2006; Sisowath et al., 2007; Ursing et al., 2006). PF3D7_1372000, a *Plasmodium* exported protein of the Poly-Helical Interspersed Sub-Telomeric (PHIST) protein family (Tarr et al., 2014; Warncke et al., 2016), was also identified among the shared top genes. Literature has reported that the parasite exported proteins are pivotal for parasite survival by interacting and interfering activities of the infected cells (Maier et al., 2008). In addition, UBC12, which plays a central role in cell cycle and DNA damage repair (Karpiyevich et al., 2019), was identified, possibly reflecting the mechanism that Plasmodium responds to ART-induced stress by delaying their cell-cycle progression and inducing a state of dormancy during early ring-stage development (van Biljon et al., 2018). Other important features whose molecular mechanisms are yet unclear were also identified, pointing to future studies that follow-up and validate these new molecular markers for ART resistance.

Limitations of the study

Although our model has achieved satisfying performances on the same population study, we noticed that during the cross-platform prediction, the performance has been impacted severely by the condition of samples in the target datasets, i.e. *in vivo*, *ex vivo*, or *in vitro*, whether treated by DHA, developmental stage (hpi). These observations imply that genes related to ART resistance expressed differently under different conditions. Although many studies have addressed the dependency between ART resistance with developmental stages (Intharabut et al., 2019; Mok et al., 2011, 2015), *in vitro* environments may also impact the ART-resistance phenotype, which needs more experimental assessments in the future.

Furthermore, although top genes were identified in this study, further experimental evidence is still needed to elucidate their roles in ART resistance. For further verification of these biomarkers, gene function perturbations could be carried out on the ART-resistant strains in both *in vivo* and *in vitro* conditions. For example, translation and ubiquitin-activating enzyme inhibitors were found to antagonize the activity of DHA *in vivo* and *in vitro* on Plasmodium falciparum strains (Bridgford et al., 2018). Moreover, atovaquone, a mitochondrial electron transport chain inhibitor, could reverse the ART resistance in Cambodian Cam3.II line *in vitro* (Mok et al., 2021). Instead of broad inhibitors that deactivate certain pathways, more targeted gene silencing methods, such as RNAi or CRISPR, would be recommended to inhibit certain top biomarkers, to elucidate the mechanisms of ART resistance.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact and materials availability
  - Data and code availability
- METHOD DETAILS
  - Data pre-processing
  - Model training
  - SHAP feature importance analysis
  - Co-expression and functional analysis of top genes
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Evaluation of prediction performances
  - C-index is equivalent to AUROC when predicting binary labels.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.103910.

ACKNOWLEDGMENTS

This work is supported by NIH R35GM133346 and NSF#1452656. The Datasets used for the analyses described in this manuscript were contributed by Michael T. Ferdig, Department of Biological Sciences at the University of Notre Dame. They were obtained as part of the Malaria DREAM Challenge through Synapse (syn16924919), managed by Sage Bionetworks. Per challenge rules, the manuscript has been edited and approved by the Challenge organizers (Dr. Michael T. Ferdig and Dr. Geoffrey Siwo).

AUTHOR CONTRIBUTIONS

YG designed and implemented the algorithm. HZ, HL, and JG carried out post-challenge analysis. YG wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

YG serves as commissioning editor of iScience and receives compensation.

Received: October 26, 2021
Revised: January 26, 2022
Accepted: February 8, 2022
Published: March 18, 2022

REFERENCES

Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.-C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L., et al. (2014). A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature *505*, 50–55.

Asenso-Okyere, K., Asante, F.A., Tarekegn, J., and Andam, K.S. (2011). A review of the economic impact of malaria in agricultural development. Agric. Econ. *42*, 293–304.

Ashley, E.A., Dhorda, M., Fairhurst, R.M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J.M., Mao, S., Sam, B., et al.; Tracking resistance to artemisinin Collaboration (TRAC) (2014). Spread of artemisinin resistance in Plasmodium falciparum malaria. N. Engl. J. Med. *371*, 411–423.

Bionetworks, S., n.d.a Synapse [WWW Document]. URL https://www.synapse.org/#!Synapse:syn16924919/wiki/(accessed 8.18.20a)

Bridgford, J.L., Xie, S.C., Cobbold, S.A., Pasaje, C.F.A., Herrmann, S., Yang, T., Gillett, D.L., Dick, L.R., Ralph, S.A., Dogovski, C., et al. (2018). Artemisinin kills malaria parasites by damaging proteins and inhibiting the proteasome. Nat. Commun. 3801. https://doi.org/10.1038/s41467-018-06221-1.

Chavchich, M., Gerena, L., Peters, J., Chen, N., Cheng, Q., and Kyle, D.E. (2010). Role of pfmdr1 amplification and expression in induction of resistance to artemisinin derivatives in Plasmodium falciparum. Antimicrob. Agents Chemother. *54*, 2455–2464.

Cheeseman, I.H., Miller, B.A., Nair, S., Nkhoma, S., Tan, A., Tan, J.C., Al Saai, S., Phyo, A.P., Moo, C.L., Lwin, K.M., et al. (2012). A major genome region underlying artemisinin resistance in malaria. Science *336*, 79–82.

Chotivanich, K., Tripura, R., Das, D., Yi, P., Day, N.P.J., Pukrittayakamee, S., Chuor, C.M., Socheat, D., Dondorp, A.M., and White, N.J. (2014). Laboratory detection of artemisinin-resistant plasmodium falciparum. Antimicrob. Agents Chemother. *58*, 3157–3161. https://doi.org/10.1128/aac.01924-13.

Conn, J.E., Grillet, M.E., Correa, M., and Sallum, M.A.M. (2018). Malaria transmission in south America—present status and prospects for elimination. Towards Malar. Elimination - A Leap Forward. https://doi.org/10.5772/intechopen.76964.

Cowman, A.F., Healer, J., Marapana, D., and Marsh, K. (2016). Malaria: biology and disease. Cell *167*, 610–624.

Dahlström, S., Ferreira, P.E., Veiga, M.I., Sedighi, N., Wiklund, L., Mårtensson, A., Färnert, A., Sisowath, C., Osório, L., Darban, H., et al. (2009). Plasmodium falciparum multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. J. Infect. Dis. *200*, 1456–1464.

Dhiman, S. (2019). Are malaria elimination efforts on right track? An analysis of gains achieved and challenges ahead. Infect. Dis. Poverty *8*, 14.

Dogovski, C., Xie, S.C., Burgio, G., Bridgford, J., Mok, S., McCaw, J.M., Chotivanich, K., Kenny, S., Gnädig, N., Straimer, J., et al. (2015). Targeting the cell stress response of Plasmodium falciparum to overcome artemisinin resistance. PLoS Biol. *13*, e1002132.

Dondorp, A.M., Nosten, F., Yi, P., Das, D., Phyo, A.P., Tarning, J., Lwin, K.M., Ariey, F., Hanpithakpong, W., Lee, S.J., et al. (2009). Artemisinin resistance in Plasmodium falciparum malaria. N. Engl. J. Med. *361*, 455–467.

Eastman, R.T., Khine, P., Huang, R., Thomas, C.J., and Su, X.-Z. (2016). PfCRT and PfMDR1 modulate interactions of artemisinin derivatives and ion channel blockers. Sci. Rep. *6*, 25379.

Fact Sheet about Malaria [WWW Document], n.d. URL https://www.who.int/news-room/fact-sheets/detail/malaria (accessed 2.10.20)

Fairhurst, R.M., and Dondorp, A.M. (2016). Artemisinin-resistant plasmodium falciparum malaria. Microbiol. Spectr. *4*. https://doi.org/10.1128/microbiolspec.EI10-0013-2016.

Ford, C.T., and Janies, D. (2020). Ensemble machine learning modeling for the prediction of artemisinin resistance in malaria. F1000Research *9*, 62. https://doi.org/10.12688/f1000research.21539.1.

Gil, J.P., and Krishna, S. (2017). pfmdr1 (Plasmodium falciparum multidrug drug resistance gene 1): a pivotal factor in malaria resistance to artemisinin combination therapies. Expert Rev. Anti. Infect. Ther. *15*, 527–543.

Gupta, B., Xu, S., Wang, Z., Sun, L., Miao, J., Cui, L., and Yang, Z. (2014). Plasmodium falciparum multidrug resistance protein 1 (pfmrp1) gene and its association with in vitro drug susceptibility of parasite isolates from north-east Myanmar. J. Antimicrob. Chemother. *69*, 2110–2117.

Hillier, C., Pardo, M., Yu, L., Bushell, E., Sanderson, T., Metcalf, T., Herd, C., Anar, B., Rayner, J.C., Billker, O., and Choudhary, J.S. (2019). Landscape of the plasmodium interactome reveals both conserved and species-specific functionality. Cell Rep. *28*, 1635–1647.e5.

Holmgren, G., Gil, J.P., Ferreira, P.M., Veiga, M.I., Obonyo, C.O., and Björkman, A. (2006). Amodiaquine resistant Plasmodium falciparum malaria in vivo is associated with selection of pfcrt 76T and pfmdr1 86Y. Infect. Genet. Evol. *6*, 309–314.

Holmgren, G., Hamrin, J., Svärd, J., Mårtensson, A., Gil, J.P., and Björkman, A. (2007). Selection of pfmdr1 mutations after amodiaquine monotherapy and amodiaquine plus artemisinin combination therapy in East Africa. Infect. Genet. Evol. *7*, 562–569.

Hunt, P., Afonso, A., Creasey, A., Culleton, R., Sidhu, A.B.S., Logan, J., Valderramos, S.G., McNae, I., Cheesman, S., do Rosario, V., et al. (2007). Gene encoding a deubiquitinating enzyme is mutated in artesunate- and chloroquine-resistant rodent malaria parasites. Mol. Microbiol. *65*, 27–40.

Hunt, P., Martinelli, A., Modrzynska, K., Borges, S., Creasey, A., Rodrigues, L., Beraldi, D., Loewe, L., Fawcett, R., Kumar, S., et al. (2010). Experimental evolution, genetic analysis and genome re-sequencing reveal the mutation conferring artemisinin resistance in an isogenic lineage of malaria parasites. BMC Genomics *11*, 499.

Imwong, M., Dondorp, A.M., Nosten, F., Yi, P., Mungthin, M., Hanchana, S., Das, D., Phyo, A.P., Lwin, K.M., Pukrittayakamee, S., et al. (2010). Exploring the contribution of candidate genes to artemisinin resistance in Plasmodium falciparum. Antimicrob. Agents Chemother. *54*, 2886–2892.

Intharabut, B., Kingston, H.W., Srinamon, K., Ashley, E.A., Imwong, M., Dhorda, M., Woodrow, C., Stepniewska, K., Silamut, K., Day, N.P.J., et al.; Tracking resistance to artemisinin Collaboration (2019). artemisinin resistance and stage dependency of parasite clearance in falciparum malaria. J. Infect. Dis. *219*, 1483–1489.

Karpiyevich, M., Adjalley, S., Mol, M., Ascher, D.B., Mason, B., van der Heden van Noort, G.J., Laman, H., Ovaa, H., Lee, M.C.S., and Artavanis-Tsakonas, K. (2019). Nedd8 hydrolysis by UCH proteases in Plasmodium parasites. PLoS Pathog. *15*, e1008086.

Koenderink, J.B., Kavishe, R.A., Rijpma, S.R., and Russel, F.G.M. (2010). The ABCs of multidrug resistance in malaria. Trends Parasitol. *26*, 440–446.

Li, D., Wang, Y., Hu, W., Chen, F., Zhao, J., Chen, X., and Han, L. (2021). Application of machine learning classifier to drug resistance analysis. Front. Cell Infect. Microbiol. *11*, 742062.

Lundberg, S., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (31st Conference on Neural Information Processing Systems) In press. https://arxiv.org/abs/1705.07874.

Maier, A.G., Rug, M., O'Neill, M.T., Brown, M., Chakravorty, S., Szestak, T., Chesson, J., Wu, Y., Hughes, K., Coppel, R.L., et al. (2008). Exported proteins required for virulence and rigidity of Plasmodium falciparum-infected human erythrocytes. Cell *134*, 48–61.

Mbacham, W.F., Ayong, L., Guewo-Fokeng, M., and Makoge, V. (2019). Current situation of malaria in Africa. Methods Mol. Biol. *2013*, 29–44.

Miller, L.H., and Su, X. (2011). Artemisinin: discovery from the Chinese herbal garden. Cell *146*, 855–858.

Mok, S., Ashley, E.A., Ferreira, P.E., Zhu, L., Lin, Z., Yeo, T., Chotivanich, K., Imwong, M., Pukrittayakamee, S., Dhorda, M., et al. (2015). Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. Science *347*, 431–435.

Mok, S., Imwong, M., Mackinnon, M.J., Sim, J., Ramadoss, R., Yi, P., Mayxay, M., Chotivanich, K., Liong, K.-Y., Russell, B., et al. (2011). Artemisinin resistance in Plasmodium falciparum is associated with an altered temporal pattern of transcription. BMC Genomics *12*, 391.

Mok, S., Stokes, B.H., Gnädig, N.F., Ross, L.S., Yeo, T., Amaratunga, C., Allman, E., Solyakov, L., Bottrill, A.R., Tripathi, J., et al. (2021). Artemisinin-resistant K13 mutations rewire Plasmodium falciparum's intra-erythrocytic metabolic program to enhance survival. Nat. Commun. *12*, 530.

Ngalah, B.S., Ingasia, L.A., Cheruiyot, A.C., Chebon, L.J., Juma, D.W., Muiruri, P., Onyango, I., Ogony, J., Yeda, R.A., Cheruiyot, J., et al. (2015). Analysis of major genome loci underlying artemisinin resistance and pfmdr1 copy number in pre- and post-ACTs in Western Kenya. Scientific Rep. *5*, 8308. https://doi.org/10.1038/srep08308.

Oakley, M.S.M., Kumar, S., Anantharaman, V., Zheng, H., Mahajan, B., Haynes, J.D., Moch, J.K., Fairhurst, R., McCutchan, T.F., and Aravind, L. (2007). Molecular factors and biochemical pathways induced by febrile temperature in intraerythrocytic Plasmodium falciparum parasites. Infect. Immun. *75*, 2012–2025.

Organization, W.H., Others, et al.. (2020). World Malaria Report 2020: 20 Years of Global Progress and Challenges.

Ould Ahmedou Salem, M.S., Mint Lekweiry, K., Bouchiba, H., Pascual, A., Pradines, B., Ould Mohamed Salem Boukhary, A., Briolant, S., Basco, L.K., and Bogreau, H. (2017). Characterization of Plasmodium falciparum genes associated with drug resistance in Hodh Elgharbi, a malaria hotspot near Malian-Mauritanian border. Malar. J. *16*, 140.

Patterson, T.A., Lobenhofer, E.K., Fulmer-Smentek, S.B., Collins, P.J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E.S., Hager, J., Tikhonova, I.R., et al. (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat. Biotechnol. *24*, 1140–1150.

Sachs, J., and Malaney, P. (2002). The economic and social burden of malaria. Nature *415*, 680–685.

Sargeant, T.J., Marti, M., Caler, E., Carlton, J.M., Simpson, K., Speed, T.P., and Cowman, A.F. (2006). Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. Genome Biol. *7*, R12.

Sastry, A.V., Dillon, N., Anand, A., Poudel, S., Hefner, Y., Xu, S., Szubin, R., Feist, A.M., Nizet, V., and Palsson, B. (2021). Machine learning of bacterial transcriptomes reveals responses underlying differential antibiotic susceptibility. mSphere *6*, e0044321.

Shaw, P.J., Chaotheing, S., Kaewprommal, P., Piriyapongsa, J., Wongsombat, C., Suwannakitti, N., Koonyosying, P., Uthaipibull, C., Yuthavong, Y., and Kamchonwongpaisan, S. (2015). Plasmodium parasites mount an arrest response to dihydroartemisinin, as revealed by whole transcriptome shotgun sequencing (RNA-seq) and microarray study. BMC Genomics *16*, 830.

Siddiqui, F.A., Boonhok, R., Cabrera, M., Mbenda, H.G.N., Wang, M., Min, H., Liang, X., Qin, J., Zhu, X., Miao, J., et al. (2020). Role of plasmodium falciparum kelch 13 protein mutations in P. Falciparum populations from Northeastern Myanmar in mediating artemisinin resistance. MBio *11*. e01134-19. https://doi.org/10.1128/mBio.01134-19.

Sidhu, A.B.S., Uhlemann, A.-C., Valderramos, S.G., Valderramos, J.-C., Krishna, S., and Fidock, D.A. (2006). Decreasing pfmdr1 copy number in plasmodium falciparum malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. J. Infect. Dis. *194*, 528–535.

Sisowath, C., Ferreira, P.E., Bustamante, L.Y., Dahlström, S., Mårtensson, A., Björkman, A., Krishna, S., and Gil, J.P. (2007). The role of pfmdr1 in Plasmodium falciparum tolerance to artemether-lumefantrine in Africa. Trop. Med. Int. Health *12*, 736–742.

Suresh, N., and Haldar, K. (2018). Mechanisms of artemisinin resistance in Plasmodium falciparum malaria. Curr. Opin. Pharmacol. *42*, 46–54.

Tabbabi, A., Alkishe, A.A., Samy, A.M., Rhim, A., and Peterson, A.T. (2020). Malaria in north Africa: a review of the status of vectors and parasites. J. Entomol. Sci. *55*, 25–37.

Takala-Harrison, S., Clark, T.G., Jacob, C.G., Cummings, M.P., Miotto, O., Dondorp, A.M., Fukuda, M.M., Nosten, F., Noedl, H., Imwong, M., et al. (2013). Genetic loci associated with delayed clearance of Plasmodium falciparum following artemisinin treatment in Southeast Asia. Proc. Natl. Acad. Sci. U. S. A. *110*, 240–245.

Talapko, J., Škrlec, I., Alebić, T., Jukić, M., and Včev, A. (2019). Malaria: the past and the present. Microorganisms *7*, 179. https://doi.org/10.3390/microorganisms7060179.

Tarr, S.J., Moon, R.W., Hardege, I., and Osborne, A.R. (2014). A conserved domain targets exported PHISTb family proteins to the periphery of Plasmodium infected erythrocytes. Mol. Biochem. Parasitol. *196*, 29–40.

Tilley, L., Straimer, J., Gnädig, N.F., Ralph, S.A., and Fidock, D.A. (2016). Artemisinin action and resistance in plasmodium falciparum. Trends Parasitol. *32*, 682–696.

Ursing, J., Zakeri, S., Gil, J.P., and Björkman, A. (2006). Quinoline resistance associated polymorphisms in the pfcrt, pfmdr1 and pfmrp genes of Plasmodium falciparum in Iran. Acta Trop. *97*, 352–356.

van Biljon, R., Niemand, J., van Wyk, R., Clark, K., Verlinden, B., Abrie, C., von Grüning, H., Smidt, W., Smit, A., Reader, J., et al. (2018). Inducing controlled cell cycle arrest and re-entry during asexual proliferation of Plasmodium falciparum malaria parasites. Sci. Rep. *8*, 16581.

Warncke, J.D., Vakonakis, I., and Beck, H.-P. (2016). Plasmodium helical interspersed Subtelomeric (PHIST) proteins, at the center of host cell remodeling. Microbiol. Mol. Biol. Rev. *80*, 905–927.

World Health Organization (2020). World Malaria Report 2019 (World Health Organization).

Zhu, L., Tripathi, J., Rocamora, F.M., Miotto, O., van der Pluijm, R., Voss, T.S., Mok, S., Kwiatkowski, D.P., Nosten, F., Day, N.P.J., et al. (2018). Tracking Resistance to Artemisinin Collaboration I, 2018. The origins of malaria artemisinin resistance defined by a genetic and transcriptomic background. Nat. Commun. *9*, 5158.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
in vivo Plasmodium falciparum transcriptomes	GEO	GSE59099
in vitro Plasmodium falciparum transcriptomes	Synapse Storage	syn16924919
ex vivo Plasmodium falciparum transcriptomes	GEO	GSE25878
ex vivo Plasmodium falciparum transcriptomes	GEO	GSE59098
in vitro Plasmodium falciparum transcriptomes	GEO	GSE61536
in vitro Plasmodium falciparum transcriptomes	GEO	GSE151189
Software and Algorithms		
ART-resistance prediction model	Github	https://github.com/GuanLab/Predict-Malaria-ART-Resistance

RESOURCE AVAILABILITY

Lead contact and materials availability

This study generated no new materials. Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yuanfang Guan (gyuanfan@umich.edu).

Data and code availability

- All Plasmodium falciparum transcriptome data used in this paper have been deposited in GEO and Synapse storage, and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at Github and is publicly available as of the date of publication at: https://github.com/GuanLab/Predict-Malaria-ART-Resistance.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data pre-processing

The *in vivo* prediction model was built based on clinical population data from the published paper by Mok et al. (2015) and provided by the Malaria DREAM challenge. The *P. falciparum* isolates were collected ~18 hours post invasion from 1,043 acute patients under varying treatment and health conditions mainly from Southeast Asia. The parasite isolate transcriptome was analysed by Bozdech two-color microarray platform, with 10,159 unique probes covering 5363 genes(Bionetworks, n.d.a). The Artemisia resistance status of the *P. falciparum* isolates was labeled as ‘Fast’ or ‘Slow’, indicating the clearance rate of *P. falciparum* after ART treatment.

The test transcriptome data was generated by Agilent HD Exon one-color microarray platform from 30 *P. falciparum* isolates collected from Thai-Myanar border from 2007 to 2012, as provided by the Malaria DREAM Challenge, which includes 63,976 unique probes covering 5440 genes including non-coding RNS(Bionetworks, n.d.a). The isolates were cultured in blood cells and treated by the artemisinin 6 and 24 hours post invasion (hpi). The IC50 of *P. falciparum* culture, *i.e.*, the drug concentration that 50% of parasites die was recorded as an indicator of ART resistance. Higher IC50 means stronger ART-resistance, therefore corresponds to a slow clearance rate.

The training and testing microarray data were then processed and normalized by different pipelines with respect to their own microarray platforms (Bionetworks, n.d.a). The two-color *in vivo* microarray data were processed by GenePix Pro v6.0 software, where features of each array were extracted with foreground intensity > 1.5 fold background intensity for either channel, and went through background correction and lowess normalization using the limma R package. Then the arrays were log normalized against co-hybridized 3D7 control, and the gene expression levels were acquired by averaging their ORF Probe intensities. The *in vitro* single-color microarray data were processed by Agilent Feature extraction and QC pipeline, then quantile normalized by the preprocessCore R package. Then samples were log normalized against NF54 control, and batch corrected by the sva R package. Then the gene expression levels were obtained by the reshape R package.

The microarray data usually contains missing values due to artifact and technical failures. If the expression level of gene *i* of sample *j* is missing, we fill into the average gene expression level of gene *i*, based on the data from the rest of samples.

$$\overline{x_i} = \frac{1}{N}\sum_{j=O}^{j=N} x_{ij}$$

In order to make a robust cross-platform model, we used rank normalization to process the raw gene expression data, specifically,

N : the total number of samples

m : the total number of genes

x_{ij} : the expression level of gene j in sample i

x_{ir} : the expression level of rth ranked gene in sample i

R_i : the expected expression level of ranked i gene in a sample

$X_i : \{x_{j1}, x_{j2}, x_{j3}, ..., x_{jm}\} = \{x_{r_1}, x_{r_2}, x_{r_3}, ..., x_{r_m}\}$where $x_{r_1} < x_{r_2} < x_{r_3} < ..., x_{r_m}$

$R_i = \frac{1}{N}\sum_{1}^{N} x_{r_i}$

$X_i' = \{R_1, R_2, R_3, ..., R_m\}$

The microarray record of sample i is transformed from X_i to X_i' . The preprocessed *in vivo* and *in vitro* data was then used in the model training and prediction.

Model training

We tested five types of base learners, including LightGBM, XGboost, random forest, GPR and linear regression. The first three base learners are tree-based and the later two are kernel-based algorithms. For LightGBM, we used gradient boosted decision trees, with 5 as the number of leaves, a learning rate of 0.05 and a total of 800 estimators, and 1000 boosting rounds. For random forest, we used a maximal depth of 2 and 100 estimators. For GPR we used dot products and a white kernel. For all other base learners, we used the default parameters. ten-fold cross validation was used to evaluate the performance of models. The ten *in vivo* models were transferred to *in vitro* data to make predictions of the ART resistance of *P. falciparum*.

For cross-platform prediction, the shared genes were used in model construction. Each *P. falciparum* strain was sampled under four different conditions (6 hpi, 24 hpi, with or without ART perturbation), and each sample carried two biological replicates. We conducted cross-platform prediction on the 4 conditions, respectively. For each condition, the average prediction values of the two biological replicates are used as the final prediction.

SHAP feature importance analysis

We conducted SHAP (SHapley Additive exPlanations) analysis to evaluate the contributions of different genes in ART resistance prediction. The SHAP value describes the average marginal contribution of a feature across all instances (Lundberg and Lee, 2017). We summed up absolute values of SHAP values of all samples for each feature. The summary plot sorting features by the sum of the absolute SHAP values over all samples are included in Figures S2 and S3.

Co-expression and functional analysis of top genes

We conducted co-expression analysis on rank normalized gene expression level among the top-ranked genes by SHAP analysis, for both *in vivo* and *in vitro* datasets. The co-expression significance between two genes is defined as the Pearson's correlation of their normalized expression level across all samples. For example, for gene *i* and *j* in all *N* samples, X_i and X_j refer to the rank normalized expression level of both genes, respectively. Then,

$$X_i = \{x_{i1}, x_{i2}, \ldots x_{in}\}$$
$$X_j = \{x_{j1}, x_{j2}, \ldots x_{jn}\}$$

where *n* refers to the total number of samples in the dataset. The co-expression level $r_{i,j}$ between two genes is:

$$r_{i,j} = cor(X_i, X_j)$$

where $r_{i,j}$ is the Pearson's correlation between gene *i* and gene *j*. The co-expression network of both *in vivo* and *in vitro* dataset was constructed based on the significantly correlated genes ($r_{i,j}$ >0.4) and visualized using *ggraph*.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation of prediction performances

Because *in vivo* data and bore binary labels, we used AUROC (Area under the Receiver Operating Curve) and AUPRC (Area under the Precision Recall Curve). For the *in vitro* data, because the evaluation is a real value, we used Spearman and Pearson's correlations and C-index, as clearance rate and IC50 do not share the same distribution (Figures 2A and 2B). The C-index is calculated as the following:

$$C - \text{index} = \frac{\sum_{i,j} 1_{p_i<p_j} \cdot 1_{IC50_i<IC50_j}}{\sum_{i,j} 1_{p_i<p_j}}$$

p_i : the predicted value of sample i, ranges from 0 to 1.

$IC50$: the $IC50$ of sample i.

$1_{p_i<p_j} = 1 if p_i<p_j$, else 0

$1 IC50_i<IC50_j = 1$ if $IC50_i<IC50$, else 0.

C-index is equivalent to AUROC when predicting binary labels.

For external validation datasets, the labels were also binary, thus we use AUROC for performance evaluation. AUPRC was not used for horizontal comparison since the baseline for each dataset is different. 95% confidence intervals of all performances were calculated by bootstrapping.

The Pearson and Spearman's correlation coefficient, AUROC and AURPC were calculated using *Python Sklearn* module. The code implementing c-index was provided in the github repository (see Data and code availability in Resource Availability section).