MACHINE LEARNING METHODS FOR INFECTIOUS DISEASES: APPLICATIONS FOR TUBERCULOSIS and MALARIA

Wouter André Deelder

Thesis submitted in accordance with the requirements for the
degree of Doctor of Philosophy

University of London
April 2022

Faculty of Epidemiology and Population Health
Department of Medical Statistics

LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

Research group affiliation: Prof. Taane G. Clark and Dr. Luigi Palla

No funding was received

I, Wouter André Deelder, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

Signed

Date: 18 April 2022

Abstract

Infectious diseases such as malaria (caused by *Plasmodium spp* parasites) and tuberculosis (TB, caused by *Mycobacterium* bacteria) are major public health challenges, being leading causes of death worldwide, particularly in low-income countries. The genomes of the underlying causal pathogens contain valuable information to guide clinical treatment and programmatic control decision making. Whole genome sequencing (WGS) has therefore emerged as an increasingly common approach to characterize genetic mutations (e.g., single nucleotide polymorphisms; SNPs) and understand the diversity of these pathogens. However, WGS leads to high dimensional datasets ("big data"). Some established statistical methods are less suited to such big data analysis, and machine learning (ML) approaches offer a promising alternative for modelling and inference.

In this thesis, I explore the application of ML methods, including deep learning, to WGS datasets for malaria parasites (*P. falciparum* and *P. vivax*) and *M. tuberculosis* bacteria. For *M. tuberculosis* (n=17k; >100k SNPs; genome size 4.4 Mbp), I applied non-parametric classification-tree and gradient-boosted-tree models to predict drug resistance across 14 anti-TB drugs. For established first-line drugs, the models had high predictive ability (area under the receiver operating curve > 0.85), and included SNPs in candidate genes linked to drug-resistance. For drugs with less established knowledge, I developed a customized decision tree approach ("Treesist-TB"), which performs TB drug resistance prediction by extracting and evaluating genomic variants across multiple studies. Treesist-TB revealed both known and novel putative SNPs for resistance and had improved predictive sensitivity compared to a widely-used TB mutation database (TB-Profiler tool).

For *P. falciparum* (n>1,100; >74k SNPs; genome size 26.8 Mbp) and *P. vivax* (n>350, >125k SNPs; genome size 23.3 Mbp), I developed an image-based convolutional neural network (CNN) approach ("DeepSweep"), with the aim of identifying genetic regions subject to recent positive selection, such as those linked to the onset of antimalarial drug resistance. DeepSweep detected genetic regions proximal to known and suspected drug resistance loci for both *P. falciparum* (e.g., *pfcrt, pfdhps* and *pfmdr1*) and *P. vivax* (e.g., *pvmrp1*), and detected signals overlapping with those from two established extended haplotype homozygosity methods. Finally, I applied ML approaches, including CNNs, to predict the geographic origin of *P. falciparum* and *P. vivax* infections at different levels of geographic granularity (continents, countries, GPS locations). Classification methods had the lowest distance errors, and >90% accuracy at a country level, thereby demonstrating the utility of ML approaches for detecting imported infections and the geo-classification of malaria parasites.

Overall, these applications demonstrate the potential of ML methods to extract new insights from large WGS datasets and assist infection control. However, there are risks in applying ML methods on WGS data "out of the box" without context-specific adaptation of the algorithms. My work demonstrates how

adaptation of standard ML methods can lead to better predictions and more interpretable results, offering greater assistance to infection control decision making.

Acknowledgements

I would firstly like to express my gratitude to my supervisors Taane Clark and Luigi Palla for taking a chance on me and giving me the opportunity to work on a PhD under their guidance. I am extremely thankful that they welcomed me with open arms in their research groups and that they provided me with generous and patient support along the way. I am also grateful for the fun and collaborative atmosphere that they created, which made the journey a lot more enjoyable.

I want to thank Ernest Diez Benavente for his guidance and support on the malaria papers.

I want to thank Jody Phelan for the guidance and support he gave me on the tuberculosis papers, and for always being available in a gracious and kind manner to help to resolve bigger and smaller barriers and bottlenecks across the research.

I want to express my gratitude to my colleague Wijnand de Wit at Dalberg for being supportive in me pursuing this PhD.

I want to thank my family for all the care and guidance they have given me throughout my life: My mother, my brothers Bas and Kees, and my sister Krista.

I would not have started this endeavour without the encouragement and support of my wife Emma Clarke-Deelder. I am very grateful that she nudged me in this direction and supported me throughout.

My son Andrew was born towards the end of my PhD. He made my life infinitely better (although the last stretches of this PhD maybe a little more time-compressed)

My father, André Martien Deelder, was a proud parasitologist and might subconsciously have influenced me to pursue this journey. I wish he could have seen the completion of this work. This thesis is dedicated to him.

Contents

Acknowledgements 5

List of Tables 7

List of Figures 7

Abbreviations 8

Additional Publications 9

Introduction 10

  Overview 10

  Infectious diseases 10

  Malaria 10

    Human Malaria Plasmodia 10

    Global malaria burden 11

    The *Plasmodium* genomes 12

    Plasmodium drug resistance 12

  Detection and prediction of drug resistance for Malaria 14

  Tuberculosis 16

    Mycobacterium tuberculosis 16

    Global tuberculosis burden 16

    The *M. tuberculosis* genome 17

    M. tuberculosis drug resistance 18

  Detection and prediction of drug resistance for Tuberculosis 19

  Whole genome sequencing (WGS) and bioinformatics 20

  Machine learning 21

  Project structure 24

Chapter 2 27

Chapter 3 49

Chapter 4 69

Chapter 5 106

Discussion 124

Conclusions 129

References 130

List of Tables

Table 1. Time of widespread usage, time and location of onset of resistance for major anti-malarial drugs and associated genes 13
Table 2. Genes associated with drug resistance for Mycobacterium Tuberculosis 20
Table 3. Overview of the genomic datasets used in this thesis 21
Table 4. Machine learning methods used in this thesis 23

List of Figures

Figure 1. Schematic illustration of the concept of selective sweeps 14
Figure 2. Treatment outcomes of TB over time 17
Figure 3. Global distribution of TB lineages 18
Figure 4. TB phylogenetic tree with color coding by (sub)lineage 18
Figure 5. Mechanisms associated with drug resistance 19

Abbreviations

- ACTs: Artemisinin-based Combination Therapies
- bp: Base pairs
- CNN: Convolutional Neural Network
- DALY: Disability-Adjusted Life-Year
- DCT: Decision Tree
- DST: Drug Sensitivity testing
- GBT: Gradient Boosted Tree
- GWAS: Genome Wide Association Study
- Indels Insertions or deletions
- Kbp: kilobase pairs
- LD: Linkage Disequilibrium
- ML: Machine Learning
- MDR: Multi-Drug Resistant
- NS : Non-Synonymous
- P. : Plasmodium
- PU : Positive Unlabeled
- SNPs: Single Nucleotide Polymorphisms
- TB: Tuberculosis
- XDR: Extensively-Drug Resistant
- WGS: Whole Genome Sequencing
- WHO: World Health Organization

Additional Publications

I contributed to other manuscripts, which were not part of my PhD:

*Jody Phelan, **Wouter Deelder**, Daniel Ward, Susana Campino, Martin L. Hibberd, Taane G Clark;* COVID-Profiler: a webserver for the analysis of SARS-CoV-2 sequencing data", BMC Bioinformatics; 2022 Apr 15;23(1):137. doi: 10.1186/s12859-022-04632-y. PMID: 35428185

Introduction

Overview

This thesis explores the potential for machine learning methods to overcome the challenges presented by traditional statistical methods in the analysis of whole genome sequencing data and how these methods may help contribute to the global fight against infectious diseases such as malaria and tuberculosis. In doing so, this introduction aims to set out the global burden of these infectious diseases and the remaining challenges in addressing them, including how pathogen drug resistance can challenge their control. It covers the growing importance of whole genome sequence data, and how they can inform the study of loci linked to drug resistance. Further, it describes the challenges encountered by traditional methods when applied to the analysis of "big" genomic datasets. Finally, this chapter introduces machine learning methods, and in particular, the subset of methods applied to the genomic datasets.

Infectious diseases

Infectious diseases are estimated to have inflicted a burden of 574 million Disability-Adjusted Life Years (DALYs) on the world's population in 2020 (1). Malaria and tuberculosis (TB) are among the highest-burden infectious diseases in terms of mortality and morbidity, responsible for 0.6 million and 1.3 million deaths in 2021, respectively (2,3). For this reason, this thesis focuses on these two diseases and their causal pathogens, with a particular focus on *Plasmodium falciparum* and *Plasmodium vivax* for malaria and *Mycobacterium tuberculosis* for TB. It explores how machine learning methods, applied to datasets of whole genome sequences, can contribute analytical insights to support better programmatic and clinical outcomes.

Malaria

Human Malaria Plasmodia

Malaria is an infectious disease caused by protozoan parasites of the genus *Plasmodium* (4). The oldest Plasmodium protozoa, extracted from mosquitoes that were entrapped in amber, stem from approximately 30 million years ago (5). There are approximately 200 Plasmodia species, which infect birds, reptiles and primates (6). Six parasite species infect humans: *P. falciparum, P. vivax, Plasmodium ovale curtesi, Plasmodium ovale wallikeri, Plasmodium malariae and Plasmodium knowlesi* (4). *P. falciparum* infections occur in tropical areas around the world, and cause most of the overall global malaria mortality. *P. vivax* is less temperature sensitive and therefore more geographically widespread, occurring across large parts of Southeast Asia and Central and South America, as well as Ethiopia (7). *P. ovale* infections primarily occur in sub-Saharan Africa and islands in the western Pacific, and *P. malariae* infections occur across large areas of Sub-Saharan Africa, South America and South East Asia, including in co-infections with *P. falciparum* (7).

Malaria parasites undergo a complex lifecycle, involving humans and Anopheles mosquitos, where the parasites ultimately reach the human blood stream after a bite from an infected mosquito. The clinical manifestations of malaria are linked to parasites invading red blood cells (erythrocytes), where they multiply until the cells burst, upon which a subsequent cycle is started with the invasion of new red blood cells. The characteristic malarial fever occurs at this phase of erythrocyte escape and invasion (4). This can lead to severe anemia, and in addition, *P. falciparum*-infected erythrocytes can adhere to walls of blood vessels, which, when occurring in cerebral microvasculature, can cause cerebral malaria, which may be fatal (4).

Global malaria burden

Malaria has affected humanity for centuries. The Roman poet Livius already described how different epidemics plagued the Romans, who had the suspicion that the disease arose from the swamps around the city (8). Despite many decades of public health efforts to reduce the global malaria burden, malaria continues to be a major public health problem. In 2020, there were an estimated 241 million malaria cases worldwide, of which 627,000 cases resulted in death (2). Sub-Saharan Africa accounts for approximately 95% of all global malaria cases and deaths. The vast majority of malaria deaths occur in children less than five years old (2), principally due to *P. falciparum* infections.

The world has made progress in decreasing the burden of malaria in the last decades, although a reversion has taken place in recent years. Between 2000 and 2015 global cases dropped from 241M to 224M and deaths declined from 896,000 to 562,000 (2). The downward trend can be attributed to increased funding (e.g., through the creation of the Global Fund to Fight HIV/AIDS, TB and Malaria and the President's Malaria Initiative) and the accompanying scale-up of long-lasting insecticide-treated nets, indoor residual spraying, rapid diagnostic tests, and artemisinin-based combination therapies (ACTs) (9). However, since 2015 the downward trend has flattened out and even reverted in some geographies. Between 2015 and 2019, the global case-load increased slightly (from 224M in 2015 to 227M in 2019) and the global deaths declined only slightly (from 562,000 in 2015 to 558,000 in 2019) (2). This development, which puts the 2030 global malaria targets as formulated by the global health community in 2016 at risk, has been attributed to a more difficult funding environment, conflict and climate change, as well as to increasing levels of resistance to ACTs, bednets and insecticides (2). Furthermore, in 2020 the global COVID-19 pandemic caused additional disruption to malaria control efforts, causing an increase in malaria cases of 14M to 241M and an increase in malaria deaths of 47,000 to 627,000 (2).

The *Plasmodium* genomes

This thesis involves the analysis of *P. falciparum* and *P. vivax* parasite genomic data. The *P. falciparum* genome is 23 Mbp in length across 14 chromosomes, apicoplast and mitochondrial DNA, with a GC content of 19.4%, and contains 5,300 genes. The first reference genome was Pf3D7 (10,11). The *P. vivax* genome has a length of 29 Mbp across 14 chromosomes, apicoplast and mitochondrial DNA, with a GC content 40.6% and contains 5,400 genes (10,11). There are two reference genomes *P. vivax* Salvador I (Pvsal1) and PvP01 (from South East Asia) (10,11).

Genomic diversity studies in *P. falciparum* and *P. vivax* have shown that there is geographic clustering, with strong genomic differences between continents, which coincide with loci and mutations linked to drug resistance, response to mosquito vectors and (evasion of) the human immune system (12,13). Molecular barcodes to classify species and geographic origin have been developed. These barcodes use only a subset of the genome due the high-dimensionality of the full genome and the associated computational cost (12,14).

Plasmodium drug resistance

Growing resistance to anti-malarial drugs poses a serious challenge to global efforts to reduce the burden of malaria. Anti-malarial drugs are an essential tool for the treatment of infected individuals and a critical pillar of malaria control programs. However, the usage of anti-malarials and the onset of resistance are intimately intertwined. Quinine was the first modern malaria drug and was developed in 1820 by Pellentier and Caventou, with formal reports of *in vivo* resistance in Brazil and South East Asia being published in the late 1950s (15). A new class of anti-malarials, four-amino quinolines, was developed in the 1940s, with chloroquine being the key member of this class (16). However, chloroquine-resistant infections (for *P. falciparum*) emerged independently in at least three locations, namely the border between Thailand and Cambodia (1957), the Venezuelan-Colombian border (1960s) and Papua New Guinea (1970s) (16), and chloroquine resistance is now widespread.

Sulfadoxine-pyrimethamine (SP), a combination drug which targets enzymes in the folate pathway, replaced chloroquine as first-line treatment for malaria in the period from the 1960s to the 1980s, and is currently used for preventative treatment in pregnancy and, in infants, as part of seasonal malaria chemoprophylaxis in sub-Saharan Africa (2,16). However, resistance to SP first emerged in the late 1960s in Thailand and subsequently spread to Sub-Saharan Africa (16). In general, resistance has emerged for each newly developed drug, with the first findings of resistance often arriving from the Mekong region in South East Asia. Most worryingly, resistance to artemisinin-combination therapies, the current first-line treatment for *P. falciparum*, has been observed in the form of delayed parasite clearance in South East Asia, posing a threat to the effectiveness of the current control paradigm (17).

Table 1: Time of widespread usage, time and location of onset of resistance for major anti-malarial drugs and associated genes (18–24)

Drug	Date of widespread release	Date of detection of first in-vivo resistance	Location of first *in vivo* resistance	*P. falciparum* genes involved in resistance
Chloroquine	1946	1957	Thai-Cambodian border	*pfmrp1,pfmdr1, pfcrt*
Sulfadoxine-pyrimethanimine (SP)	1967	1967	Thailand	*Pfdhfr, pfdhps*
Pyronaridine	1970	1985	China	
Artemisinin monotherapy	1971	1989	Vietnam	*pfkelch13*
Mefloquine	1977	1982	Thailand	*Pfmdr1*
Piperaquine	1978	1981	China	
Artesunate-mefloquine	1992	2007	Thai-Cambodian border	*pfkelch13, Pfmdr1*
Artemether-lumefantrine	1999	2006	Thai-Myanmar border	*pfkelch13*
Dihydroartemisinin-piperaquine	2007	2007	Western Cambodia	*pfkelch13*
Pyronaridine-artesunate	2012	2012	Western Cambodia	*pfkelch13*

Anti-malarial resistance is triggered by genomic mutations, which might for example alter the transport of the drug into or out of the parasite's vacuole, or alternatively change the binding target of the drug (25) (**Table 1**). For example, *P. falciparum* parasites can become resistant to chloroquine or amodiaquine through mutations in the *Pfcrt* gene, which encodes a transporter that can transfer these drugs out of the vacuole before they can exert their mechanism of action. Similarly, resistance against SP occurs through mutations in the *Pfdhps* and *Pfdhfr* genes, inhibiting the activity of two key enzymes in the folate pathway (25). The underlying mutations causing resistance for *P. vivax* are less well defined than for *P. falciparum* (29) although putative genes for resistance to chloroquine (*Pvmdr1, Pvcrt*), primaquine (*Pvmrp1)* and SP (*pvdhps, pvdhfr*) have been defined (26).

Drug resistance threatens to undo the progress made in the fight against malaria, at both a clinical and population level. New research methods, such as those explored in this thesis, can play a role in informing the fight against drug resistance by allowing us to better characterize and predict drug resistance within individual patients and within populations at large.

Detection and prediction of drug resistance for Malaria

The detection of drug-resistance against anti-malarial drugs was historically only possible through observation of *in-vivo* treatment failure (27). The arrival of genomic sequencing opened the possibility to make drug-resistance predictions based on the presence or absence of genomic markers. The lack of labelled phenotypic data for *P. vivax* and *P. falciparum*, in contrast to *M. tuberculosis* bacteria where phenotypic data is more readily available, does complicate training statistical models. It requires us to look for genomic markers of drug resistance in an indirect manner, for example by finding signatures of positive selection across the parasite genome, in the assumption that drug resistance leads to a selective advantage.

A particular signature of interest is the so-called selective sweep. Selective sweeps arise as beneficial alleles increase in frequency over time and "sweep" through populations. These sweeps leave tell-tale genomic signatures in the site-frequency spectrum, the amount of population differentiation and the pattern of linkage disequilibrium (28) (**Figure 1**).

Figure 1: Schematic illustration of the concept of selective sweeps, adapted from (29)

The detection of these selective sweeps, and signatures of positive selection in general, has historically been performed using a wide variety of methods and approaches for a wide variety of species (30–34). There are at least three common approaches to detect positive selection sweeps in non-clonal species. First, it is possible to assess the differentiation of genomic loci between populations, particularly through differences in allele frequency. For example, populations that are exposed to different drugs may be subject to different selection pressures, leading to differences in mutation frequency underlying resistance. Second, one can assess differences in site-frequency distributions or spectra. Lastly, one can assess the extent of linkage disequilibrium (or correlation between genetic markers) and extended haplotype homozygosity at loci (28,35).

These methods were originally pioneered on the human genome (30), but they have been subsequently utilized for *Plasmodium* and helped to identify genetic markers associated with drug resistance (13,36). Recently, efforts have been made to efficiently apply these methods to whole genome sequencing libraries, such as REHH, SweeD and OmegaPlus (37–39). However, these tools and methods require careful parameter definition and calculation, and the outcomes are sensitive to the SNPs included, population structure and the chosen statistical significance thresholds.

Researchers have explored the potential of applying machine learning methods to the detection of selective sweeps (40). To date, most of the methods aim to make predictions using pre-calculated population genetic statistics as features (such as Tajima's D and Fay and Wu's H statistics) (28,41,42). Thus, this approach does not solve the challenge of defining and calculating these population genetic statistics, which is a complex and time-consuming task, especially when working with many sub-populations. As will be shown, a (deep) machine learning approach might provide an interesting alternative, given its potential to learn from a relatively rudimentary set of base features that require little to no pre-definition by the user (43).

Tuberculosis

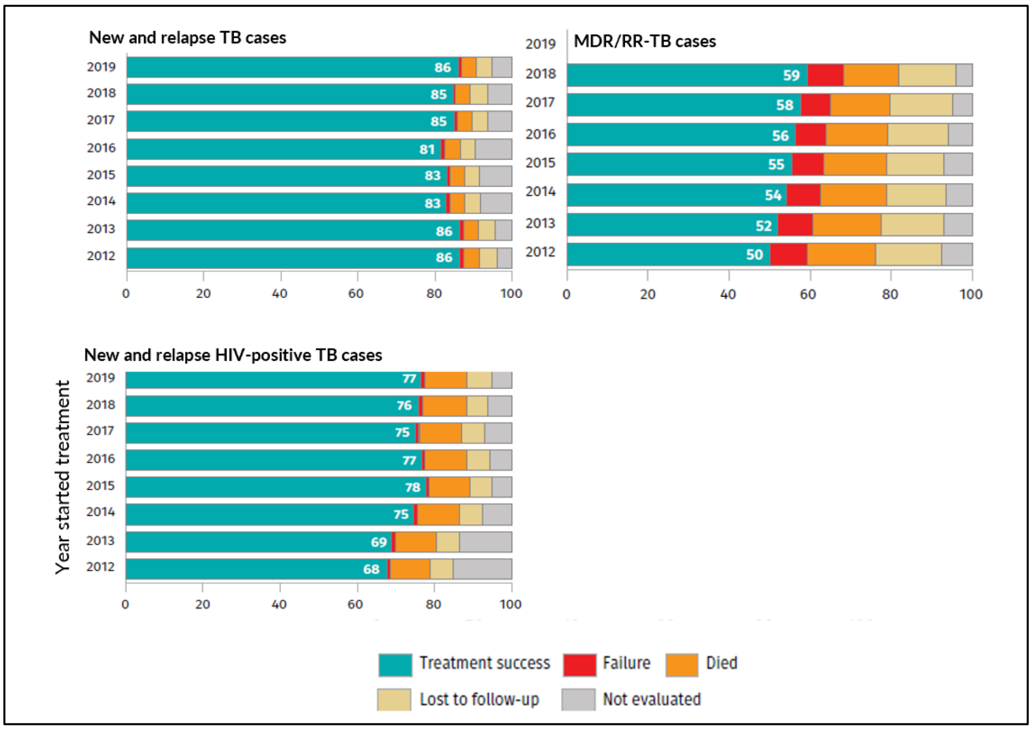
Mycobacterium tuberculosis

Tuberculosis (TB) is an infectious disease caused by members of the *M. tuberculosis* complex, which includes *M. tuberculosis, Mycobacterium bovis* and *Mycobacterium africanum* (44). This thesis focuses on *M. tuberculosis*. TB is spread through the inhalation of aerosols that contain *M. tuberculosis* bacteria. Upon infection, the bacteria will invade and replicate within alveolar macrophages. In most affected individuals, the subsequent immune response will lead to an immunological equilibrium and a latent stage in which the bacteria are encapsulated in granulomas in the lungs (so-called primary lesions). In some individuals however, an active TB infection might develop (post-primary disease), which requires treatment with appropriate antibiotics to prevent potentially fatal outcomes (45).

Global tuberculosis burden

Until recently, TB was the leading cause of death from a single pathogen (until being overtaken by COVID-19). In 2020, there were an estimated 1.3M global deaths caused by tuberculosis among HIV-negative people (3). Worryingly, this number was up from 1.2M in 2019, with reductions in access to TB screening, treatment and care due to the COVID-19 pandemic estimated to be one of the driving factors (3). People living with HIV or infected with drug-resistant TB have significantly worse treatment outcomes than other TB patients (3) (**Figure 2**). Drug resistant *M. tuberculosis* is one of the major threats to effectively control the disease, especially resistance to first-line rifampicin (RR-TB) and isoniazid drugs; in combination, called multi-drug resistance (MDR-TB). RR-TB and MDR-TB together accounted for around 130,000 cases in 2020 (3). Additional resistance to second-line drugs can lead to extensively drug-resistant strains (XDR-TB) (46). In recent years, there have been new definitions for pre-XDR (now defined as MDR-TB and resistance to any fluoroquinolone) and for XDR-TB (now defined as MDR-TB with additional resistance to any fluoroquinolone, and either bedaquiline or linezolid or both. In this thesis, the old definition of XDR-TB is used (defined as MDR-TB with additional resistance to fluoroquinolones and second-line injectables (47) for the reason that this this matches the definition in use when the isolates were collected and when the treatment decisions were made.

Figure 2: Treatment outcomes of TB over time (3)

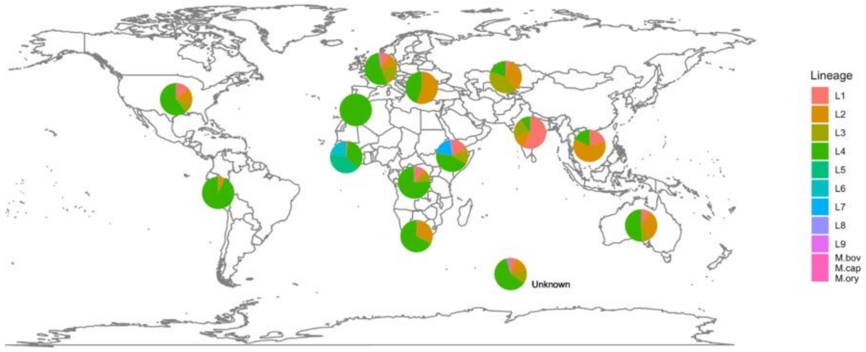


The *M. tuberculosis* genome

The *M. tuberculosis* genome is 4.4 Mbp long and encodes for approximately 4,000 genes (48). Gene expression varies over the duration of an infection, with periods of slow growth and dormancy being part of the characteristic features of *M. tuberculosis* (49). The mechanisms behind these characteristics are not fully understood but likely contribute to the complication and lengthening of TB treatment.

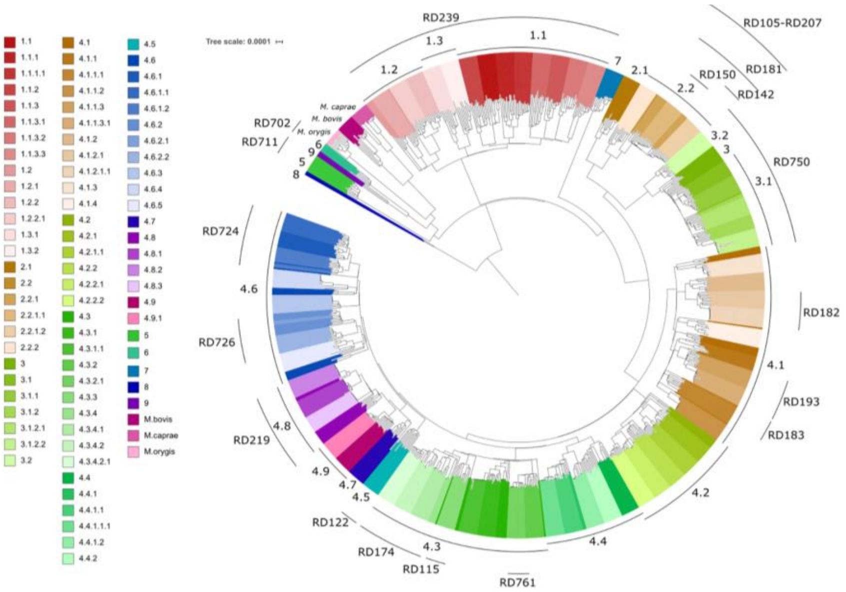
Analysis of the *M. tuberculosis* genomic diversity has confirmed phylo-geographical groupings. There are currently 9 main human lineages and 64 sub-lineages that exhibit strong geographic clustering (50). Four lineages are dominant (lineages 1 to 4) (**Figure 3, Figure 4**), but the others are in isolated parts of Africa (e.g., *M. africanum* lineages 5/6 in West Africa). Strain-specific genomic diversity is associated with differences in virulence, pathogenicity and transmissibility (50). The genomic differences have enabled the development of barcodes to facilitate the identification of lineage and sub-lineage (50).

Figure 3: Global distribution of TB lineages) (50)



The global distribution of the 35,298 *Mycobacterium tuberculosis* complex study isolates

Figure 4: TB phylogenetic tree with color coding by (sub)lineage (50)



Phylogenetic tree of *Mycobacterium tuberculosis* complex isolates. A representative tree with a maximum of 10 isolates per sub-lineage (important regions of difference (RDs) are also highlighted)

Analysis of phylogenetic data and strain-specific diversity has also led to the identification of drug resistance mutations, sometimes appearing in multiple branches of the tree, and outbreaks and transmission events may be identified by finding isolates with near-identical genetic variation, with supporting epidemiological data (51–53).

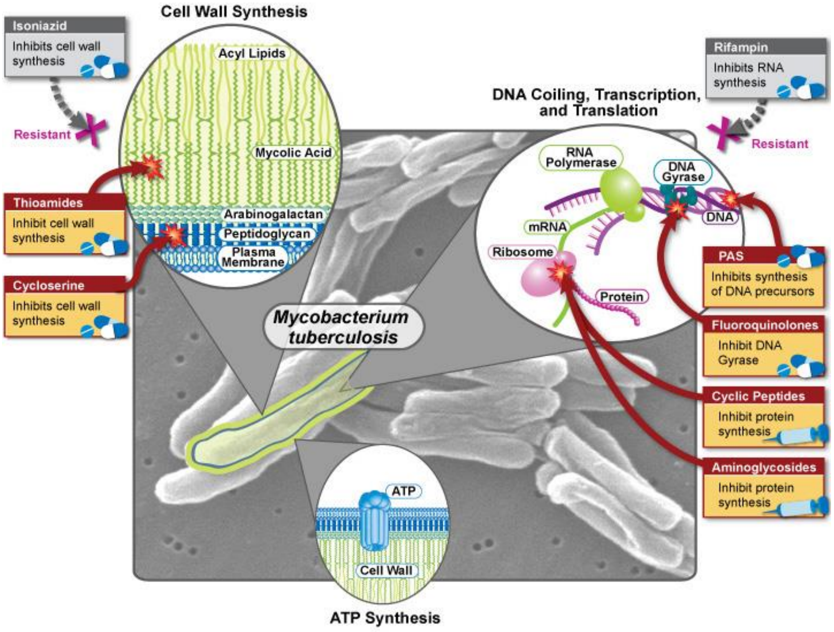
M. tuberculosis drug resistance

As stated earlier, resistance against first-line tuberculosis treatments is hindering global efforts to reduce the disease burden (WHO, 2018). Tuberculosis requires treatment with antibacterial drugs to reduce morbidity and prevent potential mortality. First-line anti-TB therapy is centred around four drugs: rifampicin (RMP), isoniazid (INH), ethambutol (EMB) and pyrazinamide (PZA) (54). However, there is increasing resistance to these drugs, especially RR-TB and MDR-TB (54). Second-line drugs are available to treat MDR cases, most importantly the fluoroquinolones (FQ; ciprofloxacin (CIP), ofloxacin (OFL), moxifloxacin (MOX)) and the injectables (INJ; amikacin (AMK), kanamycin (KAN), capreomycin

(CAP)). Treatment of drug-resistant TB is however far more complex, costly and time-consuming than the first-line protocol (itself having a duration of 6 months) and includes the usage of drugs with severe adverse effects (46).

The onset of resistance against TB drugs likely occurs predominantly through mutations (e.g. single nucleotide polymorphisms (SNPs); small insertions and deletions) that modify drug targets (e.g. the proteins synthesized by the *rpoB* gene for rifampicin) or alter the proteins involved in the activation of pro-drugs (e.g. *katG* gene for isoniazid) (55–57) (**Figure 5; Table 2**).

Figure 5: Mechanisms associated with drug resistance (58)



It is important to emphasize that the timely detection and treatment of drug-resistant TB is critical not only for clinical outcomes at the individual patient level, but also for curbing the wider drug-resistant TB epidemic. Modelling studies have indicated that the transmission of undetected or untreated DR-TB, also in comparison to resistance potentially acquired during TB treatment, is a major contributor to the overall global burden of drug-resistant TB (59). In this thesis, the aim is to help develop new methods to make better predictions on whether patients carry drug-resistant TB strains, thereby providing these patients with an earlier diagnosis and more appropriate and effective treatment.

Detection and prediction of drug resistance for Tuberculosis

The determination of the drug-resistance profile for *M. tuberculosis* isolates has historically been performed in the laboratory by means of phenotypic testing, also called drug susceptibility testing (DST). However, this method is relatively slow and expensive, and it comes with inherent challenges that affect accuracy and reproducibility (60). The phenotypic datasets for the *M. tuberculosis* studies described in this paper are available upon request.

The challenges surrounding DST has fuelled interest in genotypic screening for drug resistance. However, the ability to make predictions based sequenced genomes requires an understanding of the genomic markers associated with resistance. Our knowledge of the genomic markers is unfortunately not complete, both in terms of individual markers and in terms of possible epistatic effects between drug-resistance markers or between these markers and compensatory mutations (61).

The sequencing of individual genes, and increasingly the sequencing of the entirety of pathogen genomes, and the increased availability of large genomic datasets, has enhanced the potential to infer these genomic markers. However, these data must be analyzed with appropriate statistical methods. Genomic-Wide Association Studies (GWAS) have historically been used to this aim (62), but these models need careful correction to account for population structure (e.g. lineages and sub-lineages in the case of *M. tuberculosis*) and moreover are less well suited to detect epistatic effects. Convergent evolution analysis (63) uses simple two-way table analysis methods to detect drug resistance mutations, but requires an accurate phylogenetic tree to account for the population structure, which is computationally more difficult to generate for large numbers of samples. In this thesis, the aim is to apply machine learning methods in a manner that expands on the GWAS approach in several ways, including allowing for more complex (epistatic) interactions between covariates.

Table 2: Genes associated with drug resistance for M. Tuberculosis (64)

Drug	**Genes**
Rifampicin	*rpoB, rpoC*
Isoniazid	*fabG1, inhA, katG, kasA, ahpC*
Pyrazinamide	*pncA*
Ethambutol	*embR, embC, embA, embB*
Streptomycin	*rpsL, gid, rrs*
Amikacin	*rrs*
Capreomycin	*tlyA ,rrs*
Kanamycin	*eis, rrs*
Ciprofloxacin, Ofloxacin, Moxifloxacin	*gyrA, gyrB*
Ethionamide	*fabG1, ethA*
Cycloserine	*alr, ald*
PAS	*folC, ribD, thyX, thyA*

PAS = para-aminosalicylic acid

Whole genome sequencing (WGS) and bioinformatics

The first whole genomes were sequenced using labor intensive capillary sequencing techniques (65). Subsequently, newer methods (e.g. shotgun sequencing and subsequently next-generation sequencing) allowed to sequence whole genomes at much greater scale and at much lower cost (65). More recently, methods have also been developed to sequence genomes outside the standard laboratory settings (e.g. Oxford Nanopore Technology sequencing) (65). The resulting datasets are typically analyzed using a combination of bioinformatic tools. Typically, raw sequence data is assessed for quality, and subsequently aligned to a reference genome or *de novo* assembled (reference-free), to allow the calling of variants (e.g., SNPs, indels). This process typically results in a rectangular dataset with dimensions based on the numbers of genomic variants and samples. The characterized dataset of genomic variation subsequently requires (statistical) analysis to provide relevant information for clinical and programmatic users.

M. tuberculosis, *P. falciparum* and *P. vivax* reference genomes were first sequenced in 1998, 2002 and 2008, respectively (10,11,48). WGS has emerged as an increasingly common approach to characterize genomic isolates, in both clinical and research settings, and the number of *M. tuberculosis*, *P. falciparum* and *P. vivax* isolates that have undergone WGS has grown steadily. The datasets used in the project are summarized (**Table 3**).

Table 3: Overview of the genomic datasets used in this thesis

	P. falciparum	*P. vivax*	*M. tuberculosis*
No. Isolates	5,957	658	32,689
No. countries	27	13	30
Median Sequencing coverage	>30 fold genome coverage	>50-fold genome coverage	>50-fold genome coverage
No. SNPs	~750k	~588k	~640k
Reference genome used	Pf3D7	PvP01	H37Rv

The availability of WGS datasets gives new possibilities to detect genomic drivers of resistance and make other predictions of interest. However, new methods are likely needed to accommodate the size of these datasets and to be able to detect epistatic interactions. The focus of my thesis is to understand whether machine learning methods applied to the large SNP datasets of *M. tuberculosis* and *Plasmodium* malaria pathogens can generate new insights and improve our ability to make predictions.

Machine learning

Machine learning (ML) is a sub-field within statistical learning, with the definition that a program or algorithm (the “machine”) is said to learn from experience “E” with respect to some class of tasks “T” and performance measure “P” if its performance at tasks in “T”, as measured by “P”, improves with experience “E” (66).

Over the past decade, there has been a growing interest in machine learning, fuelled by the increased availability of large datasets and increased computational power. These changes have brought attention to a specific set of models that allow for an alternative approach to analysing large datasets, with relatively few underlying model assumptions about the distribution and functional relationships between the included variables. These models have the potential to provide greater flexibility for problems of prediction in high dimensional variable spaces, when each individual variable contains limited information and with interactions between variables (67–69).

ML subdivides into the fields of supervised, unsupervised and reinforcement learning. My work focuses on supervised learning, where the task to learn is a mapping from inputs to outputs, given a labelled training set (i.e. the aforementioned experience) of input-output pairs (70). In contrast, unsupervised learning uses an training dataset without labels and focuses on identifying and describing structures within this dataset (71), and reinforcement learning aims to map decisions to situations or states by learning to optimize rewards (72). Deep learning is a subset of machine learning where models learn in a hierarchical layer-based manner with relatively simple features as the starting point (43).

The boundaries between the field of machine learning and the wider field of statistical learning are not well defined. In practice, the focus of the machine learning community, at least initially, was heavily centred on the topic of prediction, with a strong focus on predictive accuracy and computational speed, and with only trailing interest in inferential questions and the linkage to the wider statistical field (73). The main interest of many in the machine learning community was thus, phrased more informally, in "whether and how well we can predict", and much less in "why we can predict" and "why our models work".

The focus on prediction, together with the aforementioned increased availability of both computational power and large datasets, led to a burst of activity to creatively apply, combine, modify and adapt different statistical learning models, all with the aim to improve predictive performance. The popularity of the prediction contests on Kaggle, and the methods deployed by winning teams, provide a vivid illustration of this dynamic. The reasons, from a statistical theoretical perspective, on why certain approaches were effective was not always initially understood and in some cases only became better understood over time (73).

The development of convolutional neural nets provides a good example of this dynamic. The aim to achieve best-in-class performance on standardized image datasets led to the pioneering of many adaptations and modifications, such as for example the usage of max-pooling (74), drop-out (75), and new optimizers and activation functions (76). Only over time did the theoretical underpinnings of some of these innovations crystallize, explaining why they might be effective in improving predictive performance (as well as what the reasons might be that the entire class of convolutional neural nets has shown such impressive performance on a wide range of tasks) (77).

It is possible that, over time, machine learning might become a less distinct field within statistical learning, as the theoretical foundations and linkages to other statistical fields become better understood (73).

ML methods have gained rapid traction in healthcare and biomedical settings, given the large amounts of data generated and the key interest in questions around screening, diagnosis and prediction in these settings (78). The range and depth of applications of ML methods in the field of healthcare and bioinformatics is large, including prediction of acute kidney injury (79), the prediction of skin cancer (80) and from the prediction of diabetic retinopathy (81) to the prediction of protein folding (82). An overview of the use of ML and Deep learning in bioinformatics is given by Ravi and al (83). In the fields of TB and malaria field there have been applications of ML outside the prediction of drug resistance, including the classification of digital chest x-rays for TB (84), the detection of parasites for malaria in microscope films (85–88), the prediction of TB infection from both digital x-rays (84) and serological data (89), the classification of clinical malaria outcomes based on haematological indicators (90) and in supporting drug development for both new TB drugs (91) and new anti-malarial drugs (92).

Machine learning for WGS datasets

As mentioned, ML is an emerging tool for the analysis of WGS datasets, with the high-dimensionality of the data making the application of traditional methods more difficult, especially in settings where infection control stakeholders require updated analyses to inform decision making. There are many possible ML methods that one in principle could apply to WGS datasets. In this thesis, a pre-selection of models was made that offer: a) the ability to incorporate the features that traditionally-used methods struggle to include in a computationally efficient manner (e.g. epistatic interactions); b) a degree of interpretability and transparency, also to maximize the likelihood of adoption and uptake in clinical settings; and c) the possibility to use and apply these models in low-resource settings. With those objectives in mind, in this thesis at least two models were applied on the each of the questions of interest. The aim was to include one ML model that is highly interpretable, and at least one state-of-art model which has shown very strong performance in similar settings (but often is far more complex and less easy to interpret). The comparison between these two models moreover allows for the quantification of the extent of predictive performance one forsakes by using a simpler machine learning model.

For classification tasks where labelled data was available, decision trees (DTs) were used as the simpler model. Gradient boosted trees (GBTs) were used as the more complex model, where they are a natural transition from decision tree models and they have a strong performance on other predictive tasks (93–96). For the prediction of the geographic location of malaria isolates, (penalized) linear and logistic regression models were applied as the simpler models, and their performance was compared to more complex convolutional neural net (CNN) models. Finally, for the image-based detection of positive selection and drug resistance in the malaria genome, only a CNN ML approach was adopted, given its unique fit to the task at hand. The performance of this CNN model was compared against statistical haplotype-based methods. A comparison of the approaches adopted is summarized (**Table 4**).

Table 4: Machine learning methods used in this thesis (61)

Machine learning method	**Description**
Penalized linear regression	Penalized linear regression is an adjustment of the traditional regression approach by adding a regularization factor (either a L1 or a L2 term or a combination) to reduce the number of included features and/or shrink their coefficients in order to reduce variance (by inserting some bias) and improve the interpretability of the model (71).
Penalized logistic regression	As per above, but taking a (logistic) classification approach
Decision Trees (DTs)	Decision trees are recursive, greedy, top-down partitioning algorithms (71). Although they offer the benefit of easy interpretation, decision trees can suffer from high variance due to over-fitting of spurious features on small subsets of the data.
Gradient Boosted Trees (GBTs)	Gradient boosted trees build an ensemble of individually weak learners (often short and stumpy decision trees) in an adaptive manner by optimizing the approximation of the gradient of the loss function (71,97–99)
Convolutional neural nets (CNNs)	Convolutional neural nets are a sub-set of neural Networks that are often applied to image data. Convolutional networks incorporate a mathematical operation called convolution that facilitates the detection of features in an image in a location-invariant manner (43)

Project structure

The global disease control strategy for TB rests on accurate and cost-effective detection of drug resistant TB. Current methods of drug-resistance testing are either relatively slow, expensive, or inaccurate. ML methods that can predict known and unknown forms of drug resistance in WGS isolates may serve as a valuable complementary tool. **Chapter 2** discusses the prediction of drug-resistance and the discovery of new SNPs using ML methods for *M. tuberculosis*. DT-based approaches were applied to 8,639 *M. tuberculosis* WGS isolates that have accompanying DST data across 14 anti-TB drugs.

Chapter 3 utilizes the same dataset (as in **Chapter 2**) and develops a customized decision-tree algorithm called Treesist-TB. This algorithm enhances standard DTs by allowing the incorporation of priors and constraints on the features and sub-structures that can be included in the trees. The algorithm was subsequently applied in a new ensemble-based manner across individual studies to discover genomic variants that have support across multiple studies. The overall aim is ensuring robustness to the presence of DST errors in individual studies, which can lead to genomic variants being undetected in the analysis of aggregate datasets.

Chapter 4 discusses the development of a new method to detect selective sweeps for *P. vivax* and *P. falciparum*. The detection of these signatures has historically been performed using a wide variety of methods that require a high amount of pre-processing and domain-specific expertise to extract features, or were not optimized for application on WGS libraries. Deep learning methods were applied, which do not require feature extraction. This application appears to be the first time deep learning or image-classifier-based methods have been applied to raw WGS data to detect selective sweeps. A novel image-based deep learning approach was applied that does not require extensive feature extraction using CNNs to identify selective sweeps.

Chapter 5 discusses the prediction of the geographic origins of malaria isolates for *P. falciparum* and *P. vivax*. It explores the accuracy of both regular ML approaches as well as deep learning approaches, across regression and classification methods, to make predictions at different levels of geographic granularity.

Overall, this work shows the potential of applying ML approaches to WGS pathogen datasets to make predictions that will improve clinical and programmatic decision making. It also shows the risks of not adapting and customizing algorithms to the specific context of the pathogen in question, and the additional power that can be harnessed if adaptation and customization is performed correctly. Much of my work has been published in peer review journals. Specifically, the research papers included in this thesis include:

Research paper (chapter)	Authors	Title	Status, journal and year
2	Wouter Deelder, et. al	Machine learning predicts accurately *Mycobacterium tuberculosis* drug resistance from whole genome sequencing data	Published, Frontiers in Genetics September 2018
3	Wouter Deelder, et al.	A modified decision tree approach to improve the prediction and mutation discovery for drug resistance in Mycobacterium tuberculosis	Published, BMC Genomics January 2022
4	Wouter Deelder, et al.	Using deep learning to identify recent positive selection in malaria parasite sequence data	Published, Malaria Journal June 2021
5	Wouter Deelder, et al.	Geographical classification of malaria parasites through applying machine learning to whole genome sequence data	Submitted for publication, Nature Scientific Reports

Chapter 2

RESEARCH PAPER COVER SHEET

SECTION A – Student Details

Student ID Number	1701929	**Title**	Mr.
First Name(s)	Wouter		
Surname/Family Name	Deelder		
Thesis Title	Machine learning methods for infectious diseases: applications for tuberculosis and malaria.		
Primary Supervisor	Prof. Taane Clark, Dr. Luigi Palla		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Frontiers in Genetics		
When was the work published?	September 2018		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	**N/A**
Please list the paper’s authors in the intended authorship order:	N/A
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I supported the conception and design of the study. I helped to clean and format the genomic dataset for the machine learning analysis. I implemented the machine learning algorithms and analysed the data. I wrote the first draft of the manuscript, and finalised it after receiving revisions from co-authors and reviewers.

SECTION E

Student Signature	
Date	April 11, 2022

Supervisor Signature	
Date	April 11, 2022

frontiers
in Genetics

ORIGINAL RESEARCH
published: 26 September 2019
doi: 10.3389/fgene.2019.00922

Machine Learning Predicts Accurately *Mycobacterium tuberculosis* Drug Resistance From Whole Genome Sequencing Data

Wouter Deelder[1,2], Sofia Christakoudi[1,3], Jody Phelan[1], Ernest Diez Benavente[1], Susana Campino[1], Ruth McNerney[4], Luigi Palla[1†] and Taane G. Clark[1*†]*

[1] Faculties of Epidemiology & Population Health and Infectious & Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom, [2] Dalberg Advisors, Geneva, Switzerland, [3] Epidemiology and Biostatistics Department, Imperial College London, St Mary's Campus, London, United Kingdom, [4] Department of Medicine, University of Cape Town, Cape Town, South Africa

Background: Tuberculosis disease, caused by *Mycobacterium tuberculosis,* is a major public health problem. The emergence of *M. tuberculosis* strains resistant to existing treatments threatens to derail control efforts. Resistance is mainly conferred by mutations in genes coding for drug targets or converting enzymes, but our knowledge of these mutations is incomplete. Whole genome sequencing (WGS) is an increasingly common approach to rapidly characterize isolates and identify mutations predicting antimicrobial resistance and thereby providing a diagnostic tool to assist clinical decision making.

Methods: We applied machine learning approaches to 16,688 *M. tuberculosis* isolates that have undergone WGS and laboratory drug-susceptibility testing (DST) across 14 antituberculosis drugs, with 22.5% of samples being multidrug resistant and 2.1% being extensively drug resistant. We used non-parametric classification-tree and gradient-boosted-tree models to predict drug resistance and uncover any associated novel putative mutations. We fitted separate models for each drug, with and without "co-occurrent resistance" markers known to be causing resistance to drugs other than the one of interest. Predictive performance was measured using sensitivity, specificity, and the area under the receiver operating characteristic curve, assuming DST results as the gold standard.

Results: The predictive performance was highest for resistance to first-line drugs, amikacin, kanamycin, ciprofloxacin, moxifloxacin, and multidrug-resistant tuberculosis (area under the receiver operating characteristic curve above 96%), and lowest for third-line drugs such as D-cycloserine and Para-aminosalisylic acid (area under the curve below 85%). The inclusion of co-occurrent resistance markers led to improved performance for some drugs and superior results when compared to similar models in other large-scale studies, which had smaller sample sizes. Overall, the gradient-boosted-tree models performed better than the classification-tree models. The mutation-rank analysis detected no new single nucleotide polymorphisms linked to drug resistance. Discordance between DST and genotypically inferred resistance may be explained by DST errors, novel rare mutations, hetero-resistance, and nongenomic drivers such as efflux-pump upregulation.

Edited by:
Feng Gao,
Tianjin University,
China

Reviewed by:
Samaneh Kouchaki,
University of Oxford,
United Kingdom
Debmalya Barh,
Federal University of Minas Gerais,
Brazil
Maha Rida Farhat,
Harvard University,
United States

***Correspondence:**
Luigi Palla
Luigi.palla@lshtm.ac.uk
Taane Clark
Taane.clark@lshtm.ac.uk

†Joint Last Authors

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology,
a section of the journal Frontiers in Genetics

Received: *24 May 2019*
Accepted: *02 September 2019*
Published: *26 September 2019*

Citation:
Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, Palla L and Clark TG (2019) Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance From Whole Genome Sequencing Data. Front. Genet. 10:922. doi: 10.3389/fgene.2019.00922

Conclusion: Our work demonstrates the utility of machine learning as a flexible approach to drug resistance prediction that is able to accommodate a much larger number of predictors and to summarize their predictive ability, thus assisting clinical decision making and single nucleotide polymorphism detection in an era of increasing WGS data generation.

Keywords: ***Mycobacterium tuberculosis*, MDR-TB, XDR-TB, drug resistance, machine learning**

INTRODUCTION

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* bacteria, remains a major global public health challenge, with over 10.0 million people infected with TB and an estimated 1.6 million deaths in 2017 (World Health Organization, 2018a). An increasing prevalence of drug resistance presents a serious challenge to effective TB control (World Health Organisation, 2018b). First-line anti-TB therapy is centered around four drugs: rifampicin (RIF), isoniazid (INH), ethambutol (EMB), and pyrazinamide (PZA) (World Health Organization, 2017). *M. tuberculosis* strains resistant to at least RIF and INH are termed multidrug-resistant (MDR-TB), with >550,000 new resistant cases in 2017 (World Health Organisation, 2018b). Additional resistance to second-line drugs, the fluoroquinolones [FQ; ciprofloxacin (CIP), ofloxacin (OFL), or moxifloxacin (MOX)] and injectables [INJ; amikacin (AMK), kanamycin (KAN), capreomycin (CAP)], is termed extensively drug resistant (XDR-TB), and such cases have been reported in >115 countries (World Health Organisation, 2018b). Conventional TB treatment regimens are relatively long (>6 months) and include the simultaneous application of several drugs (World Health Organization, 2017). Treatment of drug-resistant TB is even more prolonged and involves drugs with severe side effects and with lower efficacy (World Health Organization, 2018a).

Anti-TB drugs act on *M. tuberculosis via* three main mechanisms: (i) blocking enzymes involved in the synthesis of components of the cell wall (e.g., EMB), (ii) disrupting protein synthesis at the level of the ribosomes [e.g., streptomycin (STM)] and (iii) hindering various processes at a DNA level such as RNA/DNA synthesis (e.g., RIF, FQ) (Nasiri et al., 2017). While *M. tuberculosis* drug-resistance mechanisms are not fully understood, they have been observed to be driven mainly by single nucleotide polymorphisms (SNPs) or other polymorphisms (e.g., small insertions and deletions, “indels”) resulting in the modification of drug targets (e.g., *rpoB* gene for RIF, *gidB* and *rpsL* genes for STM, *embB* gene for EMB, *gyrA* and *gyrB* genes for FQ, *rrs* gene for INJ) or in the loss of an ability to activate prodrugs (e.g., *katG* gene for INH, *pncA* gene for PZA) (Gygli et al., 2017). Mutations can be located within gene coding regions or within promoters [e.g., the *inhA* promoter for INH and ethionamide (ETH) resistance] (Palomino and Martin, 2014). A resistance mutation can directly alter drug action or be compensatory *via* activation of an alternative pathway. Mutations may cause resistance to multiple drugs and contribute to complex gene–gene interactions (Safi et al., 2013; Trauner et al., 2014; Gygli et al., 2017).

Drug resistance is traditionally diagnosed using bacterial culture and phenotypic testing, where uncovering resistance to first-line treatments leads to an assessment of second-line regimens. However, this approach is relatively slow and expensive, and it has inherent inaccuracies and reproducibility challenges (Farhat et al., 2016). Whole genome sequencing (WGS) is increasingly being used as a diagnostic tool to rapidly identify a wider set of mutations to inform clinical decision making (Dheda et al., 2017). WGS can also be used to identify new putative resistance loci, for example, through genome-wide association (GWAS) and phylogenetic-tree-based convergent evolution approaches (Coll et al., 2018). Classic regression methods, with and without the incorporation of regularization techniques, have been applied within a GWAS context to improve model generalizability and prevent model overfitting. However, these methods may fail to detect interactions among covariates and might be less suited to the analysis of large and high-dimensional datasets that arise from large-scale WGS projects (Lunetta et al., 2004; Hastie et al., 2009). This issue is of special relevance, as prior studies have indicated that there are likely to be as-yet undetected epistatic effects that might influence resistance (Farhat et al., 2016).

Machine learning is concerned with the development and application of computationally intensive analytical methods to extract information from complex datasets, with an emphasis on the task of prediction. With increasing numbers *of M. tuberculosis* clinical isolates undergoing WGS and the expanding numbers of loci implicated in resistance, machine learning offers a complementary approach to regression-based GWAS, as it has a superior capability to adapt to the growing body of clinical and biological data. Compared with regression, nonparametric machine learning methods such as classification trees (CTs) and gradient-boosted trees (GBTs) have few underlying model assumptions related to the distribution and functional relationships between the included covariates or predictors. They potentially provide greater flexibility for problems of prediction in high-dimensional variable spaces, when each individual covariate may contain limited information and covariate interactions are important (Lunetta et al., 2004; Heidema et al., 2007; Hastie et al., 2009). CTs and GBTs are recursive partitioning methods that have outperformed other classification techniques in genome-wide studies (Chen and Ishwaran, 2012) and provide predictions and the ranked importance of predictors as outputs (Efron and Hastie, 2017). GBTs in particular have achieved state-of-the-art results on many standard classification benchmarks and demonstrated scalability and speed, suggesting that they may perform well in drug-resistance studies (Chen and Guestrin, 2016). We aim to leverage the great interpretability of CTs with the superior prediction performance of GBTs.

Machine learning methods have previously been applied in a TB context, including to support digital X-ray analysis (Lakhani and Sundaram, 2017) and drug development and to assess antitubercular properties of compounds (Periwal et al., 2011). In the context of predicting pathogen drug resistance, researchers have looked to apply random forest classification and GBT models (Farhat et al., 2016; Yang et al., 2018; Kouchaki et al., 2018). For TB, different statistical models have been applied to different drugs within the same study, rather than adopting a single approach across all drugs (Kouchaki et al., 2018). Our approach differs from these and other studies in one or more of the following aspects. First, our dataset is one of the largest for TB, consisting of nearly 17,000 *M. tuberculosis* isolates sourced globally, and considers phenotypic data for a wider range of drugs (n = 14), including for less often used ones such as para-aminosalisylic acid (PAS), cycloserine (CYS), and ETH. Not only do we focus on known drug-resistance SNPs or genes, but we also analyze (640K) genome-wide SNPs with an opportunity to inform new variant discovery. Therefore, our dataset provides a unique opportunity to evaluate machine learning methods, which could be rolled in a clinical setting, based on actual *M. tuberculosis* "big data." Second, we use a combination of CTs and GBTs to optimize resistance prediction and SNP discovery (Hastie et al., 2009). Third, we assess the impact and implications of including "co-occurrent resistance" markers in the prediction models. These are mutations that are known to be causing resistance to other drugs. Furthermore, we have developed a new approach to graphically interpret and rank the results of the GBT models and propose approximate novel SNP detection thresholds, supporting the detection and interpretation of putative new SNPs linked to drug resistance. In summary, we investigate the potential of applying cutting-edge CT and GBT machine learning methods to predict drug resistance and thereby support surveillance and clinical decision making, as well as assist the discovery of putative new SNPs linked to resistance.

RESULTS

M. Tuberculosis Sequence Data, Genetic Diversity, and Drug Resistance

WGS and drug susceptibility testing data were available across 16,688 isolates (**S1 Table**), which cover the four main lineages (L1, 11.1%; L2, 21.9%; L3, 17.0%; L4, 50.1%; **S2 Table**). Across the isolates, 642,580 high-quality genome-wide SNPs were identified, with the majority in genic regions (91.6%; 56.9% of mutations leading to nonsynonymous amino acid changes). The majority of SNPs (98.9%) have low minor allele frequencies (< 1%). We also included covariates representing the aggregation of nonsynonymous mutations by locus within our machine learning approach. A phylogenetic tree constructed using all genome-wide SNPs revealed the expected clustering by lineage (**Figure 1**). The CT and GBT approaches implemented also selected lineage-specific markers to account for the phylogeographic-based population stratification.

Laboratory drug susceptibility testing (DST) of anti-TB drugs found that 35.5% of isolates had a resistance phenotype (MDR-TB, 22.5%; XDR-TB, 2.1%; other, 11.0%; **Table 1**; **S2 Table**; **S3 Table**). Due to oversampling, these rates are higher than those typically seen in clinical or surveillance settings. Fourteen drugs were included in the genome-wide analysis: INH, RIF, ETH, PZA, EMB, STM, AMK, CAP, KAN, CIP, OFL, MOX, CYS, and PAS, as well as the composite MDR-TB phenotype. Phenotypic DST data were not available for every isolate across each of the 14 drugs, as only those individuals resistant to first-line treatments are typically tested for second-line resistance. Therefore, the number of samples tested ranged from >16,000 for the most commonly tested first-line drugs (INH and RIF; ≥98.0%) to <407 (≤2.4%) for less often phenotypically assessed drugs such as PAS, CYS, and CIP (**S3 Table**). Insufficient phenotypic data were available for the inclusion of the new and repurposed drugs such as bedaquiline, delamanid, and linezolid as well as for XDR-TB.

Machine Learning Models to Predict Drug Resistance

CT and GBT approaches were used to predict drug resistance and support new SNP discovery. We fitted CT models using datasets either consisting of SNPs in genes known to be linked to drug resistance (CT-KDG) or genome wide (CT-ALL). One GBT model was fitted to datasets with all genome-wide SNPs (GBT-ALL). All of these three models (CT-KDG, CT-ALL, and GBT-ALL) excluded known co-occurrent resistance markers. We fitted one additional approach (GBT-CRM) that included all genome-wide SNPs and, therefore, potential co-occurrent resistance markers in the model. Finally, for the purpose of comparison, we fitted a logistic regression (LR) model on the SNPs in genes known to be linked to drug resistance (LR-KDG). For all approaches, we also included the aggregated count of all nonsynonymous mutations per gene in the dataset, to allow the models to use this covariate as a potential starting point and potentially cover known resistance mutations that have low frequency (Phelan et al., 2019). It should be noted that the dataset did not contain large deletions, which we have found to be present in some resistant isolates, but at very low frequency overall (Coll et al., 2018). The resulting CT-KDG models included between one and four SNPs or loci. For the CT-ALL and GBT-ALL, the number of predictors selected varied from 1 to 10 and from 30 to 134, respectively (**Table 1**), and included lineage or strain-specific markers that are not causally linked to resistance. All models overlapped with respect to known drug-resistance loci (**Table 1**), confirming that they are the strongest predictors of resistance. In some cases, the CT-KDG and CT-ALL models were identical (e.g. RIF, EMB, AMK, CAP, CIP, OFL).

The Performance of the Machine Learning Models

The predictive performance of the machine learning approaches was assessed by calculating the sensitivity and specificity and the area under the receiver operating characteristic curve (AUC), assuming the laboratory DST result was the gold standard (**Table 2**). The GBT-CRM sensitivity for RIF (88.8%) and INH (91.1%) was higher than for EMB (82.8%) and PZA (69.7%). The sensitivity

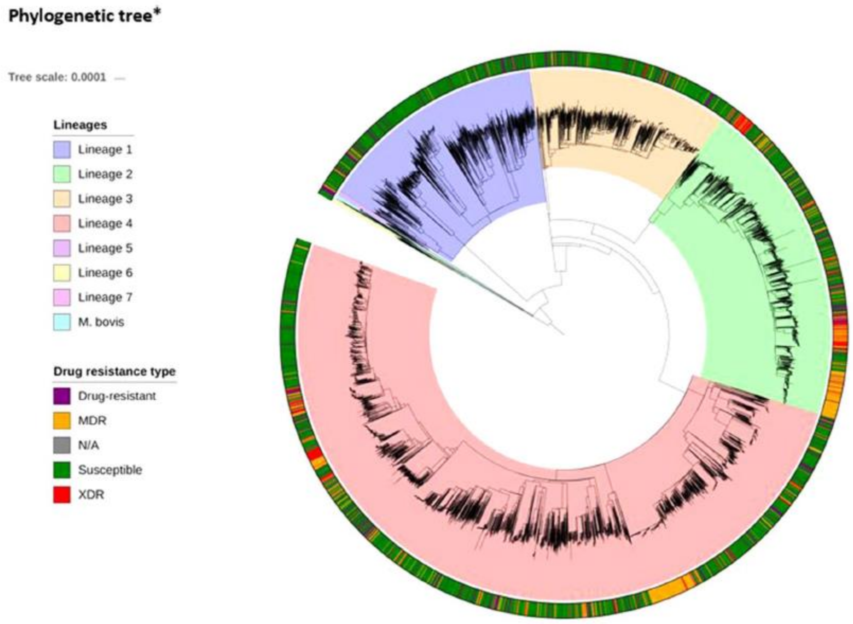


FIGURE 1 | Phylogenetic tree*(attached as separate file)* The tree includes all 16,688 isolates, complemented by additional data from lineages 5–7 and *M. bovis*. The tree was fitted using a maximum likelihood approach implemented in RAxML (Stamakis, 2014).

TABLE 1 | Drug-resistance loci identified in the machine learning models.

Drug	*N*	Resistant	%	CT-KDG (N)	CT-ALL (N)	GBT-ALL (N)	Overlapping Loci
Isoniazid	16,422	5,215	31.8	2	5	103	*katG*, fabG*
Rifampicin	16,507	4,462	27.0	1	1	39	*rpoB**
Pyrazinamide	11,968	1,813	15.1	2	4	116	*pncA*
Ethambutol	14,830	2,576	17.4	1	10	36	*embB**
Streptomycin	5,213	1,338	25.7	4	4	134	*rpsL*, rpsl, rrs*, rrs*
Amikacin	1,435	335	23.3	1	1	35	*rrs*
Capreomycin	1,731	389	22.5	1	3	44	*rrs*
Kanamycin	1,843	639	34.7	1	2	43	*rrs*
Ciprofloxacin	400	63	15.8	1	1	30	*gyrA**
Ofloxacin	1,993	506	25.4	1	1	42	*gyrA**
Moxifloxacin	885	104	11.8	1	2	36	*gyrA**
Ethionamide	940	329	35.0	3	1	60	*fabG**
Cycloserine	391	105	26.9	1	5	44	*alr*
PAS	407	43	10.6	1	1	54	*folC*
MDR-TB	–	3748	22.5	1	1	82	*rpoB*, katG, fabG*

*PAS, para-aminosalisylic acid; CT-KDG is a classification tree (CT) applied to a dataset with SNPs that are known to be associated with drug resistance [derived from Ref. (Phelan et al., 2019)]; CT-ALL and GBT-ALL are, respectively, a CT and gradient boosted tree (GBT) applied to a dataset that includes all genome-wide SNPs, except those linked to resistance for other drugs (co-occurrent resistance markers); GBT-CRM is a GBT that is applied to all genome-wide SNPs; MDR-TB is multidrug resistant TB, that is, resistance to isoniazid and rifampicin. *Total number of nonsynonymous mutations in that gene.*

for fluoroquinolones was highest for CIP (85.7%), followed by OFL (81.0%) and MOX (53.3%). The sensitivity for the injectables was highest for KAN (82.2%), followed by AMK (80.5%) and CAP (74.6%). The model sensitivity for the remaining drugs [ETH (68.1%), CYS (50.0%), and PAS (20.0%)] is substantially lower. The overall sensitivity for MDR-TB was 90.4%. The GBT-ALL model tended to outperform the CT models, with respect to sensitivity and specificity, and CT-ALL had stronger performance than

TABLE 2 | Sensitivity, specificity, and accuracy for the models (maximum value per prediction measure is bolded).

Drug	LR-KDG			CT-KDG			CT-ALL			GBT-ALL			GBT-CRM		
	Sens.	**Spec**	**Acc.**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**	**Acc**
INH	87.3	**99.1**	95.3	87.3	**99.1**	95.3	87.3	**99.1**	95.3	88.0	99.0	95.4	**91.1**	98.8	**96.3**
RIF	82.8	**99.6**	95.1	82.8	**99.6**	95.1	82.8	**99.6**	95.1	82.8	**99.6**	95.1	**88.8**	98.9	**96.2**
PZA	21.6	**100**	87.2	21.6	**100**	87.2	35.2	98.5	88.2	42.8	99.2	90.0	**69.7**	96.1	**91.8**
EMB	**84.7**	93.1	91.6	80.9	94	91.6	80.9	94.0	91.6	81.7	**94.7**	**92.4**	82.8	94.2	92.1
STM	71.6	**97.8**	91.1	72.3	96.5	90.3	71.2	97.3	90.6	72.3	97.3	90.9	**79.8**	96.0	**91.9**
AMK	**80.5**	**99.5**	**95.1**	**80.5**	**99.5**	**95.1**	**80.5**	**99.5**	**95.1**	**80.5**	**99.5**	**95.1**	**80.5**	**99.5**	**95.1**
CAP	69.6	95.5	89.6	69.6	95.5	89.6	69.6	95.5	89.6	72.1	95.8	90.4	**74.6**	**96.2**	**91.3**
KAN	74.4	**99.1**	89.7	74.4	**99.1**	89.7	**82.2**	97.8	91.8	80.8	97.8	91.3	**82.2**	98.2	**92.1**
CIP	**92.8**	**98.5**	**97.5**	**92.8**	**98.5**	**97.5**	**92.8**	**98.5**	**97.5**	85.7	**98.5**	96.2	85.7	**98.5**	96.2
OFL	80	**97.7**	93.5	80.0	**97.7**	93.5	80.0	**97.7**	93.5	**81.0**	**97.7**	**93.7**	**81.0**	97.0	93.2
MOX	**66.6**	93.2	90.9	**66.6**	93.2	90.9	46.6	**98.1**	**93.7**	53.3	96.2	92.6	53.3	97.5	**93.7**
ETH	**75.7**	75.6	75.6	**75.7**	75.6	75.6	74.2	79.6	77.7	66.6	92.6	83.5	68.1	**93.4**	**84.6**
CYS*	**57.6**	88.6	**78.4**	38.4	**98.1**	**78.4**	30.7	94.3	73.4	46.1	92.4	77.2	50.0	92.4	**78.4**
PAS	0	**100**	87.8	**20.0**	**100**	**90.2**	0	**100**	87.8	10.0	**100**	89.0	**20.0**	**100**	**90.2**
MDR	85.9	96.9	94.4	85.9	96.9	94.4	85.9	96.9	94.4	86.2	**97.5**	95.0	**90.4**	96.9	**95.5**

**No known drug-resistance SNPs for CYS were included in the KDG models; reported outcomes are the performance on the test set; RIF, rifampicin; INH, isoniazid; EMB, ethambutol; PZA, pyrazinamide; CIP, ciprofloxacin; OFL, ofloxacin; MOX, moxifloxacin; AMK, amikacin; KAN, kanamycin; CAP, capreomycin; PAS, para-aminosalisylic acid (PAS); CYS, cycloserine; ETH, ethionamide; CT-KDG is a classification tree (CT) fitted to a dataset with SNPs that are known to be associated with drug resistance [derived from Ref. (Phelan et al., 2019)]; LR-KDG is a logistic regression model applied to the same SNP set as CT-KDG; CT-ALL and GBT-ALL are, respectively, a CT and gradient boosted tree (GBT) applied to a dataset that includes all genome-wide SNPs, except those linked to resistance for other drugs (co-occurrent resistance markers); GBT-CRM is a GBT that is applied to all genome-wide SNPs; MDR is multidrug resistant TB.*

CT-KDG. The AUC values for most major first- and second-line drugs for the GBT model were above 90% (and often above 95%) (**S4 Table**). The overall predictive performance across models for CYS and PAS was relatively weak. In general, larger datasets with well-characterized PAS and CYS phenotypes will be needed to assist with identifying the full repertoire of related resistance mutations (Farhat et al., 2016; Coll et al., 2018).

Comparison Between GBT-CRM and Other Machine Learning Models

Owing to the inclusion of co-occurrent resistance markers, the GBT-CRM model was almost always the best in terms of predictive accuracy and AUC, with a marked improvement for PZA and PAS (**S1 Table**). The GBT-ALL model, which excludes co-occurrent resistance markers, but can include marker interactions and strain markers, also tended to outperform the KDG models, but to a lesser extent than GBT-CRM. The difference in predictive performance between the GBT-ALL and the KDG models was especially large for ETH and CYS.

Comparison With an in *Silico* Panel of Known Mutations and GWAS

We also compared the predictive abilities of GBT-ALL, CT-ALL, and CT-KDG models to those from the TB-Profiler mutation panel consisting of >1,300 markers across the 14 drugs (**S5 Table**) (Coll et al., 2015; Phelan et al., 2019). First, we used only those markers with minor allele frequency of >0.5% to predict resistance ("TB Panel"; **S6 Table**) and attained a performance similar to KDG models (**Table 2**). We then used the TB-Profiler (full) mutation panel and software (Phelan et al., 2019), which rules in observed frameshift mutations, large deletions, and missense mutations in known resistance genes. As TB-Profiler includes mutations occurring at low frequencies, the predicted accuracy was superior than the machine learning approaches for most drugs. For five drugs, where the resistance mechanisms are less understood, including STM, ETH, and PAS, the GBT-CRM model had a marginally better performance than the TB-Profiler (**S6 Table**). We also compared the predictive abilities of the GBT-CRM to those from an updated GWAS analysis [similar implementation to (Coll et al., 2018)] (**S6 Table**). Overall, the accuracy of both models was in the same range (<1% difference) for most drugs, with the exception for CAP, KAN, and CYS, where the performance of GWAS was distinctively greater, and with exception for PZA, MOX, and ETH, where the performance of GBT-CRM was better.

Comparison With Other Studies That Apply Machine Learning Methods

We compared our models to the results of four recent studies that have applied different machine learning models (Yang et al., 2018; Kouchaki et al., 2018; Chen et al., 2019; Yang et al., 2019). Specifically, we compared both the average and maximum of the reported results for each metric (sensitivity, specificity, AUC) for each drug across the four studies (**S7 Table**; **S8 Table**). All the comparator studies included co-occurrent resistance markers. The specificities tended to be greater for the GBT-CRM model. The sensitivities tended to be greater for one or more of the models used in the other studies. However, overall, for six drugs (PZA, AMK, CAP, KAN, CIP, and MOX), the AUC scores of the GBT-CRM were higher than for the best model for that specific drug in other studies.

Detection and Interpretation of Putative New SNPs

The CT-ALL and GBT-based approaches did not discover any putative new SNPs that met the stringent detection thresholds. We present and display a new visual approach to mutation ranking that leverages the output of the GBT-ALL model (**S1 Fig**). A number of known candidates (e.g., Rv1463 for RIF resistance) presented with marginal evidence.

DISCUSSION

With the rollout of WGS-based TB diagnosis across many countries (including UK) (PHE, 2018), there is a need to develop global TB datasets and databases (Coll et al., 2018; ReSeqTB, 2018), which in turn will require the implementation of "big data" analytical approaches (e.g., machine learning methods) to assist clinical and control program decision making. We have shown that CT and GBT machine learning approaches can play a value-adding role in predicting drug resistance and the possible detection of new putative variants. In general, the predictive performance of the CT models was inferior to the GBT approaches, but they captured the most common mutations driving resistance. When using aggregated counts of nonsynonymous mutations in known resistance genes as a predictor in the trees, the CT models did not include any known individual SNPs in that respective gene in an exclusionary manner as an additional predictor. This observation provides not only support for the validity and accuracy of the overall TB-Profiler lists but also the use of aggregation as a first parse approach to identifying relevant genes. The possible exception relates to KAN, CAP, and AMK, where the machine learning models chose a subset of the list of TB-Profiler SNPs.

The predictive performance of the GBT models, and especially the GBT-CRM model, is similar or higher than that of the models developed in other studies (Yang et al., 2018; Kouchaki et al., 2018; Chen et al., 2019; Yang et al., 2019). The performance of the more complex GBT models (GBT-ALL and GBT-CRM) in some cases is worse than TB-profiler (Phelan et al., 2019), but the comparison is affected by the fact that the latter approach uses rare alleles and deletions for prediction. For some drugs where the resistance mutations are not fully established (e.g., CYS, STM, and PAS), the GBT-CRM model had a similar or better predictive performance to the TB-profiler panel. The improved performance of the GBT-CRM over GBT-ALL and CT models may be explained by its ability to capture covariate interactions and the inclusion of co-occurrent resistance markers and strain-specific SNPs that may be informative in resistance outbreaks but in themselves may be related to transmissibility and not drug resistance. The inclusion of co-occurrent resistance markers might lead to overoptimism in the estimated performance that may not translate optimally into clinical practice. This optimism bias affects both prediction as well as detection (i.e., through mutation ranking) and may be caused by an interplay between high DST measurement errors (e.g., for pyrazinamide) (APHL, 2016), sequential testing, data from settings where drug availability is unregulated, the structure and stratification of the datasets, and differential resistance mechanisms not captured in a database (e.g., Lisboa strain types which have different MDR-TB mutations) (Coll et al., 2018). Ideally, resistance predictions should be based on underlying biological mechanisms, with co-occurring mutations having little effect, thereby assisting with the identification of novel putative markers and pathways. While our machine learning analysis suggested no novel SNPs at the importance thresholds used, in general, the approach ranks the informativeness of SNP mutations, which assists the detection of novel polymorphisms. As databases get larger with greater numbers of well-characterized resistance samples, especially for third-line drugs, there is improved potential to identify novel resistance mutations using machine learning approaches.

As expected, the overall predictive ability of INH, RIF, and MDR-TB resistance across the machine learning approaches was high (~90% sensitivity) because the underlying mutations and loci involved are well established. However, 10% of resistance cases were not identified by the models. The genotypic–phenotypic discordance, as measured with the GBT-ALL model, was higher for other first-line (e.g., EMB, ~20% and PZA, ~60%) and second-line drugs (AMK and CAP, ~20–25%; ETH, ~35%; CYS, ~55%), and large discrepancies point towards unknown genetic factors. However, other factors potentially have an effect, including laboratory DST errors or misspecified or truncated drug assay breakpoints (World Health Organization, 2018c), efflux-pump upregulation (Balganesh et al., 2012; Gygli et al., 2017), and epigenetic and hetero-resistance effects (Folkvardsen et al., 2013; Farhat et al., 2016). For example, the recent downward revision of the critical concentrations for the fluoroquinolones and injectables is likely to decrease specificity and increase sensitivity of WGS-based analysis (World Health Organization, 2018c). Future studies should aim to use quantitative minimum inhibitory concentration scores as phenotypes (Farhat et al., 2018). For heteroresistance, both resistance and wild-type mutations occur in a mixed infection. If the resistant strain has a relatively low abundance, the drug may be labeled resistant according to the DST result but sensitive in genomic sequencing (Folkvardsen et al., 2013; Farhat et al., 2016), leading to false negative results. Across the 32 drug targets in the TB-Profiler mutation library, 28 appear to have some evidence of heteroresistance within the 17k dataset (Phelan et al., 2019). With the lower error rates and higher depth of WGS, the detection of such low frequency variants is possible; therefore, combined with robust bioinformatic approaches, sequencing is being viewed as the gold standard for drug resistance characterization (Coll et al., 2018).

In summary, our approach has shown that machine learning can robustly predict drug resistance and inform on its underlying mutations. Furthermore, such approaches will be scalable when WGS becomes routine and increasingly "big data" analyses are required.

MATERIALS AND METHODS

Phenotypic and Sequencing Data

The dataset consists of 16,688 isolates (lineages 1–4) with WGS data and phenotypic DST data (see **S1 Table** for accession numbers). The laboratory drug susceptibility testing followed

WHO recommended protocols and practice [see Ref. (Coll et al., 2018)]. The raw sequence data were mapped to the H3Rv reference genome using *bwa-mem* software, and SNPs and insertions and deletions (indels) called from the consensus of GATK and *samtools* software. The final set of SNPs ($N = 642,580$) and indels included those with low levels of missing genotypes (<2%) and excluded those in the hypervariable PE/PPE gene families. Missing values were imputed using a nearest neighbor imputation approach. The dataset was augmented with covariates that aggregated the number of nonsynonymous mutations isolated in a locus.

Fitting the Machine Learning Models

CTs (Hastie et al., 2009) were created from two SNP sets: one based on those in known drug resistance genes (Coll et al., 2015) ($N = 1,421$ SNPs; "CT-KDG") and the other using all SNPs in the dataset ($N = 641,159$, "CT-ALL"). CT algorithms produce only one easy to interpret tree as output. GBT models (Friedman, 2000; Hastie et al., 2009) were fitted to a genome-wide SNP dataset (GBT-ALL), leading to an ensemble of short and stumpy decision trees constructed in an adaptive manner. The GBT models allowed us to move beyond binary inclusion of SNPs in the final model and assess, for the purpose of SNP discovery, the weight and importance of the SNPs included. The LR model was applied to the same set of SNPs as the CT-KDG model. As mentioned, we excluded known resistance markers for drugs that were not the phenotype of interest in each individual model in the logistic regression LR-KDG, CT-KDG, CT-ALL, and GBT-ALL, but included these markers in the GBT-CRM approach.

We created a split in the dataset where 80% was used as a training and validation set, and 20% was used as a test set. We applied five-fold cross-validation to the training set to calculate the prediction accuracy and used this to select the maximum depth parameter of the CT and GBT models. (Hastie et al., 2009). The penalized LR model was cross-validated on the regularization strength C for the L1 penalty. The final models were trained on the training set and were subsequently applied to the test set, with those outcomes reported in the Results section. For the CT models, the maximum depth parameter was selected as the smallest value that was within one standard error from the best performing maximum depth setting. We followed this "one-standard-error" rule to further induce the selection of parsimonious models and to mitigate the risk of over-fitting (Hastie et al., 2009). In both the GBT and CT models, the predictions in the final leaf nodes of the tree were determined by the majority class in those nodes. The reported scores (sensitivity, specificity, accuracy, positive predicted value, negative predicted value, and AUC) were calculated after fitting the model to the training dataset with the maximum depth as described per above and other parameter values (described in **S9 Table**). The GBT models are based on an ensemble of 50 trees (to facilitate a consistent comparison across drugs with regards to the mutation ranking) with a subsampling of 60% of isolates to fit each tree. These models provide a score for weight, coverage, and importance. The "weight" refers to the number of times a feature (covariate) appears in a tree/forest; "coverage" is the relative quantity of observations affected by a feature (which would be higher for covariates that are higher up in the tree), and "importance" is the average gain in the predictive accuracy when a SNP is chosen to split a tree node. SNP discovery using GBTs was assisted by construction of a two-dimensional mutation-ranking graph (see **S1 Figure**) displaying importance gain versus weight, with coverage as the bubble size. Those SNPs with high importance and weight are more likely to be predictive in a large number of trees across different subsamples of the data and, therefore, more generalizable. The suggested thresholds for the importance and weight were chosen pragmatically based on the inclusion of known and established resistance markers. These thresholds are shown as dotted lines on the graphs (**S1 Figure**).

The core packages used in the analysis included the *SHAP* (Lundberg and Lee, 2017) to visualize the relative contribution of each predictor, the decision tree classifier in *sklearn* (version 0.19.1), and the *Xgboost* implementation (version 0.70) was used to construct the CTs and GBTs (Chen and Guestrin, 2016). The default settings were used for the implementation of these machine learning algorithms, with the exception of the parameters as specified (see **S9 Table**). The plausibility of putatively causal SNPs identified was assessed through a search of the literature, including for gene function on *Mycobrowser* (Kapopoulou et al., 2011).

Comparisons to Mutation Libraries, GWAS, and Other Studies.

We compared our machine learning prediction results to those from using a set of known SNPs associated with drug resistance on a rule-in basis. A first comparison was made with predictions based on mutations in the TB-Profiler panel (Phelan et al., 2019) that were common (minor allele frequency > 0.5%) in our dataset (TB-Panel). A second comparison was made with the application of the TB-Profiler software and its full mutation library (Phelan et al., 2019) to the dataset. We also compared our results to the application of a mixed-model regression GWAS approach (Coll et al., 2018) to the ~17k dataset, as well as other studies that applied machine learning methods (Yang et al., 2018; Kouchaki et al., 2018; Chen et al., 2019; Yang et al., 2019).

DATA AVAILABILITY STATEMENT

The raw whole genome sequencing data is available from the European Nucleotide Archive (ENA) (**S1 Table**). The computing code is available upon request from the corresponding authors.

AUTHOR CONTRIBUTIONS

WD, SCa, RM, LP, and TC conceived and designed the study. JP and EB performed the bioinformatic processing of the raw sequencing data and phenotypic data. WD performed the statistical analysis, under the supervision of LP and TC. SCa performed a statistical analysis on a subset of the data, under the supervision of RM and TC. WD wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

FUNDING

JP is supported by a Newton Institutional Links Grant (British Council) (261868591). SCa is funded by Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and MR/R020973/1). TC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1).

ACKNOWLEDGMENTS

The authors would like to thank Claudio Köser for sharing his thoughts and insights on drug susceptibility testing. We gratefully acknowledge the availability of the Medical Research Council UK funded eMedLab (HDR UK) computing resource.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00922/full#supplementary-material

REFERENCES

APHL. (2016). Issues in *Mycobacterium tuberculosis* complex (MTBC) drug susceptibility testing: pyrazinamide (PZA). Available from: https://www.aphl.org/aboutAPHL/publications/Documents/ID-PZA_WhitePaper_0216.pdf.

Balganesh, M., Dinesh, N., Sharma, S., Kuruppath, S., Nair, A. V., and Sharma, U. (2012). Efflux pumps of *Mycobacterium tuberculosis* play a significant role in antituberculosis activity of potential drug candidates. *Antimicrob. Agents Chemother.* 56 (5), 2643–2651. doi: 10.1128/AAC.06003-11

Chen, T., and Guestrin, C. (2016). "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD* (New York, USA: ACM Press), 785–794. doi: 10.1145/2939672.2939785

Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics* 99 (6), 323–329. doi: 10.1016/j.ygeno.2012.04.003

Chen, M. L., Doddi, A., Royer, J., Freschi, L., Schito, M., Ezewudo, M., et al. (2019). Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. *EBioMedicine* 43, 356–369. doi: 10.1016/j.ebiom.2019.04.016

Coll, F., McNerney, R., Preston, M. D., Guerra-Assunção, J. A., Warry, A., Hill-Cawthorne, G., et al. (2015). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* 7 (1), 51. doi: 10.1186/s13073-015-0164-0

Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., et al. (2018). Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis. Nat. Genet.* 50 (2), 307–316. doi: 10.1038/s41588-017-0029-0

Dheda, K., Gumbo, T., Maartens, G., Dooley, K. E., McNerney, R., Murray, M., et al. (2017). The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir. Med.* 5 (4), 291–360. doi: 10.1016/S2213-2600(17)30079-6

Efron, B., and Hastie, T. (2017). Computer age statistical inference algorithms, evidence, and data science. Available from: https://www.ams.org/journals/bull/0000-000-00/S0273-0979-2018-01611-X/.

Farhat, M. R., Sultana, R., Iartchouk, O., Bozeman, S., Galagan, J., Sisk, P., et al. (2016). Genetic determinants of drug resistance in *Mycobacterium tuberculosis* and their diagnostic value. *Am. J. Respir. Crit. Care Med.* 194 (5), 621–630. doi: 10.1164/rccm.201510-2091OC

Farhat, M. R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C. J. (2018). Genome wide association with quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals novel resistance genes and regulatory regions. *Nat. Commun.* 10 (1), 2128. doi: 10.1038/s41467-019-10110-6

Folkvardsen, D. B., Svensson, E., Thomsen, VØ, Rasmussen, E. M., Bang, D., Werngren, J., et al. (2013). Can molecular methods detect 1% isoniazid resistance in *Mycobacterium tuberculosis*? *J. Clin. Microbiol.* 51 (5), 1596–1599. doi: 10.1128/JCM.00472-13

Friedman, J. (2000). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Gygli, S. M., Borrell, S., Trauner, A., and Gagneux, S. (2017). Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. *FEMS Microbiol. Rev.* 41 (3), 354–373. doi: 10.1093/femsre/fux011

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer. Available from: http://link.springer.com/10.1007/978-0-387-84858-7.

Heidema, A. G., Feskens, E. J. M., Doevendans, P. A. F. M., Ruven, H. J. T., van Houwelingen, H. C., Mariman, E. C. M., et al. (2007). Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs. *Genet. Epidemiol.* 31 (8), 910–921. doi: 10.1002/gepi.20251

Kapopoulou, A., Lew, J. M., and Cole, S. T. (2011). The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis* 91 (1), 8–13. doi: 10.1016/j.tube.2010.09.006

Kouchaki, S., Yang, Y., Walker, T. M., Walker, A. S., Wilson, D. J., Peto, T. E. A., et al. (2018). Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics* 35 (13), 2276–2282. doi: 10.1093/bioinformatics/bty949

Lakhani, P., and Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 284 (2), 574–582. doi: 10.1148/radiol.2017162326

Lundberg, S. M., and Lee, S.-I. (2017). Consistent feature attribution for tree ensembles. Available from: https://arxiv.org/abs/1706.06060.

Lunetta, K. L., Hayward, L. B., Segal, J., and Van Eerdewegh, P. (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* 5 (1), 32. doi: 10.1186/1471-2156-5-32

Nasiri, M. J., Haeili, M., Ghazi, M., Goudarzi, H., Pormohammad, A., Imani Fooladi, A. A., et al. (2017). New insights in to the intrinsic and acquired drug resistance mechanisms in Mycobacteria. *Front. Microbiol.* doi: 10.3389/fmicb.2017.00681

Palomino, J. C., and Martin, A. (2014). Drug resistance mechanisms in *Mycobacterium tuberculosis. Antibiot. (Basel, Switzerland)* 3 (3), 317–340. doi: 10.3390/antibiotics3030317

Periwal, V., Rajappan, J. K., Jaleel, A. U., and Scaria, V. (2011). Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. *BMC Res. Notes* 4 (1), 504. doi: 10.1186/1756-0500-4-504

PHE. (2018). Annual report: Tuberculosis in England. 2018. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/742782/TB_Annual_Report_2018.pdf.

Phelan, J. E., O'Sullivan, D. M., Machado, D., Ramos, J., Oppong, Y. E. A., Campino, S. (2019). Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* 11 (1), 41. doi: 10.1186/s13073-019-0650-x

ReSeqTB. (2018). Integrating standardized whole genome sequence analysis with a global Mycobacterium tuberculosis antibiotic resistance knowledgebase. *Sci Rep.* 8 (1), 15382. doi: 10.1038/s41598-018-33731-1

Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M. (2013). Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-Arabinose biosynthetic and utilization pathway genes. *Nat. Genet.* 45 (10), 1190–1197. doi: 10.1038/ng.2743

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30 (9), 1312-3. doi: 10.1093/bioinformatiosbtu033

Trauner, A., Borrell, S., Reither, K., and Gagneux, S. (2014). Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. *Drugs* 74 (10), 1063–1072. doi: 10.1007/s40265-014-0248-y

World Health Organization. (2017). DS TB Treatment Factsheet.

World Health Organization. (2018a). Tuberculosis Factsheet. Available from: http://www.who.int/en/news-room/fact-sheets/detail/tuberculosis.

World Health Organisation. (2018b). What is multidrug-resistant tuberculosis (MDR-TB) and how do we control it?

World Health Organization (2018c). *Technical Report on critical concentrations for drug susceptibility testing of medicines used in the treatment of drug-resistant tuberculosis*. Geneva: World Health Organization.

Yang, Y., Niehaus, K. E., Walker, T. M., Iqbal, Z., Walker, A. S., Wilson, D. J. (2018). Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* 34 (10), 1666–1671. doi: 10.1093/bioinformatics/btx801

Yang, Y., Walker, T. M., Walker, A. S., Wilson, D. J., Peto, T. E. A., Crook, D. W. (2019). DeepAMR for predicting co-occurrent resistance of *Mycobacterium tuberculosis*. *Bioinformatics* 34 (10), 1666–1671. doi: 10.1093/bioinformatics/btx801

Conflict of Interest: Author WD was employed by the company Dalberg Advisors in Switzerland. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

S1 Table

Sources of sequence data, and drug resistance phenotypes

Project	N	Susceptible	Drug-resistant	MDR-TB	XDR-TB
Mixed	8128	5005	744	2379	0
PRJNA282721	1840	1471	277	87	5
PRJEB2794	1306	1219	79	8	0
PRJEB7056	1088	874	173	41	0
PRJEB9680	1031	710	73	246	2
PRJEB10385	682	98	193	296	95
PRJEB2221	356	331	19	6	0
PRJNA183624	331	85	41	138	67
PRJEB2358	325	293	30	2	0
PRJEB7669	232	0	3	218	11
PRJNA235852	208	155	33	20	0
PRJEB5162	191	175	14	2	0
PRJNA187550	157	43	0	91	23
PRJEB11653	126	14	77	35	0
PRJNA200335	126	23	5	43	55
PRJEB14199	123	0	34	14	75
PRJEB2777	98	98	0	0	0
PRJEB7281	95	38	15	41	1
PRJEB6945	46	46	0	0	0
PRJEB2424	45	3	2	40	0
PRJEB15857	38	18	5	15	0
PRJEB2138	37	9	4	14	10
PRJNA49659	30	30	0	0	0
PRJEB7727	28	12	11	5	0
PRJNA376471	18	11	0	7	0
PRJEB6276	3	3	0	0	0
Total	*16688*	*10764*	*1832*	*3748*	*344*

* https://www.ebi.ac.uk/ena; Drug-resistant refers to non-MDR-TB/-XDR-TB resistance; MDR-TB is defined as resistance to isoniazid and rifampicin; XDR-TB is defined as MDR-TB, and resistance to any fluoroquinolone, and to any of the three second-line injectables (amikacin, capreomycin, and kanamycin).

S2 Table

Phenotypic drug susceptibility tests status by lineage

Lineage	N	%	Susceptible	Drug-resistant	MDR-TB	XDR-TB
1	1851	11.1	1492	203	150	6
2	3653	21.9	1445	479	1572	157
3	2830	17.0	2162	215	425	28
4	8354	50.1	5665	935	1601	153
Overall	16688	100.0	10764	1832	3748	344
			64.5%	11.0%	22.5%	2.1%

Drug-resistant refers to non-MDR-TB/-XDR-TB resistance; MDR-TB is defined as resistance to isoniazid and rifampicin; XDR-TB is defined as MDR-TB, and resistance to any fluoroquinolone, and to any of the three second-line injectables (amikacin, capreomycin, and kanamycin).

S3 Table

Phenotypic drug susceptibility testing results

Drug	No. tests	% of 16,688	Resistant	%
Rifampicin	16507	98.9	4462	27.0
Isoniazid	16422	98.4	5215	31.8
Ethambutol	14830	88.9	2576	17.4
Pyrazinamide	11968	71.7	1813	15.1
Streptomycin	5213	31.2	1338	25.7
Ofloxacin	1993	11.9	506	25.4
Kanamycin	1843	11.0	639	34.7
Capreomycin	1731	10.4	389	22.5
Amikacin	1435	8.6	335	23.3
Ethionamide	940	5.6	329	35.0
Moxifloxacin	885	5.3	104	11.8
PAS	407	2.4	43	10.6
Ciprofloxacin	400	2.4	63	15.8
Cycloserine	391	2.3	105	26.9

PAS = para-aminosalisylic acid

S4 Table

Predictive accuracy and Area under the ROC Curve (AUC) for models (maximum value per prediction measure is bolded)

Drug	**LR-KDG**			**CT-KDG**			**CT-ALL**			**GBT-ALL**			**GBT-CRM**		
	NPV	**PPV**	**AUC**	**NPV**	**PPV**	**AUC**	**NPV**	**PPV**	**AUC**	**NPV**	**PPV**	**AUC**	**NPV**	**PPV**	**AUC**
INH	94.2	**97.9**	93.7	94.2	**97.9**	93.4	94.2	**97.9**	93.4	94.5	97.7	95.8	**95.8**	97.4	**96.7**
RIF	94.1	**98.7**	91.2	94.1	**98.7**	91.2	94.1	**98.7**	91.2	94.1	**98.7**	95.3	**96.0**	96.8	**97.9**
PZA	86.7	**100.0**	60.7	86.7	**100.0**	60.8	88.6	82.8	73.7	89.8	91.8	87.0	**94.2**	78.0	**95.5**
EMB	**96.5**	72.9	89.9	95.7	74.5	87.4	95.7	74.5	87.4	95.9	**77.2**	94.0	96.1	75.6	**95.8**
STM	90.9	**91.8**	87.3	91.0	87.7	87.1	90.8	90.0	88.4	91.1	90.2	92.2	**93.3**	87.3	**94.0**
AMK	**94.4**	**98.1**	91.1	**94.4**	**98.1**	90.0	**94.4**	**98.1**	90.0	**94.4**	**98.1**	94.5	**94.4**	**98.1**	**96.4**
CAP	91.4	82.0	84.0	91.4	82.0	82.5	91.4	82.0	82.5	92.1	83.8	90.2	**92.8**	**85.5**	**93.4**
KAN	86.3	**98.1**	88.5	86.3	**98.1**	86.7	89.9	95.8	90.4	89.2	95.7	92.9	**90.0**	96.6	**96.8**
CIP	**98.5**	**92.8**	95.6	**98.5**	**92.8**	95.6	**98.5**	**92.8**	95.6	97.0	92.3	92.9	97.0	92.3	**99.7**
OFL	94.0	91.5	88.8	94.0	91.5	88.8	94.0	91.5	88.8	**94.3**	**91.6**	92.0	94.2	89.5	**93.3**
MOX	**96.7**	47.6	79.9	**96.7**	47.6	79.9	95.2	**70.0**	93.0	95.7	57.1	**97.4**	95.7	66.6	97.2
ETH	**85.3**	62.5	79.7	**85.3**	62.5	75.6	85.2	66.2	80.4	83.8	83.0	85.4	84.5	**84.9**	**88.4**
CYS*	-	-	-	**-**	-	**-**	73.5	72.7	69.8	77.8	75.0	80.6	**79.0**	**76.5**	**83.8**
PAS**	87.8	**-**	50.0	**90.0**	**100.0**	60.0	87.8	**-**	67.9	88.9	**100.0**	**82.6**	**90.0**	**100.0**	82.5
MDR	96.0	88.9	96.5	96.0	88.9	91.4	96.0	88.9	91.4	96.0	**91.0**	97.1	**97.2**	89.5	**97.4**

* No known drug resistance SNPs for CYS were included in the KDG models; CT-KDG is a classification tree (CT) fitted to a dataset with SNPs that are known to be associated with drug resistance (derived from (24)); LR-KDG is a Logistic Regression model applied to the same SNP set as CT-KDG; CT-ALL and GBT-ALL are respectively a CT and Gradient Boosted Tree (GBT) applied to a dataset that includes all genome-wide SNPs, except those linked to resistance for other drugs ("co-occurrent resistance markers"); GBT-CRM is a GBT that is applied to all genome-wide SNPs; PPV=Positive Predicted Value, NPV=Negative Predicted Value, AUC=Area under the ROC Curve; ** PPV for PAS for LR-KDG and CT-ALL could not be calculated as sensitivity was 0. Reported outcomes are the performance on the test-set; RIF=rifampicin, INH=isoniazid, EMB=ethambutol, PZA=pyrazinamide, CIP=ciprofloxacin,OFL= ofloxacin, MOX=moxifloxacin, AMK=amikacin, KAN=kanamycin, CAP=capreomycin, PAS=para-aminosalisylic acid (PAS), CYS=cycloserine, ETH=ethionamide; MDR is multi-drug resistant TB.

S5 Table

Summary of drugs and mutations in *TBProfiler* library* used in this study

Drug	Locus	Gene	SNPs	Indels
Rifampicin	Rv0667	*rpoB*	94	25
	Rv0668	*rpoC*	8	-
Isoniazid	*Rv1483*	*fabG1*	11	-
	Rv1484	*inhA*	13	-
	Rv1908c	*katG*	226	37
	Rv2245	*kasA*	4	-
	Rv2428	*ahpC*	21	-
Ethambutol	*Rv1267c*	*embR*	20	-
	Rv3793	*embC*	25	-
	Rv3794	*embA*	9	6
	Rv3795	*embB*	127	1
Pyrazinamide	*Rv1630*	*rpsA*	3	-
	Rv2043c	*pncA*	280	87
	Rv3601c	*panD*	10	1
Streptomycin	*Rv0682*	*rpsL*	16	-
	Rv3919c	*gid*	2	26
	rrs	*rrs*	19	-
Ethionamide	*Rv1483*	*fabG1*	3	-
	Rv1484	*inhA*	3	-
	Rv3854c	*ethA*	33	42
	Rv3855	*ethR*	2	-
Amikacin	*rrs*	*rrs*	6	-
Capreomycin	*Rv1694*	*tlyA*	16	13
	rrs	*rrs*	4	-
Kanamycin	*Rv2416c*	*eis*	10	-
	rrs	*rrs*	4	-
FQ	*Rv0005*	*gyrB*	26	-
	Rv0006	*gyrA*	21	-
PAS	*Rv2447c*	*folC*	18	-
	Rv2671	*ribD*	1	-
	Rv2754c	*thyX*	1	-
	Rv2764c	*thyA*	19	5
Cycloserine	*Rv2780*	*ald*	-	12
	Rv3423c	*alr*	3	-

* https://github.com/jodyphelan/tbdb;; indels = insertions and deletions, FQ = Fluoroquinolones, PAS = Para-aminosalicylic acid

S6 Table

Comparison between Gradient Boosted Tree model (GBT-CRM), TB-Panel, TB-Profiler and GWAS study

Drug	TB-Panel*			TB-Profiler**			GWAS***			GBT-CRM****		
	Sens	Spec	Acc.	Sens	Spec	Acc.	Sens	Spec	Acc.	Sens	Spec	Acc.
INH	88.0	97.0	94.1	93.7	98.1	96.7	92.2	98.6	96.6	91.1	98.8	96.3
RIF	84.1	99.4	95.3	95.9	98.2	97.6	92.9	98.6	97.1	88.8	98.9	96.2
PZA	19.9	98.8	86.4	87.6	96.7	95.3	39.4	98.2	89.0	69.7	96.1	91.8
EMB	84.1	93.3	91.7	92.1	91.7	91.8	89.0	92.9	92.2	82.8	94.2	92.1
STM	81.4	81.5	81.5	78.0	96.3	91.6	70.2	98.1	90.9	79.8	96.0	91.9
AMK	82.3	86.5	85.5	86.0	98.3	95.4	86.0	98.6	95.7	80.5	99.5	95.1
CAP	76.3	84.9	83	84.7	95.9	93.4	78.7	96.7	92.6	74.6	96.2	91.3
KAN	84.9	86.9	86.2	92.0	96.8	95.1	86.2	98.2	94.0	82.2	98.2	92.1
CIP	80.9	98.2	95.5	90.6	98.0	96.8	84.1	98.8	96.5	85.7	98.5	96.2
OFL	81.0	97.4	93.2	90.1	96.5	94.9	83.8	97.6	94.1	81.0	97.0	93.2
MOX	81.7	92.7	91.4	86.0	91.9	91.2	81.7	93.6	92.2	53.3	97.5	93.7
ETH	76.5	75.4	75.8	89.5	67.4	75.1	55.7	86.6	75.7	68.1	93.4	84.6
CYS	-	-	-	43.0	92.5	79.2	33.3	98.3	80.8	50.0	92.4	78.4
PAS	9.3	97.8	88.4	23.8	96.7	89.0	48.8	95.3	90.4	20.0	100.0	90.2
MDR	79.7	97.7	93.8	94.1	98.3	97.3	89.8	98.7	96.5	90.4	96.9	95.5

GBT-CRM is a Gradient Boosted Tree (GBT) that is applied to all genome-wide SNPs (including co-occurrent resistance markers); * List of TB Profiler panel mutations with minor allele frequency > 0.5% in the dataset, and applied on a “rule-in” basis; ** TB-Profiler prediction (24); *** GWAS approach as described in (12) but re-run on the 17k dataset used in this study; ****Reported outcomes for GBT-CRM is based on the performance when applied to the test-set; RIF=rifampicin, INH=isoniazid, EMB=ethambutol, PZA=pyrazinamide, CIP=ciprofloxacin,OFL= ofloxacin, MOX=moxifloxacin, AMK=amikacin, KAN=kanamycin, CAP=capreomycin, PAS=para-aminosalisylic acid (PAS), CYS=cycloserine, ETH=ethionamide, Sens=Sensitivity, Spec=Specificity, Acc=Accuracy, MDR = multi-drug resistant TB.

S7 Table

Comparison between Gradient Boosted Tree model (GBT-CRM) and average scores across other machine learning studies*

	Sens GBT-CRM	Spec GBT-CRM	AUC GBT-CRM	Sens Other (avg.)	Spec Other (avg.)	AUC - Other (avg.)	Difference (Sens)	Difference (Spec)	Difference (AUC)
INH	91.1	98.8	96.7	93.4	96.1	97.6	-2.3	2.7	-0.9
RIF	88.8	98.9	97.9	94.7	97.0	98.4	-5.9	1.9	-0.5
PZA	69.7	96.1	95.5	83.7	90.3	92.9	-14.0	5.8	2.6
EMB	82.8	94.2	95.8	93.5	93.8	97.4	-10.7	0.4	-1.6
STM	79.8	96.0	94.0	88.2	91.3	93.5	-8.4	4.8	0.5
AMK	80.5	99.5	96.4	83.4	90.3	93.2	-2.9	9.2	3.2
CAP	74.6	96.2	93.4	68.2	89.2	83.1	6.4	7.0	10.3
KAN	82.2	98.2	96.8	80.8	91.5	90.2	1.4	6.7	6.6
CIP	85.7	98.5	99.7	87.9	91.7	93.8	-2.2	6.8	5.9
OFL	81.0	97.0	93.3	81.6	92.2	91.3	-0.6	4.8	2.0
MOX	53.3	97.5	97.2	87.3	90.4	91.8	-34.0	7.1	5.4
ETH	68.1	93.4	88.4	90.6	85.6	92.2	-22.5	7.8	-3.8
MDR	90.4	96.9	97.4	96.2	96.5	99.4	-5.8	0.4	-1.9

GBT-CRM is a Gradient Boosted Tree (GBT) that is applied to all genome-wide SNPs (including co-occurrent resistance markers); * other studies with results found in references (22,23,26,27). For study (27) the performance is the DeepAMR model. Note: Not all studies included all drugs; Reported outcomes for the GBT-CRM is the performance based on the application to the test-set ; RIF=rifampicin, INH=isoniazid, EMB=ethambutol, PZA=pyrazinamide, CIP=ciprofloxacin,OFL= ofloxacin, MOX=moxifloxacin, AMK=amikacin, KAN=kanamycin, CAP=capreomycin, PAS=para-aminosalisylic acid (PAS), CYS=cycloserine, ETH=ethionamide, Sens = Sensitivity, Spec = Specificity, AUC=Area under the ROC Curve

S8 Table

Comparison between Gradient Boosted Tree model (GBT-CRM) and maximum scores in other machine learning studies*

	Sens GBT-CRM	Spec GBT-CRM	AUC GBT-CRM	Sens Other (max)	Spec Other (max)	AUC Other (max)	Difference (Sens)	Difference (Spec)	Difference (AUC)
INH	91.1	98.8	96.7	97.0	98.4	99.0	-5.9	0.4	-2.3
RIF	88.8	98.9	97.9	97.0	97.8	99.0	-8.2	1.1	-1.1
PZA	69.7	96.1	95.5	88.1	91.2	95.0	-18.4	4.9	0.5
EMB	82.8	94.2	95.8	97.0	96.0	99.0	-14.2	-1.8	-3.2
STM	79.8	96.0	94.0	90.1	94.2	95.2	-10.3	1.8	-1.2
AMK	80.5	99.5	96.4	89.5	90.8	95.0	-9.0	8.7	1.4
CAP	74.6	96.2	93.4	71.9	92.7	85.5	2.7	3.5	7.9
KAN	82.2	98.2	96.8	81.1	93.5	92.5	1.1	4.7	4.3
CIP	85.7	98.5	99.7	96.0	98.0	98.0	-10.3	0.5	1.7
OFL	81.0	97.0	93.3	96.0	93.7	95.0	-15.0	3.3	-1.7
MOX	53.3	97.5	97.2	95.0	93.0	95.0	-41.7	4.5	2.2
ETH	68.1	93.4	88.4	90.6	85.6	92.2	-22.5	7.8	-3.8
MDR	90.4	96.9	97.4	96.3	98.0	100	-5.9	-1.1	-2.6

GBT-CRM is a Gradient Boosted Tree (GBT) that is applied to all genome-wide SNPs (including co-occurrent resistance markers); * other studies with results found in references (22,23,26,27). For study (27) the performance is the DeepAMR model. Note: Not all studies included all drugs; Reported outcomes for the GBT-CRM is the performance based on the application to the test-set ; RIF=rifampicin, INH=isoniazid, EMB=ethambutol, PZA=pyrazinamide, CIP=ciprofloxacin,OFL= ofloxacin, MOX=moxifloxacin, AMK=amikacin, KAN=kanamycin, CAP=capreomycin, PAS=para-aminosalisylic acid (PAS), CYS=cycloserine, ETH=ethionamide, Sens = Sensitivity, Spec = Specificity, AUC=Area under the ROC Curve

S9 Table

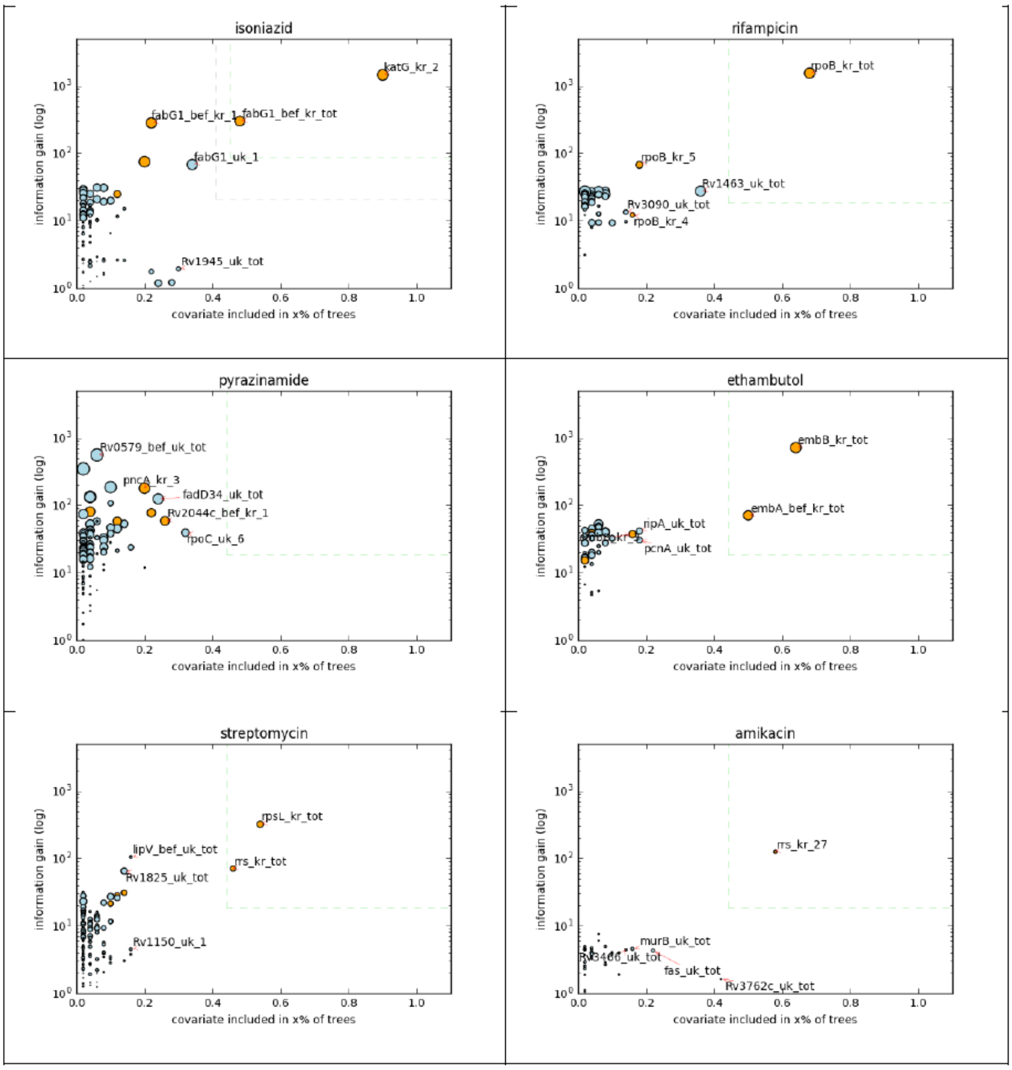
The machine learning parameter settings

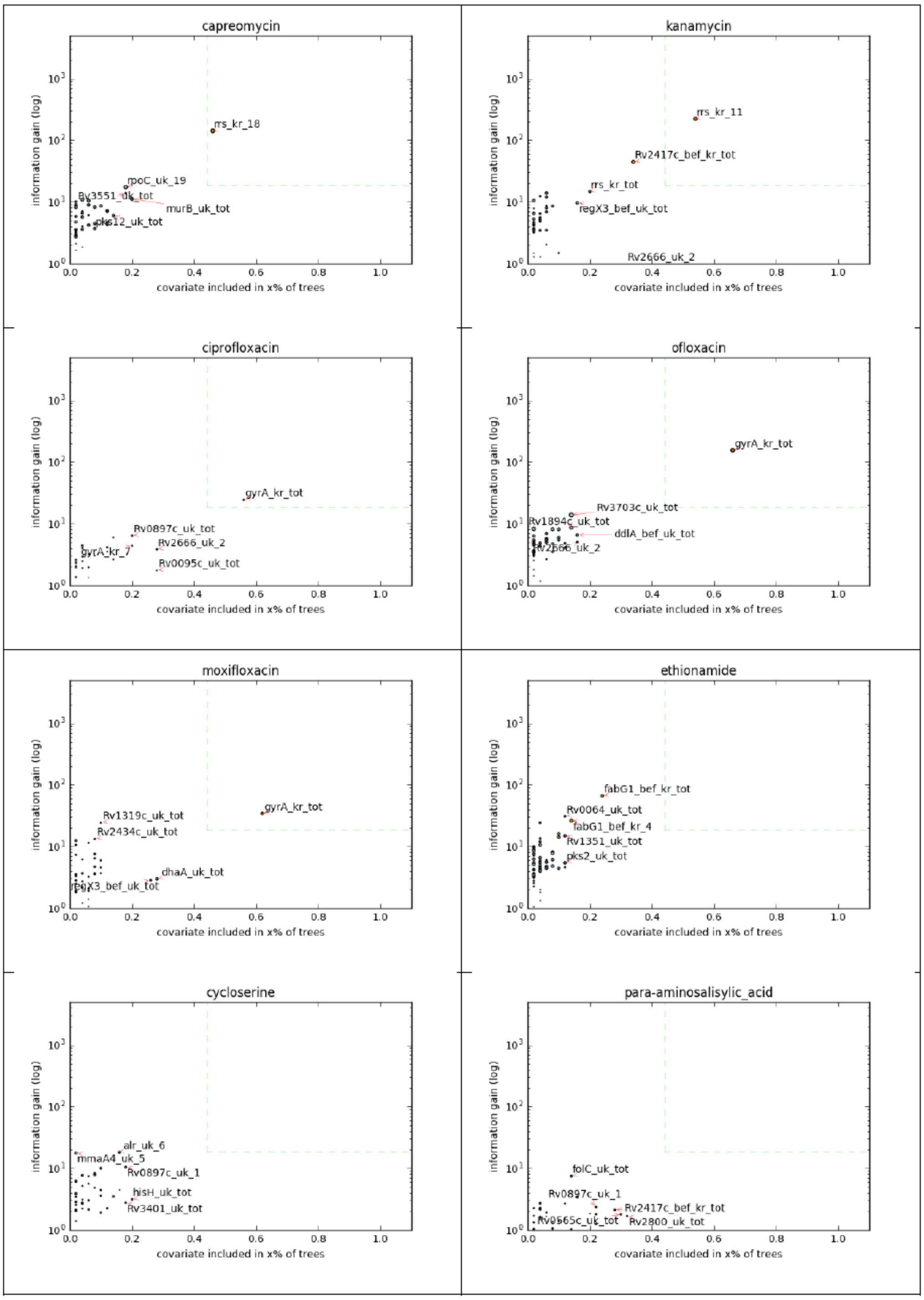
Classifier	Parameters
Decision Tree Classifiers (“CT”)	Function to measure the quality of a split = Gini, Minimum of samples required before making splits=3, Minimum of samples required for leaf nodes=3, The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf nod =0, The number of covariates to consider when looking for the best split=None, max_leaf_nodes=None, Minimum impurity decrease required for splits=0.0, Minimum impurity threshold=0, Class weighting=None.
Gradient Boosted Tree Classifier (“GBT”)	Boosting learning rate=0.1 Booster='gbtree' Min. loss reduction required for further partition on a leaf node =0 Minimum sum of instance weight(hessian) needed in a child =1 Maximum delta step we allow each tree’s weight estimation to be=0 Subsample ratio of columns when constructing each tree=1 Subsample ratio of columns for each split, in each level=1 L1 regularization term on weights=0 L2 regularization term on weights=1 Global bias=0.5.
Logistic Regression (“LR”)	Penalty=”L1” Tolerance=0.0001 Maximum iterations=100

* see Methods for those parameters that were chosen by cross-validation

S1 Figure

A two-dimensional mutation ranking across drugs created from the outputs of the gradient boosted tree (GBT) models, using the proportion of GBT trees within the overall ensemble they appear in and the information gain associated with their presence. The orange points refer to previously known SNPs (TB-profiler), with the dotted green box as a suggested detection threshold determined by optimizing the discrimination between previously known SNPs and other SNPs across drugs.




capreomycin
rrs_kr_18
rpoC_uk_19
Rv3551_uk_tot
murB_uk_tot
pks12_uk_tot
kanamycin
rrs_kr_11
Rv2417c_bef_kr_tot
rrs_kr_tot
regX3_bef_uk_tot
Rv2666_uk_2
ciprofloxacin
gyrA_kr_tot
Rv0897c_uk_tot
Rv2666_uk_2
gyrA_kr_7
Rv0095c_uk_tot
ofloxacin
gyrA_kr_tot
Rv3703c_uk_tot
Rv1894c_uk_tot
ddlA_bef_uk_tot
Rv2666_uk_2
moxifloxacin
gyrA_kr_tot
Rv1319c_uk_tot
Rv2434c_uk_tot
dhaA_uk_tot
regX3_bef_uk_tot
ethionamide
fabG1_bef_kr_tot
Rv0064_uk_tot
fabG1_bef_kr_4
Rv1351_uk_tot
pks2_uk_tot
cycloserine
alr_uk_6
mmaA4_uk_5
Rv0897c_uk_1
hisH_uk_tot
Rv3401_uk_tot
para-aminosalisylic_acid
folC_uk_tot
Rv0897c_uk_1
Rv2417c_bef_kr_tot
Rv0565c_uk_tot
Rv2800_uk_tot
information gain (log)
10^0
10^1
10^2
10^3
covariate included in x% of trees
0.0
0.2
0.4
0.6
0.8
1.0

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1701929	**Title**	Mr.
First Name(s)	Wouter		
Surname/Family Name	Deelder		
Thesis Title	Machine learning methods for infectious diseases: applications for tuberculosis and malaria.		
Primary Supervisor	Prof. Taane Clark, Dr. Luigi Palla		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	BMC Genomics		
When was the work published?	January 2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	N/A
Please list the paper's authors in the intended authorship order:	N/A
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I helped to conceive and design the study. I developed the idea of the adjusted decision tree algorithm and the consensus-decision-rule. I implemented the machine learning algorithms and analysed the data. I wrote the first draft of the manuscript, and finalised it after receiving revisions from co-authors and reviewers.

SECTION E

Student Signature	
Date	April 11, 2022

Supervisor Signature	
Date	April 11, 2022

Deelder *et al. BMC Genomics* (2022) 23:46
https://doi.org/10.1186/s12864-022-08291-4

BMC Genomics

RESEARCH **Open Access**

A modified decision tree approach to improve the prediction and mutation discovery for drug resistance in *Mycobacterium tuberculosis*

Wouter Deelder[1,2], Gary Napier[1], Susana Campino[1], Luigi Palla[1,3], Jody Phelan[1†] and Taane G. Clark[1,4*†]

Abstract

Background: Drug resistant *Mycobacterium tuberculosis* is complicating the effective treatment and control of tuberculosis disease (TB). With the adoption of whole genome sequencing as a diagnostic tool, machine learning approaches are being employed to predict *M. tuberculosis* resistance and identify underlying genetic mutations. However, machine learning approaches can overfit and fail to identify causal mutations if they are applied out of the box and not adapted to the disease-specific context. We introduce a machine learning approach that is customized to the TB setting, which extracts a library of genomic variants re-occurring across individual studies to improve genotypic profiling.

Results: We developed a customized decision tree approach, called Treesist-TB, that performs TB drug resistance prediction by extracting and evaluating genomic variants across multiple studies. The application of Treesist-TB to rifampicin (RIF), isoniazid (INH) and ethambutol (EMB) drugs, for which resistance mutations are known, demonstrated a level of predictive accuracy similar to the widely used TB-Profiler tool (Treesist-TB vs. TB-Profiler tool: RIF 97.5% vs. 97.6%; INH 96.8% vs. 96.5%; EMB 96.8% vs. 95.8%). Application of Treesist-TB to less understood second-line drugs of interest, ethionamide (ETH), cycloserine (CYS) and para-aminosalisylic acid (PAS), led to the identification of new variants (52, 6 and 11, respectively), with a high number absent from the TB-Profiler library (45, 4, and 6, respectively). Thereby, Treesist-TB had improved predictive sensitivity (Treesist-TB vs. TB-Profiler tool: PAS 64.3% vs. 38.8%; CYS 45.3% vs. 30.7%; ETH 72.1% vs. 71.1%).

Conclusion: Our work reinforces the utility of machine learning for drug resistance prediction, while highlighting the need to customize approaches to the disease-specific context. Through applying a modified decision learning approach (Treesist-TB) across a range of anti-TB drugs, we identified plausible resistance-encoding genomic variants with high predictive ability, whilst potentially overcoming the overfitting challenges that can affect standard machine learning applications.

Keywords: *Mycobacterium tuberculosis*, Ethionamide, Cycloserine, PAS, Drug resistance, Machine learning

Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, is a pressing global health problem, with >10 million cases and 1.4 million associated deaths in 2019 [1]. First-line TB treatment uses combinations of the drugs rifampicin (RIF), isoniazid (INH), ethambutol (EMB)

---

*Correspondence: Taane.clark@lshtm.ac.uk
†Jody Phelan and Taane G. Clark are Joint authors
[4] Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK
Full list of author information is available at the end of the article

and pyrazinamide (PZA) [2]. Drug-resistance requires switching to second-line therapies combined in customized treatment protocols, which might include fluoroquinolones and injectable drugs, as well as ethionamide (ETH), cycloserine (CYS) and para-aminosalisylic acid (PAS), among others. Historically, a cascade of resistance has been defined, from resistance to RIF (RR-TB), to additional resistance to INH leading to multidrug resistance (MDR-TB), further leading to an extensively drug resistant (XDR-TB) class that is MDR-TB with additional resistance to fluoroquinolones and second-line injectables. Recently, there was a new definition of pre-XDR (MDR-TB and resistance to any fluoroquinolone) and an updated definition of XDR-TB (pre-XDR and resistance to at least one additional Group A drug, including levofloxacin or moxifloxacin, bedaquiline and linezolid) [3]. These updates provide a framework for increasing progression of the severity of disease linked to resistance to additional anti-TB drugs [3].

The mechanisms that cause *M. tuberculosis* drug resistance are linked to genomic variants in drug targets or pro-drug activators, including single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels), some occurring in gene-gene interactions. Pro-drug activators convert mycobacterial enzymes that convert pro-drugs, such as INH and ETH, into their active form. If these enzymes (e.g., catalase peroxidase (KatG) for INH) are not essential, their coding genes can acquire mutations such as frameshifts which lead to loss of function, and consequently, the respective drug is not converted and resistance is caused. However, not all resistance mechanisms are well understood [4–6], especially for second-line drugs (e.g. PAS). Drug-resistance has been traditionally assessed through bacterial culture-based phenotypic drug susceptibility testing (DST), which can be time-consuming and resource intensive, with reproducibility and inhibitory concentration cut-off challenges for particular drugs [7]. Whole-genome sequencing (WGS) offers an alternative approach to infer resistance through the identification of associated genomic mutations [8], called "genotypic resistance" profiling. TB-Profiler software [9, 10] uses a curated library of >1000 mutations to predict genotypic resistance across 14 anti-TB drugs. The use of WGS can reaffirm known resistance mutations and uncover new candidates through genome-wide association studies (GWAS) and convergent evolution analysis [11]. However, GWAS approaches typically focus on single variants at a time in regression models, whereas resistance phenotype prediction from WGS is a classification problem with high-dimensional input and potential complex interactions, a standard task in machine learning [12]. Therefore, the ongoing generation of large datasets using WGS is highly suited to the application of machine learning methods to improve "genotypic resistance" profiling [12].

The application of machine learning methods to *M. tuberculosis* has shown some impressive performances in genotypic profiling [13–17]. However, these models have several drawbacks that could affect their application in clinical settings, including their interpretability and an optimism bias related to the inclusion of non-associated cross-resistance and bacterial lineage markers; both leading to reduced predictive performance in hospital and other clinical settings [15]. The performance of machine learning models has also been relatively poor for a subset of second-line drugs (CYS, PAS, ETH), which in general are less often studied and analysed [11, 15]. The generally lower performance for CYS, PAS and ETH suggests that mechanisms of resistance are less well understood, and that potentially rare alleles are being missed and excluded from models [15]. Our study aims to attempt to detect new genomic variants that might cause resistance for CYS, PAS, and ETH. The approach involves a customized (decision tree) machine learning algorithm, called Treesist-TB, which detects genomic variants in individual studies within the aggregated datasets, and can model variant interactions. It attempts to be robust to the presence of DST errors in some of the individual studies, which can lead to genomic variants being undetected in the analysis of the aggregate dataset.

Results

Genomic and phenotypic data

WGS data was available for 32,689 (32 k) *M. tuberculosis* samples, which covered the main lineages 1 (9.6%), 2 (25.2%), 3 (11.4%) and 4 (51.0%) (S1 Table). Most samples were pan-susceptible (77.9%), but RR-TB (1.3%), MDR-TB (13.0%) and XDR-TB (2.3%) phenotypes were also represented. Phenotypic DST data was not available for all isolates, with limited data generation for PAS ($n=1114$, 8.8% resistant), CYS ($n=833$, 18.0% resistant), and ETH ($n=2138$, 32.2% resistant) (S2 Table; Table 1), as these drugs are mostly prescribed to and assessed in patients with RR-TB and MDR-TB.

Application of Treesist-TB to first-line drugs

Treesist-TB is a python-based machine learning algorithm that fits customized decision trees across individual studies and combines the extracted features to make final resistance predictions. It can also, if desired, be run assuming all data is from a single study (referred to as a "single optimised tree"). The algorithm was first applied to well-understood first-line drugs, using a subset of isolates that had complete DSTs (RIF: $n=2045$, 8.1% resistant, 7 studies; INH $n=1835$, 16.2% resistant; 6 studies;

Table 1 Predictive performance across algorithms

Drug	Total tests	% resistance	TB-Profiler				Treesist-TB[a]			
			Sens	**Spec**	**Acc**	**AUC**	**Sens**	**Spec**	**Acc**	**AUC**
INH	1835	16.2	86.2	98.4	96.5	92.3	84.2	99.2	96.8	91.7
RIF	2045	8.1	90.3	98.2	97.6	94.2	86.1	98.5	97.5	92.3
EMB	1999	3.5	71.4	96.7	95.8	84.1	57.1	98.2	96.8	77.7
PAS	1114	8.8	38.8	95.7	90.7	67.2	64.3	90.6	88.2	77.4
CYS	833	18.0	30.7	95.2	83.6	62.9	45.3	93.7	85.0	69.5
ETH	2118	32.2	71.1	78.6	76.2	74.8	72.1	75.8	74.6	73.9
			Regular Decision Tree				**Single optimized Tree[b]**			
			Sens	**Spec**	**Acc**	**AUC**	**Sens**	**Spec**	**Acc**	**AUC**
INH	1835	16.2	85.6	100	97.7	92.9	80.2	99.2	96.1	89.8
RIF	2045	8.1	81.2	100	98.5	91.5	87.3	99.8	98.8	93.6
EMB	1999	3.5	32.9	99.7	97.3	82.9	34.3	99.5	97.2	83
PAS	1114	8.8	64.3	100	96.9	85.5	50	97.8	93.6	74.1
CYS	833	18.0	33.3	99.4	87.5	67.3	35.3	98	86.7	66.7
ETH	2118	32.2	48.8	94.3	79.7	77.5	49.6	92.5	78.7	76.2

INH Isoniazid, *RIF* Rifampicin, *PAS* para-aminosalisylic acid, *CYS* cycloserine, *ETH* ethionamide, *EMB* Ethambutol, *Sens* Sensitivity, *Spec* Specificity, *Acc* Accuracy, *AUC* Area under the ROC Curve

[a] default application of Treesist-TB

[b] application of Treesist-TB with a single combined study dataset

EMB: $n=1999$, 3.5% resistant, 5 studies; S2 Table) across second-line drugs.

We fitted a default Treesist-TB tree assuming individual studies, as well as, for comparison purposes, single optimised and regular decision trees. The single optimized trees were simpler and contained fewer implausible sub-structures than regular decision trees (S2 Figure) while maintaining relevant structures such as double mutations and gene-gene interactions. In particular, the optimized trees contain fewer genes (INH: 27 vs. 5; RIF: 6 vs. 4; EMB: 5 vs. 4) but generally more individual variants (INH: 29 vs. 6; RIF: 15 vs. 20; EMB: 6 vs. 5) than regular decision trees. However, single optimized trees do include some unlikely features that might arise from overfitting on DST errors or other artefacts in the aggregated dataset (S2 Figure), so we applied the default Treesist-TB algorithm, which incorporates information from individual sub-studies.

The Treesist-TB algorithm identified several predictive genomic variants for resistance of RIF ($n=20$; 7 unreported in the TB-Profiler library), INH ($n=20$, 13 unreported) and EMB ($n=10$, 2 unreported) (S1 Figure, S2 Figure, Table 2; S4 Table). These included mutations in established loci such as *rpoB* ($n=18$, RIF), *katG* ($n=17$, INH), and *embB* ($n=7$, EMB). A confirmation analysis of the Treesist-TB mutations in the set of validation isolates ($n=\sim30$ k of 32 k, not analysed by Treesist-TB), revealed that none were present in susceptible strains, but they were frequent in both MDR-TB (median (maximum): RIF 1.6% (65.3%); INH <0.1% (79.2%); EMB 2.1 (23.8)) and XDR-TB (RIF 0.8% (70.7%); INH <0.1% (78.6%); EMB 3.1% (35.3%)) isolates. The predictive accuracy of resistance from Treesist-TB was similar to the TB-Profiler tool (Treesist-TB vs. TB-Profiler: RIF 97.5% vs. 97.6%; INH 96.8% vs. 96.5%; EMB 96.8% vs. 95.8%), and like those from the single optimized and regular decision trees (Table 1), whose models include mutations not associated with resistance.

Application to selected second-line drugs

Given the strong performance for first-line drugs, Treesist-TB was then implemented for PAS, CYS and ETH, for which predictive accuracy has historically been poor and resistance mutations are only partially known [10]. Again, for comparison purposes, we fitted a single optimized tree for each drug and contrasted the performance and structure with regular classification trees (S3 Figure; Table 2). The results revealed that the optimized trees contain both fewer genes (PAS: 33 vs. 4; CYS: 7 vs. 3; ETH: 11 vs. 3) and variants (PAS: 37 vs. 7; CYS: 7 vs. 3; ETH: 13 vs. 5) than regular decision trees. The single optimized trees were simpler and contained fewer implausible sub-structures than regular decision trees, which appeared to be over-fitted (S2 Figure).

For PAS, the default application of Treesist-TB detected 11 genomic variants across three genes (*folC* 6, *Rv2670c* 1, *thyX* 4) (Table 2). Six of the variants are unreported in TB-Profiler, occurring in *folC* (R49Q, Ser98G),

Table 2 The Treesist-TB inferred variants

Drug	Gene	# variants in the 32 k dataset*	Treesist-TB Mutations**
RIF	*rpoB*	757	**N163K**, V170F, L430P, Q432K, Q432L, D435Y, D435V, S441L, **H445D, H445D, H445N**, H445Y, **H445R,** H445L, S450L, L452P, I491F
RIF	*rpoC*	700	**N1239D, Q1289A**
INH	*ahpC*	31	**-57C > T, −48G > A**
INH	*fabG1*	26	**-126G > A**
INH	*katG*	648	**Y597D, T568P, A476V,** S315T, S315N, S302R, **W300C**, G297V, **P193fs, L159F**, **G156D, A144V, D142G,** L141F, N138D, A109V, **Y98C**
EMB	*embA*	743	**-31delC,** −16C > T, **−16C > A**
EMB	*embB*	762	M306V, M306L, M306I, G406A, Q497K, Q497R, D1024N
PAS	*Rv2670c*	191	**A5V**
PAS	*folC*	262	Q153G, Q153A, S150G, **S98G, R49Q,** I43T
PAS	*thyX*	148	**-4C > T, −9G > A**, −16C > T, **−18G > T**
CYS	*alr*	239	**Y388D, L283P,** L113R, **T20M**
CYS	*rpoC*	700	**D485Y**, I491T
ETN	*ethA*	494	**W455*, K448fs, P436fs, A352fs, P334A, F320S, L295fs, C294*, R279*, Q269*, M260I, W256*, C253F, T236fs, Y235fs, W228*, N226fs, K224*, A222V, S208L, R207G, V202F, L194P, T186P, P164R, P160fs, C137R, C137R, W116*, K103fs, W45*, K37fs, L35R, Q24*, D6fs**
ETN	*fabG1*	26	**-118C > G, −34C > T, −15C > T, −8T > C, −8T > A**
ETN	*gyrA*	764	A90V, S91P, D94A, D94G
ETN	*inhA*	108	I21T, **R27W,** I194T, **P251R**
ETN	*mshA*	250	**A133fs, H175fs, V237L, A422V**

* 32 k *M. tuberculosis* isolates [18]

** **Bolded** if not in TB-Profiler in https://github.com/jodyphelan/tbdb/blob/master/tbdb.csv; * stop codon

INH Isoniazid, *RIF* Rifampicin, *PAS* para-aminosalisylic acid, *CYS* cycloserine, *ETH* ethionamide, *EMB* Ethambutol

** Mutations underlined if they are in > 5% of MDR-TB or XDR-TB strains in the 32 k *M. tuberculosis* isolates

Rv2670c (A5V), and *thyX* (three indels: -4C > T, −9G > A, −18G > T) (S5 Table). These PAS mutations were present in XDR-TB samples in the validation set (frequency: median 0.2%, max. 6.1%) (S5 Table). For PAS, compared to TB-Profiler, the Treesist-TB mutation set leads to a higher sensitivity (64.3% vs. 38.8%), lower specificity (90.6% vs. 95.7%) and similar overall accuracy (88.2% vs. 90.7%) for drug resistance prediction (Table 1). For CYS, Treesist-TB identified six variants across two genes (*rpoC 2,* 1 unreported; *alr* 4, 3 unreported). *RpoC* is a locus linked to compensatory effects in RIF resistance. The CYS mutations were present in XDR-TB samples in the validation set (frequency: median < 0.1, max. 8.5%) (S5 Table). Compared to TB-Profiler, the set of Treesist-TB mutations had a higher sensitivity (45.3% vs. 30.7%), and similar specificity (93.7% vs. 95.2%) and overall accuracy (85.0% vs. 83.6%) for resistance prediction. For ETH, Treesist-TB identified 52 genomic variants, more than half in *ethA* (35; 67.3%), with others found across four genes (*inhA* 4, *gyrA* 4, *mshA* 4, *fabG1 promoter* 5). Most variants are not present in the TB-Profiler library (*ethA* 34, *inhA* 2, *mshA* 4, *fabG1 promoter* 5). *EthA*, *fabG1 promoter* and *inhA* are established ETH related loci, but *gyrA* is linked to fluoroquinolone resistance, and *mshA* is known to encode a glycosyl-transferase enzyme involved in mycothiol biosynthesis that can affect ETH activation. These mutations for ETH were present in XDR-TB samples in the validation set (frequency: median < 0.1%, max. 36.5%) (S5 Table). For ETH, compared to TB-Profiler, Treesist-TB has a marginally higher sensitivity (72.1% vs. 71.1%), lower specificity (75.8% vs. 78.6%) and a similar overall accuracy (74.6% vs. 76.2%) for drug resistance prediction.

Discussion

The relatively poor knowledge of underlying mutations for second-line anti-TB drug resistance will make prospects for WGS-informed clinical and infection control more difficult. Whilst machine learning has the promise to fill any gaps in "genetic" knowledge, some implementations for *M. tuberculosis* "genotypic profiling" have led to over-optimistic predictive abilities and models with mutations that are not biologically plausible or unrelated to the resistance of interest. Our work describes a decision tree machine learning approach, called Treesist-TB, which attempts to account for inter-study differences and constrains the size of models, thereby minimising the risk of over-fitting due to phylogenetic or false resistance-associated mutations. Its application to RIF, INH and EMB drugs, with known resistance mechanisms, detected both established and unreported mutations in functional pathways, and had predictive abilities similar to other machine learning implementations and the TB-Profiler tool. Application of Treesist-TB to CYS, PAS and ETH drugs, whose underlying resistance variants are less established and are less often studied, detected putative non-synonymous SNPs and frameshift mutations in activation pathways. For the PAS drug, genomic variants were found in the *folC* gene, which interrupts bioactivation within the folate biosynthetic pathway [19]. Similarly, mutations were found in the *alr* gene encoding alanine racemase that compensates for the inhibitory effect of CYS [20]. Finally, for ETH, the majority of mutations were detected in the *ethA* gene that activates ETH by the NADPH-specific flavin adenine dinucleotide-containing monooxygenase EthA [21]. Importantly, integrated WGS and DST studies for relatively new anti-TB drugs (e.g., bedaquiline, clofazimine and delamanid) are much-needed, as current low sample sizes make the determination of mutations underlying their resistance difficult [22].

Treesist-TB detects SNPs by working with the largest datasets possible, where some of the reported performance problems for second-line drugs are due to the exclusion of rare alleles. More importantly, Treesist-TB considers individual sub-studies that make up the large dataset, implicitly adjusting for potential DST or mislabelling errors in individual studies, which are potentially more common in some laboratories or drug assays. Treesist-TB also incorporates existing knowledge on which sub-structures in the decision trees are biologically less plausible, such as reversion mutations, and can prune these structures. If required, the approach can give preference to known resistance genomic variants in tree model building and control its complexity by placing a ceiling on the number of previously unknown resistance mutations. In this sense, Treesist-TB can take advantage of prior knowledge and insights specific to TB drug resistance, thereby providing a counterweight against the increasing usage of machine learning "out of the box", which can lead to models that do not generalize well in clinical practice.

Our analysis revealed that standard machine learning approaches could, even after regular cross-validation, overfit in subtle manners that lead to an upward bias in performance and not translate into a high out-of-training-set performance. Although, a robust simulation study that considers a number of machine learning approaches is beyond the scope of our work, previous studies have shown that some implementations have boosted performance through the selection of cross-resistance markers that are unlikely to be causally related to resistance to the drug under investigation [15]. These unrelated markers might get selected as features by machine learning models due to the unique structure of TB datasets, including arising from *M. tuberculosis* phylogenetic structures and sequential drug testing practices. Similarly, fitted tree structures with features that are biologically unrelated to resistance might lead to impressive performance within the training set, but may be inappropriate for predictions in clinical practice. These problems will be exacerbated for more complex models that have a greater number of parameters, such as convolutional neural nets [23].

Conclusions

In general, with the increasing application of WGS data in a clinical or research setting, there is a need for robust and interpretable machine learning models that take advantage of the resulting large and growing datasets, whilst being robust to data errors. One important application is in antimicrobial resistance (AMR) genotypic profiling, which could ultimately replace phenotypic DST approaches. However, any AMR models derived must be reliable in terms of prediction, generalize across clinical settings, and adapt to increasing data and knowledge. In addition, such models need to account for the idiosyncrasies of pathogens and infections, where *M. tuberculosis* is highly clonal and has no horizontal gene transfer, but for other pathogens there may be plasmid derived AMR. In conclusion, we have developed Treesist-TB, which can assist with identifying mutations and prediction drug resistance in a TB context. Through providing software for its implementation, the utility beyond TB can be evaluated, and the approach potentially refined for other AMR settings.

Materials and methods

Phenotypic and sequencing data

The main dataset consists of 32,689 (32 k) isolates with whole genome sequencing (WGS) and phenotypic drug susceptibility test (DST) data (see S1 Table [18];). The laboratory DST followed WHO recommended protocols and practice (see [11]). XDR-TB was defined using the recently replaced definition, that is, being MDR-TB with additional resistance to fluoroquinolones and second-line injectables. This is because the isolates were collected, processed, and resistance patterns interpreted for treatment options before the new definitions were introduced [3]. DST data was not available for every isolate across all drugs, as only those individuals resistant to first-line treatments are typically tested for second-line resistance. All isolates with PAS, CYS and ETH DST were included in the analysis (see S2 Table for sample sizes). A subset with complete INH, RIF and EMB DST data and with similar characteristics in terms of sample size and number of individual studies were chosen for Treesist-TB benchmark analysis (S2 Table). The residual 31 k isolates were used for validation through the analysis of mutation frequencies across susceptible and resistance groups. The raw sequence data were mapped to the H3Rv reference genome using *bwa-mem* software, and genomic variants (SNPs, indels) were called from the consensus of GATK and *samtools* software. Most genomic variants (98.9%) have low minor allele frequencies (< 1%), and we excluded SNPs in hypervariable PE/PPE gene families and with synonymous mutations (see [18]).

Treesist-TB model

The Treesist-TB model is a major extension of a simple decision tree approach (sklearn implementation, v0.23.1) with the following modifications: (1) incorporation of prior parameters on which features to prioritize in the tree building in case of ties; 2) incorporation of tree pruning to limit interactions in the tree that are *a priori* determined to be unlikely (e.g. double mutations that compensate resistant mutations and restore drug sensitivity); (3) incorporation of prior parameters for the maximum number of genes (not genomic variants) in a tree. Although Treesist-TB is compatible with regular cross-validation methods (e.g., leave k-fold out), these approaches may lead to unstable results for trees in general. To prevent trees from having excessive depth, the setting of priors for the maximum number of new genes outside known resistance genes (not variants) has been implemented. We extracted a set of genomic variants using a consensus rule that variants were only included when in genes that were more than once detected across sub-datasets (S1 Figure).

Model fitting

The predictive performance of the final models fitted to the entire dataset was measured using sensitivity, specificity, accuracy, and area under the ROC curve (AUC) metrics, assuming DST results as the gold standard. We compared the performance of the (default) Treesist-TB model primarily with the TB-Profiler software and mutation library (> 1000 SNPs, indels or large deletions) [9, 10]. In addition, for comparison, we fitted a regular decision tree model and Treesist-TB (labelled as “Single optimized Tree”) on aggregate datasets. The depth of the regular decision tree was set by 5-fold cross-validation up to a maximum of 15.

Packages

The pipeline was implemented in Python (v3.7), building on the tree algorithm from sklearn (v0.23.1). The plausibility of putatively causal genomic variants identified was assessed using *Mycobrowser* [24].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08291-4.

Additional file 1.

Acknowledgements
We thank Aleksei Ponomarev and Gabriel Marzinotto for coding support and public code that was used in the development of Treesist-TB.

Authors’ contributions

WD, JP and TGC conceived and designed the study. GN and JP performed the bioinformatic processing of the raw sequencing data and phenotypic data; WD developed the algorithm and performed the statistical analysis, under supervision of SC, LP, JP, and TGC. WD wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

Funding
JP is funded by a Newton Institutional Links Grant (British Council, no. 261868591). TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). SC is funded by Medical Research Council UK grants (ref. MR/M01360X/1, MR/R025576/1, and MR/R020973/1).

Availability of data and materials

The raw whole genome sequencing data is available from the European Nucleotide Archive (ENA) (see [18]). Computing code is available at https://github.com/WDee/Treesist-TB.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors approve the publication.

Competing interests

There are no competing interests. WD was employed by the company Dalberg Advisors in Switzerland. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details
[1]London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Dalberg Advisors, 7 Rue de Chantepoulet, CH-1201 Geneva, Switzerland. [3]Department of Public Health and Infectious Diseases, University of Rome La Sapienza, Rome, Italy. [4]Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK.

Received: 25 June 2021 Accepted: 3 January 2022

References

1. World Health Organization. Tuberculosis Factsheet 2018.
2. World Health Organization. DS TB Treatment Factsheet 2017.

3. World Health Organization. Meeting report of the WHO expert consultation on drug-resistant tuberculosis treatment outcome definitions, 17–19 November 2020. In: World Health Organization [Internet]. 2020 p. 14. Available: https://apps.who.int/iris/handle/10665/340284
4. Trauner A, Borrell S, Reither K, Gagneux S. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. Drugs. 2014;74:1063–72. https://doi.org/10.1007/s40265-014-0248-y.
5. Safi H, Lingaraju S, Amin A, Kim S, Jones M, Holmes M, et al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes. Nat Genet. 2013;45:1190–7. https://doi.org/10.1038/ng.2743.
6. Gygli SM, Borrell S, Trauner A, Gagneux S. Antimicrobial resistance in mycobacterium tuberculosis: mechanistic and evolutionary perspectives. FEMS Microbiol Rev. 2017;41:354–73. https://doi.org/10.1093/femsre/fux011.
7. Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic determinants of drug resistance in *mycobacterium tuberculosis* and their diagnostic value. Am J Respir Crit Care Med. 2016;194:621–30. https://doi.org/10.1164/rccm.201510-2091OC.
8. Dheda K, Gumbo T, Maartens G, Dooley KE, McNerney R, Murray M, et al. The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. Lancet Respir Med. 2017;5:291–360. https://doi.org/10.1016/S2213-2600(17)30079-6.
9. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7:51. https://doi.org/10.1186/s13073-015-0164-0.
10. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med. 2019;11:41. https://doi.org/10.1186/s13073-019-0650-x.
11. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant mycobacterium tuberculosis. Nat Genet. 2018;50:307–16. https://doi.org/10.1038/s41588-017-0029-0.
12. Libiseller-Egger J, Phelan J, Campino S, Mohareb F, Clark TG. Robust detection of point mutations involved in multidrug-resistant mycobacterium tuberculosis in the presence of co-occurrent resistance markers. PLoS Comput Biol. 2020;16. https://doi.org/10.1371/journal.pcbi.1008518.
13. Kouchaki S, Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Wren J, editor. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty949.
14. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics. 2018;34:1666–71. https://doi.org/10.1093/bioinformatics/btx801.
15. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, et al. Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data. Front Genet. 2019;10. https://doi.org/10.3389/fgene.2019.00922.
16. Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, et al. DeepAMR for predicting co-occurrent resistance of mycobacterium tuberculosis. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz067.
17. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Deep learning predicts tuberculosis drug resistance status from whole-genome sequencing data. bioRxiv. 2018:275628. https://doi.org/10.1101/275628.
18. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of mycobacterium tuberculosis lineages for epidemiological and clinical studies. Genome Med. 2020;12:114. https://doi.org/10.1186/s13073-020-00817-3.
19. Minato Y, Thiede JM, Kordus SL, McKlveen EJ, Turman BJ, Baughn AD. Mycobacterium tuberculosis folate metabolism and the mechanistic basis for Para-aminosalicylic acid susceptibility and resistance. Antimicrobial agents and chemotherapy. American society for. Microbiology. 2015:5097–106. https://doi.org/10.1128/AAC.00647-15.
20. Chen J, Zhang S, Cui P, Shi W, Zhang W, Zhang Y. Identification of novel mutations associated with cycloserine resistance in mycobacterium tuberculosis. J Antimicrob Chemother. 2017;72:3272–6. https://doi.org/10.1093/jac/dkx316.
21. Vilchèze C, WR JJR. Resistance to isoniazid and Ethionamide in mycobacterium tuberculosis: genes, Mutations, and Causalities. Microbiol Spectr. 2014;2. https://doi.org/10.1128/microbiolspec.mgm2-0014-2013.
22. Gómez-González PJ, Perdigao J, Gomes P, Puyen ZM, Santos-Lazaro D, Napier G, et al. Genetic diversity of candidate loci linked to mycobacterium tuberculosis resistance to bedaquiline, delamanid and pretomanid. Sci Rep. 2021;11. https://doi.org/10.1038/s41598-021-98862-4.
23. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer New York; 2009. https://doi.org/10.1007/978-0-387-84858-7.
24. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis. 2011;91:8–13. https://doi.org/10.1016/J.TUBE.2010.09.006.

Publisher's Note

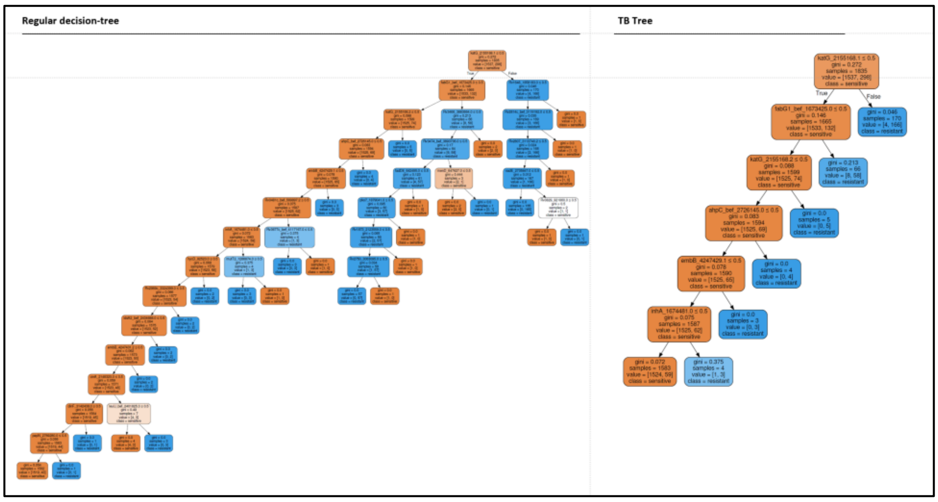
S1 Figure

Heatmap of included variant features in different sub-studies with colour-coding for subset of shared variants. The studies are highlighted on the x-axis. The y-axis has format of (gene, genomic position). The darker green colour are genomic variants that are in genes that were detected more than once across different studies (i.e. the gene has two or more filled cells in different columns in the diagram)

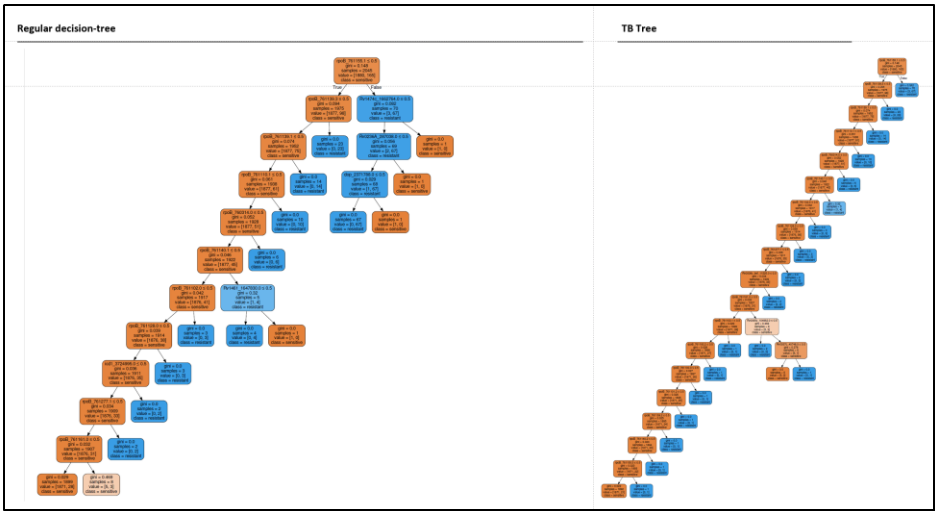
S2 Figure

Tree diagrams of regular classification tree (left) and Treesist-TB (right). The nodes in the trees are colour coded as blue (resistant) and orange (susceptible). Each node indicates the splitting variable, the improvement in purity (Gini), the total number of samples and the split over the left and the right nodes

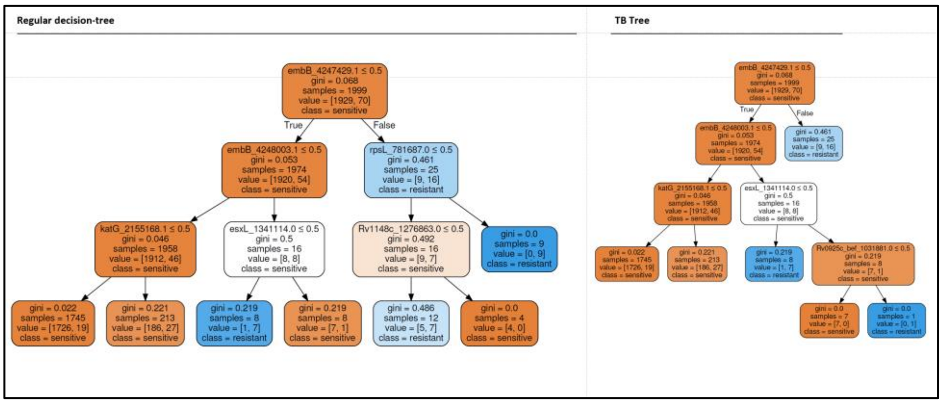
A) Isoniazid



B) Rifampicin



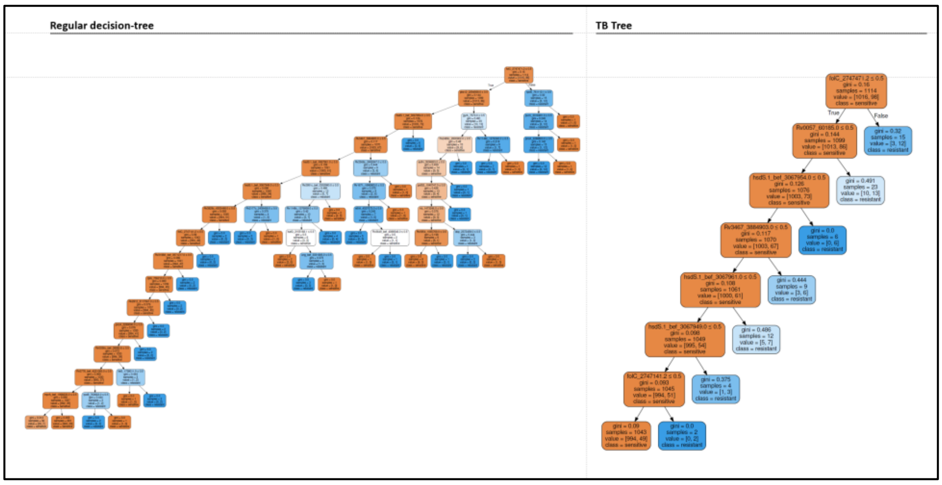
C) Ethambutol



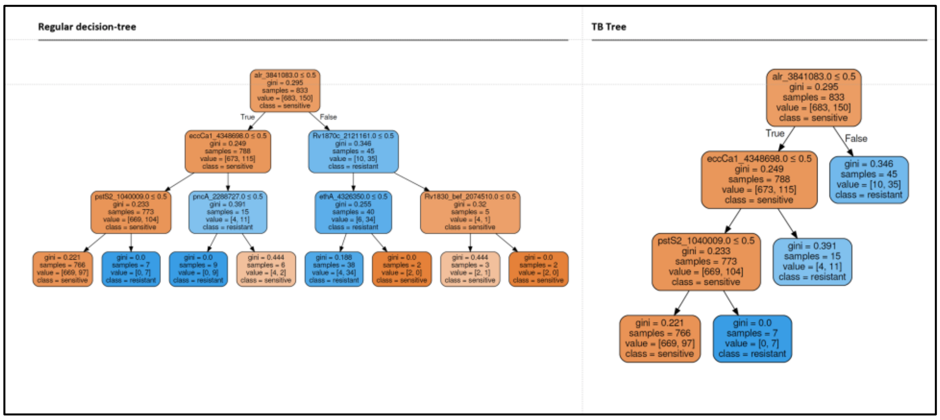
S3 Figure

Tree diagrams of regular classification tree (left) and Treesist-TB (right). The nodes in the trees are colour coded as blue (resistant) and orange (susceptible). Each node indicates the splitting variable, the improvement in purity (Gini), the total number of samples and the split over the left and the right nodes

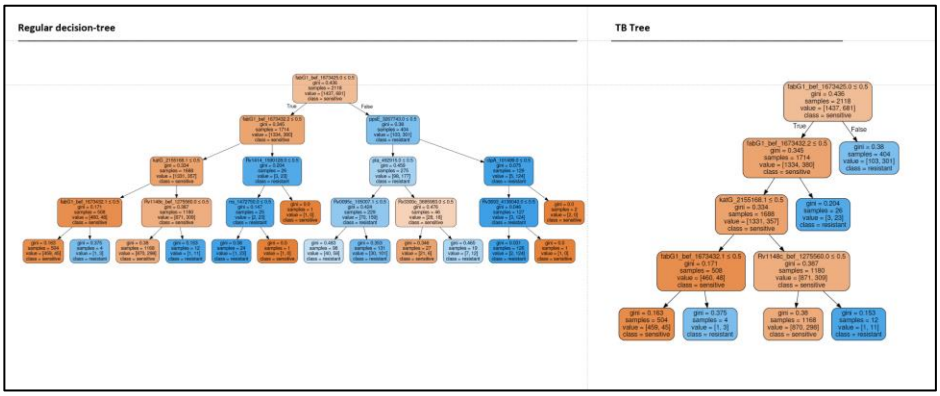
(A) Para-aminosalisylic acid



(B) Cycloserine



(C) Ethionamide



S1 Table
Phenotypic drug susceptibility tests status by lineage

Lineage	N	%	Susc.	RR-TB	MDR-TB	XDR-TB	Other DR
1	3155	9.6	2758	14	164	8	211
2	8260	25.2	5243	200	1913	473	431
3	3745	11.4	3053	19	439	37	197
4	16700	51.0	13686	191	1732	239	852
5	253	0.7	245	0	2	0	6
7	148	0.4	142	0	2	0	4
8	52	0.1	52	0	0	0	0
9	3	0	3	0	0	0	0
Other	373	1.1	283	0	3	0	87
Total (%)	32689	100	25465 (77.9%)	424 (1.3%)	4255 (13.0%)	757 (2.3%)	1788 (5.5%)

RR-TB rifampicin resistant; MDR-TB is defined as resistance to isoniazid and rifampicin; XDR-TB is defined as MDR-TB, with additional resistance to a fluoroquinolone and second-line injectable drug (pre-2021 definition).

S2 Table

Sources of sequence data, and drug resistance phenotypes

Drug	Total tests	# susc.	# resist.	% resist.	# studies	Study countries	Lineage	PMID
INH	1835	1537	298	16.2	6	Malawi, Philippines, Thailand	1-6	26116186,25336729, 25854485,25336729,31234910,31243306
RIF	2045	1880	165	8.1	7	Russia, Peru, South Korea, Philippines, Thailand	1-6	PMC3939361,26116186, 25854485, 31234910,31243306,27005572,27005572
EMB	1999	1929	70	3.5	5	Russia, Philippines, Thailand	1-6	PMC3939361,26116186, 25854485, 31234910,31243306
PAS	1114	1016	98	8.8	7	Portugal, South Africa, Uzbekistan, Russia, China, Belarus	1-4	30321294,29358649,29460750,26116186,28109869,PMC6685394,27903602
CYS	833	683	150	18.0	5	Portugal, China, Belarus, South Africa	1-4	30321294,29358649,28109869,27903602,30948181
ETN	2118	1437	681	32.2	16	Russia, Tunisia, China, Pakistan, Portugal, South Africa, Belarus	1-4	30321294,29358649,29460750,26116186,28109869,PMC6685394,27903602,30948181,30789128,PMC3939361,29358649,29358649,26418737,23995137,PMC3939361,PMC3939361

Susc. Susceptible; INH = Isoniazid, RIF = Rifampicin, PAS=para-aminosalisylic acid, CYS=cycloserine, ETH=ethionamide, EMB = Ethambutol

S3 Table

Summary of drugs and loci in TB-Profiler library

Drug	Locus	Gene	No. variants*	TB-Profiler SNPs	TB-Profiler Indels
Rifampicin	Rv0667	*rpoB*	115	94	25
	Rv0668	*rpoC*	92	8	-
Isoniazid	*Rv1483*	*fabG1*	22	11	-
	Rv1484	*inhA*	22	13	-
	Rv1908c	*katG*	93	226	37
	Rv2245	*kasA*	16	4	-
	Rv2428	*ahpC*	32	21	-
Ethambutol	*Rv1267c*	*embR*	34	20	-
	Rv3793	*embC*	85	25	-
	Rv3794	*embA*	112	9	6
	Rv3795	*embB*	125	127	1
Ethionamide	*Rv1483*	*fabG1*	20	3	-
	Rv1484	*inhA*	15	3	-
	Rv3854c	*ethA*	208	33	42
	Rv3855	*ethR*	25	2	-
PAS	*Rv2447c*	*folC*	23	18	-
	Rv2671	*ribD*	7	1	-
	Rv2754c	*thyX*	7	1	-
	Rv2764c	*thyA*	25	19	5
Cycloserine	*Rv2780*	*ald*	54	-	12
	Rv3423c	*alr*	22	3	-

* Number of genomic variants in the individual studies used

S4 Table

Frequency of Treesist-TB inferred variants in rifampicin, isoniazid, and ethambutol across 32k *Mycobacterium tuberculosis* isolates

Drug	Gene	Mutation	TB-Profiler**	Susc. %	MDR-TB %	XDR-TB %	Other resist. %
RIF	*rpoB*	N163K	no	-	<0.1	-	-
RIF	*rpoB*	V170F	Yes	-	1.0	<0.1	0.3
RIF	*rpoB*	L430P	Yes	-	1.7	1.3	0.7
RIF	*rpoB*	Q432K	yes	-	0.3	-	0.1
RIF	*rpoB*	Q432L	yes	<0.1	0.3	0.3	<0.1
RIF	*rpoB*	D435Y	yes	-	3.4	2.0	1.9
RIF	*rpoB*	D435V	yes	-	7.9	11.0	0.7
RIF	*rpoB*	S441L	yes	-	0.6	0.3	0.1
RIF	*rpoB*	H445D	yes	-	4.1	1.8	0.9
RIF	*rpoB*	H445N	yes	-	1.3	0.4	0.4
RIF	*rpoB*	H445Y	yes	-	5.5	2.5	2.2
RIF	*rpoB*	H445R	yes	-	2.1	0.9	0.2
RIF	*rpoB*	H445L	yes	-	1.4	0.8	0.2
RIF	*rpoB*	S450L	yes	<0.1	65.3	70.7	4.9
RIF	*rpoB*	L452P	yes	-	2.9	5.9	0.6
RIF	*rpoB*	I491F	yes	-	1.4	0.6	0.6
RIF	*rpoC*	N1239D	no	<0.1	-	-	-
RIF	*rpoC*	E1289A	no	<0.1	-	-	-
INH	*fabG1*	-126G>A	no	<0.1	16.8	34.6	12.6
INH	*katG*	Y597D	no	-	-	-	<0.1
INH	*katG*	T568P	no	<0.1	<0.1	-	-
INH	*katG*	A476V	no	<0.1	-	-	-
INH	*katG*	S315T	yes	<0.1	79.2	78.6	28.8
INH	*katG*	S315N	yes	-	1.8	1.3	1.1
INH	*katG*	S302R	yes	-	<0.1	<0.1	0.1
INH	*katG*	W300C	no	-	-	-	<0.1
INH	*katG*	G297V	yes	<0.1	<0.1	-	<0.1
INH	*katG*	P193fs	no	-	-	-	<0.1
INH	*katG*	L159F	no	<0.1	-	-	-
INH	*katG*	G156D	no	-	<0.1	-	-
INH	*katG*	A144V	no	<0.1	-	-	-
INH	*katG*	D142G	no	<0.1	<0.1	-	<0.1
INH	*katG*	L141F	yes	<0.1	<0.1	-	0.1
INH	*katG*	N138D	yes	-	<0.1	-	<0.1
INH	*katG*	A109V	yes	-	<0.1	-	<0.1
INH	*katG*	Y98C	no	<0.1	<0.1	-	0.2
INH	*ahpC*	-4359G>A	no	-	0.4	-	<0.1

INH	*ahpC*	-48G>A	yes	-	1.2	1.2	0.4
EMB	*embA*	-31delC	no	-	0.2	<0.1	-
EMB	*embA*	-16C>T	yes	-	2.1	4.4	0.2
EMB	*embA*	-16C>A	no	<0.1	0.8	0.7	<0.1
EMB	*embB*	M306V	yes	-	23.8	35.3	1.6
EMB	*embB*	M306L	yes	-	1.3	1.2	0.3
EMB	*embB*	M306I	yes	<0.1	20.2	26.7	3.0
EMB	*embB*	G406A	yes	-	6.6	6.5	0.4
EMB	*embB*	Q497K	yes	-	1.2	0.9	0.3
EMB	*embB*	Q497R	yes	-	5.6	8.0	0.5
EMB	*embB*	D1024N	yes	-	2.0	1.8	0.1

* from (50); INH = Isoniazid, RIF = Rifampicin, EMB = Ethambutol; RR-TB rifampicin resistant; MDR-TB multidrug resistant; XDR-TB Extensively drug resistant; ** underlined if mentioned in www.who.int/publications/i/item/9789240028173 as a high confidence (group 1) resistance mutation (sourced Nov. 2021)

S5 Table

Frequency of Treesist-TB inferred variants in para-aminosalisylic acid, cycloserine, and ethionamide across 32k *Mycobacterium tuberculosis* isolates*

Drug	Gene	mutation	TB-Profiler**	Susc. %	MDR-TB %	XDR-TB %	Other resist. %
PAS	*folC*	E153G	yes	-	0.3	0.4	0.0
PAS	*folC*	E153A	yes	-	0.2	0.3	<0.1
PAS	*folC*	S150G	yes	-	0.9	1.4	0.3
PAS	*folC*	S98G	no	-	0.0	0.3	0.0
PAS	*folC*	R49Q	no	-	0.8	0.3	0.2
PAS	*folC*	I43T	yes	-	0.7	3.1	0.2
PAS	*Rv2670c*	A5V	no	<0.1	4.5	6.1	0.8
PAS	*thyX*	-4C>T	no	<0.1	0.4	1.7	<0.1
PAS	*thyX*	-9G>A	no	<0.1	0.6	0.5	0.1
PAS	*thyX*	-16C>T	yes	<0.1	1.8	3.6	0.5
PAS	*thyX*	-18G>T	no	-	<0.1	0.2	<0.1
CYS	*rpoC*	D485Y	no	-	0.5	1.5	<0.1
CYS	*rpoC*	I491T	yes	-	1.3	4.3	<0.1
CYS	*alr*	Y388D	no	-	0.5	1.1	-
CYS	*alr*	L283P	no	<0.1	-	-	-
CYS	*alr*	L113R	yes	-	0.8	8.5	<0.1
CYS	*alr*	T20M	no	-	0.1	0.4	<0.1
ETH	*gyrA*	A90V	yes	<0.1	4.6	32.0	1.4
ETH	*gyrA*	S91P	yes	-	0.9	8.7	0.6
ETH	*gyrA*	D94A	yes	<0.1	2.3	12.8	0.4
ETH	*gyrA*	D94G	yes	<0.1	5.9	36.5	2.3
ETH	*mshA*	A133fs	no	-	<0.1	-	-
ETH	*mshA*	H175fs	no	-	<0.1	-	-
ETH	*mshA*	V237L	no	-	<0.1	-	-
ETH	*mshA*	A422V	no	-	<0.1	-	-
ETH	*fabG1*	-23G>C	no	-	<0.1	-	-
ETH	*fabG1*	-107G>A	no	<0.1	0.4	1.3	<0.1
ETH	*fabG1*	-126G>A	no	<0.1	16.8	34.6	12.6
ETH	*fabG1*	-133A>G	no	-	1.3	4.2	0.8
ETH	*fabG1*	-133A>T	no	-	1.4	5.7	0.3
ETH	*inhA*	I21T	yes	-	1.2	1.0	0.3
ETH	*inhA*	R27W	no	-	<0.1	-	-
ETH	*inhA*	I194T	yes	-	2.0	5.3	0.3
ETH	*inhA*	P251R	no	1.3	1.6	1.4	1.1
ETH	*ethA*	W455	no	-	0.1	0.6	<0.1
ETH	*ethA*	K448fs	no	-	0.1	<0.1	<0.1
ETH	*ethA*	P436fs	no	-	<0.1	-	-
ETH	*ethA*	A352fs	no	-	0.2	0.6	0.1
ETH	*ethA*	P334A	no	0.4	0.9	1.0	0.4

ETH	*ethA*	F320S	no	-	-	<0.1	-
ETH	*ethA*	L295fs	no	-	0.2	<0.1	-
ETH	*ethA*	C294	no	-	<0.1	-	-
ETH	*ethA*	R279	no	-	<0.1	0.2	-
ETH	*ethA*	Q269	yes	-	0.2	<0.1	-
ETH	*ethA*	M260I	no	<0.1	<0.1	-	0.1
ETH	*ethA*	W256	no	<0.1	0.6	2.0	0.1
ETH	*ethA*	C253F	no	-	-	0.2	-
ETH	*ethA*	T236fs	no	-	<0.1	-	-
ETH	*ethA*	Y235fs	no	-	0.2	0.6	<0.1
ETH	*ethA*	W228	no	<0.1	-	<0.1	-
ETH	*ethA*	N226fs	no	<0.1	<0.1	<0.1	-
ETH	*ethA*	K224	no	<0.1	<0.1	-	-
ETH	*ethA*	A222V	no	-	<0.1	-	-
ETH	*ethA*	S208L	no	<0.1	<0.1	-	-
ETH	*ethA*	R207G	yes	-	<0.1	0.2	<0.1
ETH	*ethA*	V202F	no	-	<0.1	0.2	-
ETH	*ethA*	L194P	no	-	<0.1	-	-
ETH	*ethA*	T186P	no	-	<0.1	<0.1	<0.1
ETH	*ethA*	P164R	no	-	<0.1	-	-
ETH	*ethA*	P160fs	no	-	0.4	<0.1	0.2
ETH	*ethA*	C137R	no	-	0.3	0.2	<0.1
ETH	*ethA*	C137R	no	-	0.3	0.2	<0.1
ETH	*ethA*	W116	no	<0.1	<0.1	0.3	-
ETH	*ethA*	K103fs	no	-	<0.1	-	-
ETH	*ethA*	W45	no	-	0.1	0.6	<0.1
ETH	*ethA*	K37fs	no	-	1.4	1.8	0.1
ETH	*ethA*	L35R	no	-	0.2	-	-
ETH	*ethA*	Q24	no	<0.1	0.8	0.9	0.1
ETH	*ethA*	D6fs	no	-	<0.1	-	-

* from (50); - refers to a frequency of zero; PAS=para-aminosalisylic acid, CYS=cycloserine, ETH=ethionamide; RR-TB rifampicin resistant; MDR-TB multidrug resistant; XDR-TB Extensively drug resistant; ** underlined if mentioned in www.who.int/publications/i/item/9789240028173 as a high confidence (group 1) resistance mutation (sourced Nov. 2021)

Chapter 4

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1701929	**Title**	Mr.
First Name(s)	Wouter		
Surname/Family Name	Deelder		
Thesis Title	Machine learning methods for infectious diseases: applications for tuberculosis and malaria.		
Primary Supervisor	Prof. Taane Clark, Dr. Luigi Palla		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Malaria Journal		
When was the work published?	June 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	N/A
Please list the paper's authors in the intended authorship order:	N/A
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I helped to conceive and design the study. I developed the haplo-imaging code library. I implemented the machine learning algorithms and analysed the data. I wrote the first draft of the manuscript, and finalised it after receiving revisions from co-authors and reviewers.

SECTION E

Student Signature	
Date	April 11, 2022

Supervisor Signature	
Date	April 11, 2022

Deelder *et al. Malar J (2021) 20:270*
https://doi.org/10.1186/s12936-021-03788-x

Malaria Journal

RESEARCH **Open Access**

Using deep learning to identify recent positive selection in malaria parasite sequence data

Wouter Deelder[1,2], Ernest Diez Benavente[1], Jody Phelan[1], Emilia Manko[1], Susana Campino[1], Luigi Palla[1,3†] and Taane G. Clark[1*†]

Abstract

Background: Malaria, caused by *Plasmodium* parasites, is a major global public health problem. To assist an understanding of malaria pathogenesis, including drug resistance, there is a need for the timely detection of underlying genetic mutations and their spread. With the increasing use of whole-genome sequencing (WGS) of *Plasmodium* DNA, the potential of deep learning models to detect loci under recent positive selection, historically signals of drug resistance, was evaluated.

Methods: A deep learning-based approach (called "*DeepSweep*") was developed, which can be trained on haplotypic images from genetic regions with known sweeps, to identify loci under positive selection. *DeepSweep* software is available from https://github.com/WDee/Deepsweep.

Results: Using simulated genomic data, *DeepSweep* could detect recent sweeps with high predictive accuracy (areas under ROC curve > 0.95). *DeepSweep* was applied to *Plasmodium falciparum* ($n = 1125$; genome size 23 Mbp) and *Plasmodium vivax* ($n = 368$; genome size 29 Mbp) WGS data, and the genes identified overlapped with two established extended haplotype homozygosity methods (within-population iHS, across-population Rsb) (~60–75% overlap of hits at $P < 0.0001$). *DeepSweep* hits included regions proximal to known drug resistance loci for both *P. falciparum* (e.g. *pfcrt*, *pfdhps* and *pfmdr1*) and *P. vivax* (e.g. *pvmrp1*).

Conclusion: The deep learning approach can detect positive selection signatures in malaria parasite WGS data. Further, as the approach is generalizable, it may be trained to detect other types of selection. With the ability to rapidly generate WGS data at low cost, machine learning approaches (e.g. *DeepSweep*) have the potential to assist parasite genome-based surveillance and inform malaria control decision-making.

Keywords: *Plasmodium falciparum*, *Plasmodium vivax*, Population genomics, Drug resistance, Machine learning, Positive selection

Background

Malaria, caused by *Plasmodium* parasites, is a major global health burden, with an estimated 229 million cases and 409,000 deaths in 2019 alone [1]. *Plasmodium falciparum* causes almost half of all malaria cases, and the majority of deaths are children in sub-Saharan Africa; *Plasmodium vivax* accounts for 65% of malaria cases in Asia and South America [1]. Malaria control involves a

---

*Correspondence: Taane.clark@lshtm.ac.uk
†Luigi Palla and Taane G. Clark Joint senior authors
[1] London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Full list of author information is available at the end of the article

combination of case management using diagnosis and treatment, and prevention using insecticide-treated nets, indoor residual spraying, and intermittent preventive treatment.

Resistance to anti-malarial medicines is a threat to the global efforts to control and eliminate malaria. Resistance originates from *Plasmodium* genetic mutations that increase in frequency over time and "sweep" through populations. During the past fifty years, several first-line treatments for *P. falciparum* malaria, including chloroquine and sulfadoxine-pyrimethamine (SP), have been rolled-out and then subsequently replaced due to the emergence of resistance. Recently, resistance to artemisinin has been reported in the form of delayed parasite clearance in Southeast Asia, posing a threat to the current first-line artemisinin-based combination therapy [2, 3]. For *P. vivax*, the spread of resistance to chloroquine, primaquine, mefloquine, and SP has been reported in various regions of the world [4, 5]. The underlying mutations causing resistance for *P. vivax* are less well defined than for *P. falciparum* [4–6].

Protecting and monitoring the efficacy of antimalarial treatments is a top priority for malaria endemic countries. There is a need to not only continuously monitor for drug resistance, which includes clinical reporting, but also to screen the parasite genome for known resistance mutations (e.g. in *P. falciparum*: *pfcrt (PF3D7_0709000), pfdhfr (PF3D7_0417200), pfdhps (PF3D7_0810800), pfmdr1 (PF3D7_0523000),* and *pfkelch13 (PF3D7_1343700)* [3]) and to identify potentially novel loci under putative positive selection. These insights are being facilitated by the characterization of genomic variation using whole-genome sequencing (WGS) across many *Plasmodium* isolates, and the subsequent application of statistical and population genomics methods to detect sweeps. In particular, sweeps can be identified through statistical approaches considering population differentiation, site-frequency spectra, or linkage disequilibrium and extended haplotype homozygosity (e.g. the within population integrated haplotype score (iHS), and the between population ratio (Rsb)) [7]. Whilst these methods have been developed for the human genome [8], they have been applied to *Plasmodium* and identified known genetic mutations contributing to drug resistance [9, 10]. Recently tools have been developed for the efficient computation of these statistics from WGS libraries, such as REHH, SweeD and OmegaPlus [11–13], but they require parameter optimization and their results are sensitive to the SNPs included, population definition, and to the statistical significance thresholds used to make inferences.

In recent years, researchers have explored the possibility of augmenting traditional approaches to the detection of selective sweeps with machine learning methods [14]. To date, sweep detection algorithms have been applied to pre-calculated population genetic statistics (e.g. Tajima's D, Fay and Wu's H) [7]. Gradient boosted decision trees and random tree classifiers have been trained on simulated data and applied to human 1000 Genomes Project data [15]. However, these methods do not solve the challenge of defining and calculating the population genetic statistics used as predictors of selection, a task which can be complex and time-consuming, especially when there are multiple sub-populations for cross-comparison. Deep (machine) learning methods may provide a viable alternative, and allow algorithms to learn through a hierarchy of features, where their definition and relationships can be inferred by the algorithm rather than externally defined [16]. The application of neural networks and deep learning has been explored within population genetics [17–19]. More generally, these methods are gaining traction in healthcare and biomedical settings, where enormous amounts of data are being generated, which contain extremely valuable signals and information, at a pace far surpassing what "traditional" methods of analysis can process [19].

The detection of recent positive selection seems amenable to deep learning approaches, where learning to recognize features in raw SNP data, such as the length and shape of shared haplotypes in genes with known sweeps within and between populations, may help to identify sweeps across the genome. The work presented applies a deep learning image-classification approach, which does not require prior extraction or selection of population-genetic statistics, to classify selective sweeps from "haplotypic" images. Using large *P. falciparum* (n = 1125) and *P. vivax* (n = 368) WGS datasets, partitioned into training and validation sets, the analysis shows that a deep learning approach (called "*DeepSweep*") calibrates well with other haplotype-based methods and other studies, and has the potential to detect novel signatures of positive selection.

Methods

Deep learning approach

DeepSweep is a deep learning model to detect instances of positive selection. It creates and analyses standardized images of the nearby genomic region around a given SNP. In brief, for each SNP of interest, and across all isolates, *DeepSweep* selects neighbouring SNPs at regularly spaced intervals, and subsequently sorts the remaining genomic matrix in alignment with the longest common haplotype, grouped for each population and for the reference and alternative alleles. The intuition is that SNPs that have undergone recent selective sweeps have a different haplotype structure resulting in distinct images (Additional file 1: Figure S1).

Model structure

DeepSweep uses a convolutional neural network (CNN) architecture, implemented using the Keras library (version 2.2.4) [20] in Python. The model was based on the AlexNet Classifier architecture, widely used for image analysis [21]. Through optimization, it was aimed to fit the smallest sized model (in terms of number of trainable parameters) that showed good predictive performance with low validation loss and high validation accuracy, but also detected features of interest, avoided overfitting, and minimized computational burden. Informally, overfitting is the training of a model that is too specifically tailored to (artefacts in) the training dataset and does not generalize well to unseen data. Statistically, within the framework of the bias–variance trade-off of a model, overfitting occurs where there is excessive variance resulting from an algorithm modelling the random noise in the training data [22]. The approach optimized over various hyper-parameters, including the number of convolutional layers (ranging from 1 to 5 layers), the number of filters (ranging from 2 to 96) and convolutional field sizes (ranging from 3 × 3 to 40 × 40). Regularization techniques (e.g. dropout [22]) were applied to prevent overfitting and support transferability. The model was trained to reduce binary cross-entropy between actual labels and estimated probabilities on images of known- and non-sweeps. The model structure was validated for 500 epochs. The final model has one convolutional layer, two dense layers, four convolutional filters, and a large convolutional field (40 × 9). The haplo-imaging algorithm and the machine learning analyses (Additional file 1: Figures S1, S2) were conducted in Python (version 2.7). The core packages for the machine learning were SnpEff (for annotating effect size) [23], SnpSift (for filtering VCF files) [24], PyVCF (for adjusting and creating VCF files) [25], SciPy and matplotlib (for image manipulation) and Tensorflow (version 1.15).

Simulated data

Sequence data was generated using SFS_Code software [26], which is a forward population genetic simulator. Simulated data corresponded to four sweep types ((i) recent—strong; (ii) recent—weak; (iii) historic—strong; (iv) partial) and compared to a Wright-Fisher "neutral" setting. The parameter settings are outlined (Additional file 1: Table S1), and lead to plausible scenarios for *Plasmodium* parasites [10]. For each comparison, 160 simulated datasets (128 training; 32 validation) were generated, each dataset with 4 populations of 100 parasite sequences (50% sweep, 50% neutral) and a locus length of at least 1kbp, where the mutation under selection was in the centre of the region. For the combined analysis of the sweep types, 640 simulated datasets (512 training, 128 validation) were used. These data were subsequently transformed into the aforementioned "haplotype images" that serve as input to the image classifier (Additional file 1: Figure S1). These haplo-images showed qualitatively discernible differences in features, with stronger or more recent sweeps leading to more "block-like" features (Additional file 1: Figure S3). The image classifier was trained on the simulated data, and classification accuracy and reduction of binomial loss were estimated. Simulated data was also used to illustrate the impact of changes in a subset of hyperparameters and confirmed that the final model had low validation loss and high validation accuracy (Additional file 1: Table S2).

Plasmodium sequencing data

Publicly available raw Illumina WGS data for *P. falciparum* (n = 1125) [27] and *P. vivax* (n = 368) [28], representing 11 malaria endemic countries (Additional file 1: Table S3; accession numbers in Additional file 1: Tables S4, S5). All samples were assessed by estMOI software [29] as either monoclonal or polyclonal samples with only a major dominant clone, to minimize the effects on analysis of multiplicity of infection. The *P. falciparum* and *P. vivax* sequences were mapped to the *Pf3D7* (23Mbp) and *PvP01* (29Mbp) reference genomes, respectively, using *bwa-mem* software (version 0.7.12; using default parameter settings) [30]. From the resulting alignments, SNPs and insertions and deletions (indels) were called from the consensus of *GATK* (version 4.1.4.1) [31] and *samtools* (version 1.9) [32] software (using default parameter settings), as applied in previous studies [4, 10]. SNPs were retained if they had < 10% missing alleles and a minor allele count greater than 4. The resulting dataset comprised of parasite genomes of *P. falciparum* (1,125 isolates, 74,757 SNPs) and of *P. vivax* (368 isolates, 126,596 SNPs). The number of missing values was 1,179,202 (2.9%) for *P. vivax* and 649,337 (1.2%) for *P. falciparum*. Missing alleles were imputed using the isolate with the longest shared haplotype around the missing position. An overview of the analytical approach is summarized (Additional file 1: Figure S2). The SnpEff tool (https://pcingola.github.io/SnpEff/) was used to annotate SNP variants and predict their effects on genes.

For *DeepSweep* model training, the presumed positive examples of positive selection are regions surrounding SNPs that are linked to drug resistance with an established scientific literature. For *P. falciparum*, these included regions around established SNPs in *pfcrt* (K76T, I356T; chloroquine), *pfdhfr* (N51I, C59R, S108N, I164L, S306F)/*pfdhps* (I431V, S436A, A437G, K540E/N, A581G, S613S) (SP), *pfmdr1* (N86Y; mefloquine, chloroquine), and *pfkelch13* (F446I, Y493H, P574L, R539T, and C580Y; artemisinin) [27]. For *P. vivax*, these included regions around some known SNPs in *pvdhps* (A553G, G383A, S382C/A) / *pvdhfr* (N50I, F57I/L, S/K58R, T61M, N117T/S) (putative SP) and *pvmdr1* (F1076L, Y976F, S698G, S513R; putative chloroquine) [4, 6]. This could be considered a relatively small number of training exemplars, which may lead to an increased risk that the implemented machine learning algorithm overfits due to potential artefacts in the training data. Therefore, for each *Plasmodium* species, "leave-one-group-out" cross-validation was implemented to understand the influence of individual training genes, where each single gene of the positive training examples was omitted in turn, with the model trained on the remaining genes [33]. The final model was fit on 80% of the data (split by SNPs), with 20% left as a hold-out set. The *DeepSweep* approach was compared to traditional haplotype-based statistics (iHS [34] and Rsb [35]), as calculated with the REHH package [36].

Results

Simulation study

Across the 4 different types of sweep simulations, the predictive accuracy was highest for more recent strong selection (97.1%), followed by weak selection (96.8%) and historic selection (88.2%) and partial selection (86.7%) (Table 1, Additional file 1: Figure S4). The total sensitivity across all sweeps combined was 89.1%, with a specificity of 93.8%, and an overall classification accuracy of 91.4%. The areas under the ROC curve were high for all simulations involving recent selection (>0.95; maximum 1), consistent with the high predictive ability of *DeepSweep*. The simulation results showed the potential utility of the approach when combining data across populations with common sweeps at difference stages.

Plasmodium falciparum DeepSweep analysis

The dataset comprised of 1,125 isolates and 74,757 SNPs. Most of these SNPs are in genic regions (76.5%), with 63.0% non-synonymous amino acid changes. Most SNPs have low minor allele frequencies (SNPs with MAF < 1%: 94.6%) (Additional file 1: Figure S5). The image classifier was trained on regions covering the established resistance SNPs in five genes, and found the models validated well using a leave-one-group-out approach. In particular, the overall accuracy was 83.6% (standard deviation 6.0%), where the performance was lower when *pfdhfr* was omitted (75.0%) and was higher when *pfdhps* (92.3%) was left out. One interpretation is that *pfdhfr* is under stronger selection than *pfdhps*, which would be consistent with *pfdhfr* N51I, C59R, S108N, I164L and S306F mutations underpinning key haplotypes underlying SP resistance [37]. The final model was fitted on 80% of the data, with 20% of the data used as a validation set, and demonstrated a strong performance both in terms of classification accuracy and reduction of binomial loss (Additional file 1: Figure S6). The trained classifier was then used to make predictions for the entire dataset of *P. falciparum* SNPs.

Table 1 Model performance based on simulated data

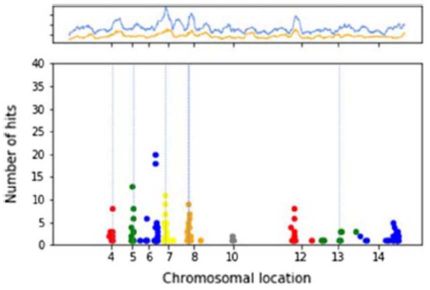
	Acc %	Sens %	Spec %	AUC
Stronger selection—recent sweep	97.1	93.8	100	1
Stronger selection—historic sweep	88.2	93.8	83.3	0.858
Weaker selection—recent sweep	96.8	100	93.3	1
Partial sweep	86.7	87.5	85.7	0.951
All sweeps combined	91.4	89.1	93.8	0.944

Acc accuracy, *Sens.* Sensitivity, *Spec.* specificity, *AUC* Area under the ROC Curve

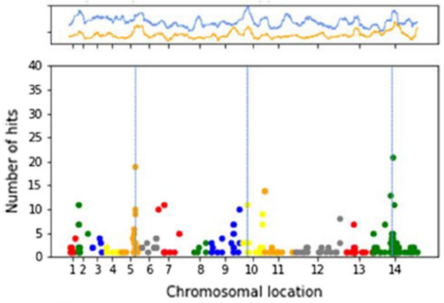
The deep learning model identified 387 SNPs in 160 genes (or ~ 2.9% of genes) as putatively under positive selection pressure in the wider dataset (Fig. 1). Further analysis focused on the subset of 11 genes that have > 6 hits (Table 2; see Additional file 1: Table S6 for the 26 genes with > 3 SNPs). Several peaks were in the vicinity of known drug-resistance genes in the training set, with nearby genes likely to be swept along (e.g. on *pfdhfr* on chromosome 4, *pfmdr1* on chromosome 5, *pfcrt* on chromosome 7, *pfdhps* on chromosome 8 and *pfkelch13* on chromosome 13). There is an additional peak on chromosome 6 that includes *Pk4* (*PF3D7*_0628200) and the HECT domain (*PF3D7*_0628100). Transcription of *Pk4* has been related to artemisinin-induced latency [38], and the HECT domain is thought to alter quinine and quinidine response, and likely co-selected with *pfcrt* [39]. There is a small peak on chromosome 10 (*PF3D7_1013500*) in the close vicinity of the gene encoding the autophagy-related protein 18 (*PF3D7_1012900*), which has been associated with artemisinin resistance. There is a peak on chromosome 12 (*PF3D7_1223500*) which has been putatively associated with SP resistance [40]. Smaller peaks were observed on chromosome 14 around *PF3D7_1462400*, which has been associated with chloroquine resistance [41].

Plasmodium vivax DeepSweep analysis

The dataset comprised of 368 isolates and 126,596 SNPs. Most of these SNPs are in genic regions (77.6%), with 42.5% non-synonymous amino acid changes. Many SNPs have low minor allele frequencies (SNPs with MAF < 1%: 77.6%) (Additional file 1: Figure S5). The image classifier was trained on the sixteen SNP mutations in the three genes. Using a leave-one-group-out validation approach, the overall accuracy was 79.7% (standard deviation 17.6%), and the performance was lower when



(a) P. falciparum



(b) P. vivax

Fig. 1 Number of *DeepSweep* Hits per locus across the 14 chromosomes and the relationship to the number of Rsb (top panel; blue line) and iHS (top panel; orange line) hits. **a** *P. falciparum*. Blue line is the running average of Rsb hits ($p < 0.0001$) over the nearest 100 locations. The orange line is the running average of iHS hits ($p < 0.0001$) over the nearest 100 locations. The vertical blue lines indicate *pfdhfr* (Chr. 4: 749,001, *pfmdr1* (Chr. 5: 960,020), *pfcrt* (Chr. 7: 404,770), *pfdhps* (Chr. 8: 549,408); *pfkelch13* (Chr. 13: 1,724,817). The tick-marks on the x-axis are chromosomal mid-points. **b** *P. vivax*. Top panel: Blue line is the running average of Rsb hits ($p < 0.0001$) over the nearest 100 locations. The orange line is the running average of iHS hits ($p < 0.0001$) over the nearest 100 locations. The vertical blue lines indicate *pvdhfr* (Chr. 5), *pvmdr1* (Chr. 10), *pvdhps* (Chr. 14). The tick-marks on the x-axis are chromosomal mid-points

pvmdr1 was omitted (57.1%) and was higher when *pvdhfr* was left out (100%). This difference is consistent with *pvmdr1* residues being strongly associated with chloroquine resistance [5] and, although, *pvdhfr* may contribute to SP drug resistance, there are very few published studies that associate genotypes of this locus with anti-folate susceptibility phenotypes [6]. As with *P. falciparum*, the trained model had strong performance both in terms of classification accuracy and reduction of binomial loss (Additional file 1: Figure S6). The model identified 577 hits in 237 genes (or ~4.3% of genes) as putatively under positive selection pressure in the wider dataset (Fig. 1). Further analysis focused on the subset of 19 genes that have > 6 hits (Table 3; see Additional file 1: Table S7 for the 35 genes with > 3 SNPs). Several loci are near the

Table 2 *Plasmodium falciparum* loci identified by *DeepSweep* (DS; with > 6 SNPs)

Chr	Gene ID (*PF3D7_*)	DS hits	iHS hits	Rsb hits
6	627800*	20	11	39
6	628100*	18	1	30
5	522400**	13		8
7	709100**	11		38
7	708200**	9		14
8	809600**	9	3	29
4	417400**	8		37
5	522900**	8		
12	1223500*	8		11
7	709300**	7		46
8	811200**	7		11

Chr Chromosome; iHS and Rsb counts defined as the number of SNPs in a gene that have an |iHS| or |Rsb| score with a p-value < 0.0001; *pfdhfr* (Chr. 4: 749,001, *pfmdr1* (Chr. 5: 960,020), *pfcrt* (Chr. 7: 404,770), *pfdhps* (Chr. 8: 549,408; * previously identified; ** close to known gene

Table 3 *Plasmodium vivax* loci identified by *DeepSweep* (DS; with > 6 SNPs)

Chr	Gene ID (*PVP01_*)	DS Hits	iHS hits	Rsb hits
14	1430700	21		
5	526800**	19	4	12
11	1101300	14		2
14	1428700**	13	1	5
2	202000	11		4
7	709800	11		0
10	1011000**	11		1
14	1432900	11	1	1
5	526400**	10	1	12
7	701100	10		8
9	948800	10		5
5	526300**	9	2	4
10	1034400	9		
12	1271500	8		
2	203000*	7		33
9	939900	7	2	1
10	1033900	7		
13	1317300	7		15
14	1418100	7		1

Chr Chromosome, iHS and Rsb Counts defined as the number of SNPs in a gene that have an iHS or Rsb score with a p-value < 0.0001; *pvdhfr* (Chr. 5), *pvmdr1* (Chr. 10), *pvdhps* (Chr. 14); * previously identified; ** close to known gene training genes (*pvdhfr* on chromosome 5, *pvmdr1* on chromosome 10, *pvdhps* on chromosome 14). Further, there was a peak around the gene encoding for the multidrug associated protein 1 (*pvmrp1*), which is a putative resistance candidate [4]. On chromosome 7, there was a peak around a gene coding for cysteine repeat modular protein 1, which is expressed in both vertebrate and mosquito hosts for host tissue targeting and invasion. This locus has been identified as presenting high population differentiation and under directional selective pressure in South America [4]. Finally, there was a larger region that was identified on chromosome 14, which contains *pvdhps* and a number of other genes that have been found in other analyses [4].

Comparison with established positive selection approaches

An analysis using the established REHH approach was performed, which involved the calculation of the integrated haplotype score (iHS) within populations and the associated Rsb values between pairs of populations (Additional file 1: Tables S8, S9). Although the REHH and *DeepSweep* methods have a different ranking of the strongest hits, there was an overall positive correlation between the number of hits from Rsb and *DeepSweep* (Pearson correlation: *P. falciparum* 0.49, *P. vivax* 0.20; Additional file 1: Figure S7). However, *DeepSweep* detected several novel loci that were not identified by REHH. These included loci on chromosomes 6 (*PF3D7_0611800*), 8 (*PF3D7_0811600*) and 14 (*PF3D7_1461800*) for *P. falciparum* (Additional file 1: Table S6), and on chromosomes 6 (PIR protein), 7 (cysteine repeat modular protein) and chromosome 14 for *P. vivax* (Additional file 1: Table S7). *PF3D7_0611800* has been linked to increased cytoadherence [42], *PF3D7_0811600* has previously been linked to SP resistance [40] and the genes coding for the PIR protein and the cysteine repeat protein have been associated with immune response and host invasion [43, 44]. There were several loci that were detected by EHH methods but not by *DeepSweep* (Additional file 1: Tables S8, S9). Some of the top hits included genes that are linked to immune response and host invasion (e.g. *PF3D7_1133400* AMA1, *PF3D7_1335100* MSP7). Other hits are housekeeping genes that are less likely to be under selective pressure (e.g. *PF3D7_0731800* (alpha/beta hydrolase), *PF3D7_1475900* (KELT protein), *PVP01_0202900* (18S) and *PVP01_1003700* (PPT)).

Discussion

The application of whole genome sequencing (WGS) is gaining traction across malaria endemic countries. With the resulting development of *Plasmodium* parasite genomic databases ("big data"), there is an opportunity for the implementation of machine learning methods to inform disease control. The detection of genomic signatures of selective sweeps resulting from the spread of mutations associated with anti-malarial drug resistance is one application of WGS data. This work presents a supervised (deep) learning approach (*DeepSweep*), which after being trained on haplotypic "images" of established drug resistance genes in *P. falciparum* and *P. vivax* parasites, resulted in the identification of loci known to be under recent positive selection. Whilst the strength of sweep signals per locus found by *DeepSweep* correlated with established EHH methods (e.g. between population Rsb), the machine learning approach has the advantage of not requiring a rigid definition and calculation of population-genetic statistics, incorporating information within and across populations, and relatively lower requirements for the pre-processing of raw SNP data. Like other machine learning approaches, it has the potential to scale up to large numbers of samples, and is parallelizable across genomic regions, thereby making it a potentially useful "big data" tool. In the absence of sufficient computational power, it is possible to develop sampling strategies that can select the subset of the data and samples that contain the highest density of information relevant to *DeepSweep*. Different model structures were assessed, but performance could be improved by further fine tuning of model hyperparameters (e.g. the number and size of the convolutional filters).

DeepSweep detected a set of loci not detected by the EHH methods, potentially because a deep learning approach can holistically incorporate information from the raw SNP data, which could be fragmented across separate populations and genomic windows, for the calculation of population-genetic statistics. Indeed, the simulation study demonstrated the potential of including haplo-images with not only single, but multiple populations, to allow the algorithm to take advantage of features that are common across regions and be robust to different stages of the sweeps. However, *DeepSweep* does require "representative" positive training examples, and in the context applied, assumes that the training drug resistance related loci have undergone or are undergoing selective sweeps in some of the populations. This assumption is not unrealistic given that some antimalarial drugs have been rolled out in different populations at different times resulting in differential stages of selective sweeps [40]. The *DeepSweep* and EHH approaches, as well as alternative methods (e.g. HaploPS [45]), can be considered complementary and could be run in parallel. However, as these approaches will increasingly use WGS, there are general challenges that affect variant-calling and ascertainment (e.g. extreme genome GC content), which can impact on the density and accuracy of genomic variant inputs, as well as the final population genomic analysis. Typically, WGS analysis leads to a dense set of well supported variants in robust genomic regions, with the application of calling algorithms incorporating information on known high quality polymorphisms [6]. Further, highly variable or problematic regions, such as *var* genes in *P. falciparum*, are typically removed from analysis [46]. In general, *DeepSweep* appeared to perform well across different GC content settings (*P. falciparum* 19%, *P. vivax* 58%), as well as in a simulated data setting which did not impose any constraint on GC content. However, in general, it is important to evaluate the quality of genomic variants used in an analysis. A further consideration is that most approaches use haplotype data, which in the human context require phasing from genotypes. Whilst the *Plasmodium* life cycle involves haploid asexual stages, complex clinical infections can complicate and confound population genetic analyses, and therefore analysis was restricted to infections with a dominant clone. However, it may be possible to extend *DeepSweep* to process individual parasite sequences for samples with multiplicity of infection. Irrespective, any novel loci identified should be confirmed through functional work [47]. Further, complementary methods that look at isolate relatedness, as determined by identity by descent (e.g. IsoRelate [48]), could also be implemented. New loci detected by *DeepSweep* that were not identified by other methods (e.g. on chromosomes 6, 8 and 14 for *P. falciparum* and on chromosomes 6, 7 and 14 for *P. vivax*) provide interesting candidates for confirmation studies.

A potential future opportunity is to apply models across species, for example, to detect *P. falciparum* loci after being trained on *P. vivax* signatures, and vice-versa. Such an application could assist to detect regions where drug resistance loci are unknown or less established, such as *P. vivax*. However, the impacts of differences in sample size and degree of polymorphism between species need to be considered. Relatedly, "real data" was used for training, but an alternative may be to use coalescent or forward-in-time simulation to create positive and negative labelled exemplars. However, there is a risk that images might not be representative of actual selective sweeps in nature. The deep learning algorithm has applications beyond positive selection, including for other evolutionary signatures (e.g. balancing selection) or application to other organisms (e.g. mosquitoes and humans).

Conclusions

The *DeepSweep* approach and the wider application of deep learning using genomic images constitutes a novel approach that shows promising results. It provides a robust, accessible and scalable approach for the identification of genomic regions under positive selection, and could assist with detecting established and new types of drug resistance. Thereby, providing insights into transmission dynamics and informing malaria control decision-making.

Abbreviations

AUC: Area under the ROC curve; CNN: Convolutional Neural Network; EHH: Extended Haplotype Homozygosity; Indels: Insertions and Deletions; iHS: Integrated Haplotype Score; ROC: Receiver Operating Characteristic; SNP: Single Nucleotide Polymorphism; SP: Sulfadoxine-Pyrimethamine; WGS: Whole Genome Sequencing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12936-021-03788-x.

Additional file 1: Table S1. Simulation parameters for the data generation using SFS_Code software. **Table S2**. Performance of Convolutional Neural Network (CNN) model structures on simulated datasets. **Table S3**. Sample origin by geographic location. **Table S4**. The 1,125 high-quality P. falciparum isolates used in this study. **Table S5**. The 368 high-quality P. vivax isolates used in the study. **Table S6**. Plasmodium falciparum loci identified by DeepSweep (DS; with >3 SNPs). **Table S7**. Plasmodium vivax loci identified by DeepSweep (DS; with >3 SNPs). **Table S8**. Plasmodium falciparum loci with the most iHS and Rsb hits. **Table S9**. Plasmodium vivax loci with the most iHS and Rsb hits. **Figure S1**. The creation of haplo-images. **Figure S2**. Workflow. **Figure S3**. Exemplar images of simulated isolates undergoing different types of sweeps or neutral evolution. **Figure S4**. Model performance on simulated datasets. **Figure S5**. Distribution of the minor allele frequencies across the SNPs. **Figure S6**. Model performance for Plasmodium falciparum and P. vivax on training and validation datasets. **Figure S7**. Relationship between -log10 p-value of Rsb hits and number of DeepSweep hits.

Acknowledgements

Oleg Tsybulniak and Aleksei Ponomarev provided support on Python coding.

Authors' contributions

WD, SC, LP and TGC conceived and designed the study. EDB, JP and EM performed the bioinformatic processing of the raw sequencing data. WD performed the population genetic and statistical analysis, under the supervision of LP and TGC. WD wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

Funding

TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC UK (Grant no. BB/R013063/1). SC is funded by the BloomsburySET and Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and MR/R020973/1).

Availability of data and materials

The WGS data is available from the European Nucleotide Archive (ENA) (see Additional file 1: Table S4, S5 for project accessions). Computing code is available from https://github.com/WDee/Deepsweep.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

WD was employed by the company Dalberg Advisors in Switzerland. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details

[1]London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Dalberg Advisors, 7 Rue de Chantepoulet, CH-1201 Geneva, Switzerland. [3]Department of Public Health and Infectious Diseases, University of Rome La Sapienza, Rome, Italy.

Received: 24 March 2021 Accepted: 29 May 2021

References

1. WHO. World Malaria Report. Geneva, World Health Organization, 2020.
2. Fairhurst RM, Dondorp AM. Artemisinin-resistant *Plasmodium falciparum* malaria. Microbiol Spectr. 2016;4:https://doi.org/10.1128/microbiolspec.ei10-0013-2016
3. Zhao Y, Liu Z, Myat Thu Soe, Wang L, Soe TN, Wei H, et al. Genetic variations associated with drug resistance markers in asymptomatic *Plasmodium falciparum* infections in Myanmar. Genes (Basel). 2019;10:692
4. Benavente ED, Ward Z, Chan W, Mohareb FR, Sutherland CJ, Roper C, et al. Genomic variation in *Plasmodium vivax* malaria reveals regions under selective pressure. PLoS One. 2017;12:e0177134
5. Ngassa Mbenda HG, Wang M, Guo J, Siddiqui FA, Hu Y, Yang Z, et al. Evolution of the *Plasmodium vivax* multidrug resistance 1 gene in the Greater Mekong Subregion during malaria elimination. Parasit Vectors. 2020;13:67.
6. Diez Benavente E, Manko E, Phelan J, Campos M, Nolder D, Fernandez D, et al. Distinctive genetic structure and selection patterns in *Plasmodium vivax* from South Asia and East Africa. Nat Commun. 2021;12:3160.
7. Nielsen R. Molecular Signatures of Natural Selection SNP: single nucleotide polymorphism. Annu Rev Genet. 2005;39:197–218.
8. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120.
9. Ocholla H, Preston MD, Mipando M, Jensen ATR, Campino S, MacInnis B, et al. Whole-genome scans provide evidence of adaptive evolution in Malawian *Plasmodium falciparum* isolates. J Infect Dis. 2014;210:1991–2000.
10. Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG. Imputation-based population genetics analysis of *Plasmodium falciparum* malaria parasites. PLoS Genet. 2015;11:e1005131.
11. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour. 2017;17:78–90.
12. Pavlidis P, Živković D, Stamatakis A, Alachiotis N. SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol. 2013;30:2224–34.
13. Alachiotis N, Stamatakis A, Pavlidis P. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics. 2012;28:2274–5.
14. Hahn MW. Molecular population genetics. Oxford University Press (OUP); 2018.
15. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. 2015;31:3946–52.
16. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.
17. Chan J, Perrone V, Spence JP, Jenkins PA, Mathieson S, Song YS. A likelihood-free inference framework for population genetic data using exchangeable neural networks. Adv Neural Inf Process Syst. 2018;31:8594–605.
18. Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol Biol Evol. 2019;36:220–38.
19. Sanchez T, Cury J, Charpiat G, Jay F. Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. bioRxiv. 2020; 2020.01.20.910539.
20. Chollet F. Keras. Github; 2015. Available from: https://github.com/fchollet/keras
21. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60:84–90.
22. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929–58.
23. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6:80–92.
24. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program. SnpSift Front Genet. 2012;3:35.
25. Casbon J. PyVCF-A Variant Call Format Parser for Python. Github; 2012. Available from: https://github.com/jamescasbon/PyVCF
26. Hernandez RD. A flexible forward simulator for populations subject to selection and demography. Bioinformatics. 2008;24:2786–7.
27. Ravenhall M, Benavente ED, Sutherland CJ, Baker DA, Campino S, Clark TG. An analysis of large structural variation in global *Plasmodium falciparum* isolates identifies a novel duplication of the chloroquine resistance associated gene. Sci Rep. 2019;9:8287.
28. Diez Benavente E, Campos M, Phelan J, Nolder D, Dombrowski JG, Marinho CRF, et al. A molecular barcode to inform the geographical origin and transmission dynamics of *Plasmodium vivax* malaria. PLoS Genet. 2020;16:e1008576.
29. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. EstMOI: Estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics. 2014;30:1292–4.
30. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; arXiv:1303.3997v2.
31. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
32. Li H. Improving SNP discovery by base alignment quality. Bioinformatics. 2011;27:1157–8.
33. Mordelet F, Vert JP. ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics. 2011;12:389.
34. Voight BF, Kudaravalli S, Wen X, Pritchard JK, Diamond J, Jobling M, et al. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e72.
35. Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 2007;5:1587–602.
36. Gautier M, Vitalis R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics. 2012;28:1176–7.
37. Turkiewicz A, Manko E, Sutherland CJ, Benavente ED, 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:e1009268
38. Zhang M, Gallego-Delgado J, Fernandez-Arias C, Waters NC, Rodriguez A, Tsuji M, et al. Inhibiting the *Plasmodium* eIF2α kinase PK4 prevents artemisinin-induced latency. Cell Host Microbe. 2017;22:766-776.e4.
39. Sanchez CP, Liu C-H, Mayer S, Nurhasanah A, Cyrklaff M, Mu J, et al. A HECT ubiquitin-protein ligase as a novel candidate gene for altered quinine and quinidine responses in *Plasmodium falciparum*. PLoS Genet. 2014;10:e1004382.
40. Ravenhall M, Benavente ED, Mipando M, Jensen ATR, Sutherland CJ, Roper C, et al. Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the *Plasmodium falciparum* population in Malawi. Malar J. 2016;15:575.
41. Pulcini S, Staines HM, Lee AH, Shafik SH, Bouyer G, Moore CM, et al. Mutations in the *Plasmodium falciparum* chloroquine resistance transporter, PfCRT, enlarge the parasite's food vacuole and alter drug sensitivities. Sci Rep. 2015;5:14552.

42. Sedillo J. Pathogenic mechanisms and signaling pathways in *Plasmodium falciparum*. Grad Theses Dissertation, University of South Florida. 2014.
43. França CT, He W-Q, Gruszczyk J, Lim NTY, Lin E, Kiniboro B, et al. *Plasmodium vivax* reticulocyte binding proteins are key targets of naturally acquired immunity in young Papua New Guinean children. PLoS Negl Trop Dis. 2016;10:e0005014.
44. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, et al. Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. Nat Genet. 2016;48:953–8.
45. Liu X, Ong RTH, Pillai EN, Elzein AM, Small KS, Clark TG, et al. Detecting and characterizing genomic signatures of positive selection in global populations. Am J Hum Genet. 2013;92:866–81.
46. Benavente ED, Oresegun DR, de Sessions PF, Walker EM, Roper C, Dombrowski JG, et al. Global genetic diversity of var2csa in *Plasmodium falciparum* with implications for malaria in pregnancy and vaccine development. Sci Rep. 2018;8:15429.
47. Mohring F, Hart MN, Rawlinson TA, Henrici R, Charleston JA, Diez Benavente E, et al. Rapid and iterative genome editing in the malaria parasite *Plasmodium knowlesi* provides new tools for *P. vivax* research. Elife. 2019;8:e45829.
48. Henden L, Lee S, Mueller I, Barry A, Bahlo M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLoS Genet. 2018;14:e1007279.

Publisher's Note

S1 Table

Simulation parameters for the data generation using SFS_Code software

Scenario	**No. of Simulated datasets***	**Selection Coefficient**	**Time window*****
Stronger selection – recent sweep	80	2000	0.20
Stronger selection – historic sweep	80	2000	0.40
Weaker selection – recent sweep	80	400	0.20
Partial sweep**	80	100	0.15
Neutral	80 x 4	0	0.20

* Each simulated dataset consisted of 4 populations of 100 individual parasites (50% sweep; 50% neutral), and a locus length of 1kbp, with the mutation inserted at position 500. The mutation rate per site was set to 0.01 and the recombination rate per site was 0.02; ** The partial sweeps were created with rejection sampling, where only sweeps that had a derived allele frequency of between 33% and 80% were kept; *** The window is the time between the introduction of the mutation and sampling expressed in 2N generations, with N being the population size.

S2 Table

Performance of Convolutional Neural Network (CNN) model structures on simulated datasets

Model	Description of model and changes*	Trainable parameters	Validation Loss**	Validation Accuracy (%)
1	Model used in our study (one convolutional layer with 4 filters, with respective filter size of (40,9) followed by two drop-out and dense layers with ReLu activation)	4,525	0.33	93.8
2	Two convolutional layers with 4 filters in each layer, with respective kernel size of 11 and 5	78,277	0.42	75.0
3	Three convolutional layers with 4 filters in each layer, with respective kernel sizes of 11, 5 and 3	71,705	0.49	62.5
4	Increase of number of convolutional filters to 8	15,193	0.26	87.5
5	Decrease of number of convolutional filters to 2	1,495	0.59	68.8
6	Increase of drop-out rate to 0.4 (from 0.2)	4,525	0.32	87.5
7	Decrease of drop-out rate to 0.1 (from 0.2)	4,525	0.48	75.0

* Model 1 is the final model used across our datasets, and Models 2 - 7 are deviations from this; ReLu: Rectified Linear Unit. ** Validation loss as measured by binary cross-entropy. The performance (validation loss, validation accuracy) was measured by training on 64 simulated isolates with the same characteristics as other simulations (e.g. undergoing different forms of sweeps as well as neutral evolution) with all performance metrics measured on a validation set of 16 isolates.

S3 Table

Sample origin by geographic location

Country	*P. falciparum* N*	*P. falciparum* %	*P. vivax* N**	*P. vivax* %
Cambodia	351	31.2	32	8.7
Malawi	221	19.6	-	
Ghana	202	18.0	-	
Vietnam	187	16.6	-	
Thailand	164	14.6	128	34.8
Peru	-	-	58	15.8
Malaysia	-	-	50	13.6
Colombia	-	-	30	8.2
Papua New Guinea	-	-	26	7.1
Mexico	-	-	20	5.4
Ethiopia	-	-	24	6.5
Total	1,125	100	368	100

* see **S4 Table** for a list of sequence data accession numbers

** see **S5 Table** for a list of sequence data accession numbers

S4 Table. The 1,125 high-quality *Plasmodium falciparum* isolates used in this study

Country	Identifier	Country	Identifier	Country	Identifier	Country	Identifier
Cambodia	ERS009721	Cambodia	ERS014167	Cambodia	ERS028699	Cambodia	ERS032695
Cambodia	ERS010057	Cambodia	ERS014168	Cambodia	ERS028700	Cambodia	ERS032696
Cambodia	ERS010059	Cambodia	ERS014169	Cambodia	ERS028701	Cambodia	ERS032697
Cambodia	ERS010062	Cambodia	ERS014170	Cambodia	ERS028702	Cambodia	ERS045932
Cambodia	ERS010066	Cambodia	ERS014171	Cambodia	ERS028703	Cambodia	ERS045933
Cambodia	ERS010067	Cambodia	ERS014172	Cambodia	ERS028704	Cambodia	ERS045934
Cambodia	ERS010216	Cambodia	ERS014173	Cambodia	ERS028705	Cambodia	ERS045935
Cambodia	ERS010217	Cambodia	ERS017698	Cambodia	ERS028706	Cambodia	ERS045936
Cambodia	ERS010318	Cambodia	ERS017699	Cambodia	ERS028707	Cambodia	ERS045937
Cambodia	ERS010319	Cambodia	ERS017700	Cambodia	ERS028708	Cambodia	ERS045938
Cambodia	ERS010320	Cambodia	ERS017702	Cambodia	ERS028709	Cambodia	ERS045939
Cambodia	ERS010321	Cambodia	ERS017705	Cambodia	ERS028710	Cambodia	ERS050859
Cambodia	ERS010322	Cambodia	ERS017706	Cambodia	ERS028712	Cambodia	ERS050860
Cambodia	ERS010323	Cambodia	ERS017707	Cambodia	ERS028713	Cambodia	ERS050865
Cambodia	ERS010324	Cambodia	ERS023736	Cambodia	ERS028714	Cambodia	ERS050866
Cambodia	ERS010325	Cambodia	ERS023737	Cambodia	ERS028715	Cambodia	ERS050870
Cambodia	ERS010327	Cambodia	ERS023738	Cambodia	ERS028716	Cambodia	ERS050872
Cambodia	ERS010330	Cambodia	ERS023739	Cambodia	ERS028717	Cambodia	ERS050885
Cambodia	ERS010331	Cambodia	ERS023740	Cambodia	ERS028718	Cambodia	ERS050890
Cambodia	ERS010332	Cambodia	ERS023741	Cambodia	ERS028719	Cambodia	ERS052777
Cambodia	ERS010333	Cambodia	ERS023742	Cambodia	ERS028720	Cambodia	ERS052778
Cambodia	ERS010334	Cambodia	ERS023743	Cambodia	ERS028721	Cambodia	ERS052781
Cambodia	ERS010335	Cambodia	ERS023744	Cambodia	ERS028722	Cambodia	ERS052784
Cambodia	ERS010336	Cambodia	ERS023745	Cambodia	ERS028723	Cambodia	ERS052785
Cambodia	ERS010337	Cambodia	ERS023746	Cambodia	ERS028724	Cambodia	ERS052790
Cambodia	ERS010346	Cambodia	ERS023747	Cambodia	ERS028725	Cambodia	ERS071770
Cambodia	ERS010511	Cambodia	ERS023748	Cambodia	ERS032018	Cambodia	ERS071771
Cambodia	ERS010516	Cambodia	ERS023749	Cambodia	ERS032022	Cambodia	ERS071773
Cambodia	ERS010590	Cambodia	ERS023750	Cambodia	ERS032026	Cambodia	ERS071774
Cambodia	ERS010592	Cambodia	ERS024123	Cambodia	ERS032029	Cambodia	ERS071776
Cambodia	ERS010598	Cambodia	ERS025100	Cambodia	ERS032037	Cambodia	ERS071777
Cambodia	ERS010669	Cambodia	ERS025257	Cambodia	ERS032050	Cambodia	ERS071778
Cambodia	ERS010672	Cambodia	ERS025258	Cambodia	ERS032060	Cambodia	ERS071779
Cambodia	ERS010673	Cambodia	ERS025259	Cambodia	ERS032109	Cambodia	ERS071781
Cambodia	ERS010779	Cambodia	ERS025262	Cambodia	ERS032135	Cambodia	ERS071782
Cambodia	ERS010786	Cambodia	ERS025263	Cambodia	ERS032243	Cambodia	ERS071784
Cambodia	ERS013843	Cambodia	ERS025265	Cambodia	ERS032246	Cambodia	ERS071787
Cambodia	ERS013844	Cambodia	ERS025266	Cambodia	ERS032249	Cambodia	ERS071791
Cambodia	ERS014154	Cambodia	ERS025267	Cambodia	ERS032255	Cambodia	ERS071794
Cambodia	ERS014155	Cambodia	ERS025268	Cambodia	ERS032258	Cambodia	ERS071796
Cambodia	ERS014156	Cambodia	ERS025269	Cambodia	ERS032683	Cambodia	ERS071797
Cambodia	ERS014157	Cambodia	ERS025271	Cambodia	ERS032686	Cambodia	ERS071799
Cambodia	ERS014158	Cambodia	ERS025272	Cambodia	ERS032687	Cambodia	ERS071800
Cambodia	ERS014159	Cambodia	ERS025273	Cambodia	ERS032688	Cambodia	ERS071809
Cambodia	ERS014161	Cambodia	ERS025274	Cambodia	ERS032689	Cambodia	ERS071812
Cambodia	ERS014162	Cambodia	ERS025275	Cambodia	ERS032690	Cambodia	ERS071815
Cambodia	ERS014163	Cambodia	ERS025276	Cambodia	ERS032691	Cambodia	ERS071824
Cambodia	ERS014164	Cambodia	ERS028696	Cambodia	ERS032692	Cambodia	ERS088705

Cambodia	ERS014165	Cambodia	ERS028697	Cambodia	ERS032693	Cambodia	ERS088706
Cambodia	ERS014166	Cambodia	ERS028698	Cambodia	ERS032694	Cambodia	ERS140927
Cambodia	ERS140928	Cambodia	ERS174484	Cambodia	ERS010778	Ghana	ERS009734
Cambodia	ERS140935	Cambodia	ERS174485	Cambodia	ERS010782	Ghana	ERS010081
Cambodia	ERS140936	Cambodia	ERS174486	Cambodia	ERS010788	Ghana	ERS010083
Cambodia	ERS141406	Cambodia	ERS174488	Cambodia	ERS017561	Ghana	ERS010084
Cambodia	ERS141408	Cambodia	ERS174489	Cambodia	ERS024125	Ghana	ERS010085
Cambodia	ERS141414	Cambodia	ERS174492	Cambodia	ERS024126	Ghana	ERS010086
Cambodia	ERS141415	Cambodia	ERS174511	Cambodia	ERS024127	Ghana	ERS010087
Cambodia	ERS141417	Cambodia	ERS174512	Cambodia	ERS024128	Ghana	ERS010088
Cambodia	ERS141474	Cambodia	ERS174567	Cambodia	ERS024129	Ghana	ERS010089
Cambodia	ERS141480	Cambodia	ERS174569	Cambodia	ERS024130	Ghana	ERS010090
Cambodia	ERS141501	Cambodia	ERS174570	Cambodia	ERS024133	Ghana	ERS010124
Cambodia	ERS141502	Cambodia	ERS174571	Cambodia	ERS024135	Ghana	ERS010125
Cambodia	ERS142851	Cambodia	ERS174572	Cambodia	ERS024136	Ghana	ERS011021
Cambodia	ERS143415	Cambodia	ERS174573	Cambodia	ERS031999	Ghana	ERS011022
Cambodia	ERS143416	Cambodia	ERS174574	Cambodia	ERS032058	Ghana	ERS011023
Cambodia	ERS143417	Cambodia	ERS174575	Cambodia	ERS032071	Ghana	ERS011025
Cambodia	ERS143418	Cambodia	ERS174576	Cambodia	ERS032077	Ghana	ERS011026
Cambodia	ERS143420	Cambodia	ERS174577	Cambodia	ERS032082	Ghana	ERS011027
Cambodia	ERS143423	Cambodia	ERS174578	Cambodia	ERS032091	Ghana	ERS013064
Cambodia	ERS143425	Cambodia	ERS174580	Cambodia	ERS032093	Ghana	ERS013065
Cambodia	ERS143430	Cambodia	ERS175810	Cambodia	ERS032136	Ghana	ERS013066
Cambodia	ERS143432	Cambodia	ERS193641	Cambodia	ERS032637	Ghana	ERS013067
Cambodia	ERS143434	Cambodia	ERS199592	Cambodia	ERS010789	Ghana	ERS013068
Cambodia	ERS143436	Cambodia	ERS199597	Cambodia	ERS010790	Ghana	ERS013069
Cambodia	ERS143443	Cambodia	ERS199602	Cambodia	ERS024137	Ghana	ERS013071
Cambodia	ERS143444	Cambodia	ERS199607	Cambodia	ERS024139	Ghana	ERS013072
Cambodia	ERS143446	Cambodia	ERS199612	Cambodia	ERS024140	Ghana	ERS013073
Cambodia	ERS143447	Cambodia	ERS199617	Cambodia	ERS024142	Ghana	ERS013074
Cambodia	ERS143452	Cambodia	ERS199622	Cambodia	ERS024143	Ghana	ERS013075
Cambodia	ERS143453	Cambodia	ERS199627	Cambodia	ERS024144	Ghana	ERS013076
Cambodia	ERS143456	Cambodia	ERS199632	Cambodia	ERS024145	Ghana	ERS013077
Cambodia	ERS143457	Cambodia	ERS199637	Cambodia	ERS024146	Ghana	ERS013078
Cambodia	ERS143459	Cambodia	ERS224869	Cambodia	ERS024147	Ghana	ERS013079
Cambodia	ERS143461	Cambodia	ERS224889	Cambodia	ERS024148	Ghana	ERS013080
Cambodia	ERS143468	Cambodia	ERS224899	Cambodia	ERS024149	Ghana	ERS013081
Cambodia	ERS143470	Cambodia	ERS224904	Cambodia	ERS024150	Ghana	ERS013082
Cambodia	ERS143483	Cambodia	ERS224909	Cambodia	ERS024151	Ghana	ERS013091
Cambodia	ERS143486	Cambodia	ERS224914	Cambodia	ERS024152	Ghana	ERS013092
Cambodia	ERS143490	Cambodia	ERS336365	Cambodia	ERS032003	Ghana	ERS013093
Cambodia	ERS143491	Cambodia	ERS336376	Cambodia	ERS032017	Ghana	ERS013094
Cambodia	ERS143492	Cambodia	ERS024141	Cambodia	ERS032031	Ghana	ERS013095
Cambodia	ERS143500	Cambodia	ERS032014	Cambodia	ERS032039	Ghana	ERS013096
Cambodia	ERS143504	Cambodia	ERS009740	Cambodia	ERS032056	Ghana	ERS013097
Cambodia	ERS143512	Cambodia	ERS009744	Cambodia	ERS032106	Ghana	ERS013098
Cambodia	ERS143515	Cambodia	ERS009745	Cambodia	ERS032108	Ghana	ERS013099
Cambodia	ERS164616	Cambodia	ERS010054	Cambodia	ERS032638	Ghana	ERS013100
Cambodia	ERS164617	Cambodia	ERS010155	Cambodia	ERS157463	Ghana	ERS013101
Cambodia	ERS164619	Cambodia	ERS010156	Ghana	ERS009723	Ghana	ERS017384

Cambodia	ERS164622	Cambodia	ERS010670	Ghana	ERS009725	Ghana	ERS017385
Cambodia	ERS164623	Cambodia	ERS010671	Ghana	ERS009727	Ghana	ERS017386
Cambodia	ERS164624	Cambodia	ERS010776	Ghana	ERS009728	Ghana	ERS017387
Cambodia	ERS174483	Cambodia	ERS010777	Ghana	ERS009730	Ghana	ERS017388
Ghana	ERS017389	Ghana	ERS022963	Ghana	ERS032185	Malawi	ERS032653
Ghana	ERS017390	Ghana	ERS022964	Ghana	ERS032188	Malawi	ERS032654
Ghana	ERS017391	Ghana	ERS022965	Ghana	ERS032189	Malawi	ERS032657
Ghana	ERS017392	Ghana	ERS022966	Ghana	ERS032190	Malawi	ERS040098
Ghana	ERS017393	Ghana	ERS022967	Ghana	ERS032191	Malawi	ERS040099
Ghana	ERS017394	Ghana	ERS022968	Ghana	ERS032192	Malawi	ERS040100
Ghana	ERS017395	Ghana	ERS022969	Ghana	ERS032195	Malawi	ERS040101
Ghana	ERS017396	Ghana	ERS022970	Ghana	ERS032196	Malawi	ERS040103
Ghana	ERS017397	Ghana	ERS022971	Ghana	ERS032199	Malawi	ERS053866
Ghana	ERS017398	Ghana	ERS022972	Ghana	ERS032201	Malawi	ERS053871
Ghana	ERS017399	Ghana	ERS022973	Ghana	ERS032202	Malawi	ERS053875
Ghana	ERS017400	Ghana	ERS022974	Ghana	ERS032204	Malawi	ERS053876
Ghana	ERS017401	Ghana	ERS022975	Ghana	ERS032205	Malawi	ERS053877
Ghana	ERS017402	Ghana	ERS022976	Ghana	ERS032212	Malawi	ERS053938
Ghana	ERS022744	Ghana	ERS022977	Ghana	ERS032213	Malawi	ERS053940
Ghana	ERS022745	Ghana	ERS022978	Ghana	ERS032215	Malawi	ERS053942
Ghana	ERS022746	Ghana	ERS022979	Ghana	ERS032218	Malawi	ERS053944
Ghana	ERS022747	Ghana	ERS022981	Ghana	ERS032219	Malawi	ERS053945
Ghana	ERS022748	Ghana	ERS022982	Ghana	ERS032220	Malawi	ERS053947
Ghana	ERS022749	Ghana	ERS022984	Ghana	ERS032221	Malawi	ERS053948
Ghana	ERS022750	Ghana	ERS022986	Ghana	ERS032222	Malawi	ERS053949
Ghana	ERS022751	Ghana	ERS031998	Ghana	ERS032223	Malawi	ERS053950
Ghana	ERS022754	Ghana	ERS032001	Ghana	ERS032226	Malawi	ERS053952
Ghana	ERS022755	Ghana	ERS032002	Ghana	ERS032227	Malawi	ERS053953
Ghana	ERS022756	Ghana	ERS032007	Ghana	ERS032229	Malawi	ERS053954
Ghana	ERS022757	Ghana	ERS032011	Ghana	ERS032230	Malawi	ERS053955
Ghana	ERS022758	Ghana	ERS032012	Ghana	ERS032231	Malawi	ERS053956
Ghana	ERS022760	Ghana	ERS032028	Ghana	ERS032232	Malawi	ERS053957
Ghana	ERS022761	Ghana	ERS032030	Ghana	ERS032233	Malawi	ERS053958
Ghana	ERS022762	Ghana	ERS032032	Ghana	ERS032236	Malawi	ERS053960
Ghana	ERS022764	Ghana	ERS032033	Ghana	ERS032238	Malawi	ERS053878
Ghana	ERS022765	Ghana	ERS032034	Ghana	ERS032239	Malawi	ERS053880
Ghana	ERS022768	Ghana	ERS032035	Ghana	ERS032667	Malawi	ERS053890
Ghana	ERS022770	Ghana	ERS032038	Ghana	ERS032668	Malawi	ERS053891
Ghana	ERS022771	Ghana	ERS032044	Ghana	ERS032669	Malawi	ERS053895
Ghana	ERS022773	Ghana	ERS032047	Ghana	ERS032670	Malawi	ERS053897
Ghana	ERS022774	Ghana	ERS032053	Ghana	ERS032671	Malawi	ERS055901
Ghana	ERS022942	Ghana	ERS032054	Ghana	ERS032672	Malawi	ERS055903
Ghana	ERS022948	Ghana	ERS032067	Ghana	ERS032673	Malawi	ERS055904
Ghana	ERS022949	Ghana	ERS032170	Ghana	ERS032674	Malawi	ERS055905
Ghana	ERS022950	Ghana	ERS032171	Ghana	ERS032675	Malawi	ERS055906
Ghana	ERS022952	Ghana	ERS032172	Malawi	ERS032647	Malawi	ERS055907
Ghana	ERS022953	Ghana	ERS032173	Malawi	ERS032660	Malawi	ERS055908
Ghana	ERS022954	Ghana	ERS032174	Malawi	ERS032661	Malawi	ERS055909
Ghana	ERS022955	Ghana	ERS032176	Malawi	ERS032662	Malawi	ERS055911
Ghana	ERS022956	Ghana	ERS032177	Malawi	ERS032665	Malawi	ERS055913

Ghana	ERS022957	Ghana	ERS032178	Malawi	ERS032666	Malawi	ERS055914
Ghana	ERS022958	Ghana	ERS032180	Malawi	ERS032648	Malawi	ERS053902
Ghana	ERS022959	Ghana	ERS032181	Malawi	ERS032649	Malawi	ERS053903
Ghana	ERS022960	Ghana	ERS032182	Malawi	ERS032650	Malawi	ERS053904
Ghana	ERS022961	Ghana	ERS032183	Malawi	ERS032651	Malawi	ERS164642
Ghana	ERS022962	Ghana	ERS032184	Malawi	ERS032652	Malawi	ERS164677
Malawi	ERS164686	Malawi	ERS168607	Malawi	ERS188108	Malawi	ERS188141
Malawi	ERS168594	Malawi	ERS168608	Malawi	ERS188115	Malawi	ERS188148
Malawi	ERS168595	Malawi	ERS168609	Malawi	ERS188122	Thailand	ERS009703
Malawi	ERS168596	Malawi	ERS168610	Malawi	ERS188129	Thailand	ERS009705
Malawi	ERS168597	Malawi	ERS168611	Malawi	ERS188136	Thailand	ERS009706
Malawi	ERS168598	Malawi	ERS168612	Malawi	ERS188067	Thailand	ERS009707
Malawi	ERS168599	Malawi	ERS168614	Malawi	ERS188081	Thailand	ERS009709
Malawi	ERS168600	Malawi	ERS168615	Malawi	ERS188088	Thailand	ERS009710
Malawi	ERS168601	Malawi	ERS168616	Malawi	ERS188095	Thailand	ERS009713
Malawi	ERS168602	Malawi	ERS168617	Malawi	ERS188102	Thailand	ERS009714
Malawi	ERS168603	Malawi	ERS193667	Malawi	ERS188109	Thailand	ERS009715
Malawi	ERS168604	Malawi	ERS193672	Malawi	ERS188116	Thailand	ERS009716
Malawi	ERS168605	Malawi	ERS193623	Malawi	ERS188123	Thailand	ERS009717
Malawi	ERS168618	Malawi	ERS193628	Malawi	ERS188130	Thailand	ERS009718
Malawi	ERS168619	Malawi	ERS193638	Malawi	ERS188137	Thailand	ERS009722
Malawi	ERS168620	Malawi	ERS193643	Malawi	ERS188144	Thailand	ERS009956
Malawi	ERS168621	Malawi	ERS193648	Malawi	ERS188068	Thailand	ERS009957
Malawi	ERS168622	Malawi	ERS193653	Malawi	ERS188075	Thailand	ERS009958
Malawi	ERS168623	Malawi	ERS193658	Malawi	ERS188082	Thailand	ERS009959
Malawi	ERS168624	Malawi	ERS193663	Malawi	ERS188089	Thailand	ERS009960
Malawi	ERS168625	Malawi	ERS193668	Malawi	ERS188096	Thailand	ERS009961
Malawi	ERS168627	Malawi	ERS193673	Malawi	ERS188103	Thailand	ERS009962
Malawi	ERS168628	Malawi	ERS193678	Malawi	ERS188110	Thailand	ERS009963
Malawi	ERS168629	Malawi	ERS193624	Malawi	ERS188117	Thailand	ERS009964
Malawi	ERS168630	Malawi	ERS193629	Malawi	ERS188124	Thailand	ERS009968
Malawi	ERS168631	Malawi	ERS193634	Malawi	ERS188131	Thailand	ERS009969
Malawi	ERS168632	Malawi	ERS193639	Malawi	ERS188138	Thailand	ERS010141
Malawi	ERS168633	Malawi	ERS193644	Malawi	ERS188145	Thailand	ERS010190
Malawi	ERS168634	Malawi	ERS193654	Malawi	ERS188077	Thailand	ERS010308
Malawi	ERS168635	Malawi	ERS193659	Malawi	ERS188084	Thailand	ERS010314
Malawi	ERS168636	Malawi	ERS193664	Malawi	ERS188091	Thailand	ERS010349
Malawi	ERS168637	Malawi	ERS193669	Malawi	ERS188105	Thailand	ERS010353
Malawi	ERS168638	Malawi	ERS193674	Malawi	ERS188112	Thailand	ERS010478
Malawi	ERS168639	Malawi	ERS193679	Malawi	ERS188119	Thailand	ERS010514
Malawi	ERS168640	Malawi	ERS193625	Malawi	ERS188133	Thailand	ERS010522
Malawi	ERS168641	Malawi	ERS193630	Malawi	ERS188140	Thailand	ERS010524
Malawi	ERS168642	Malawi	ERS188069	Malawi	ERS188147	Thailand	ERS010525
Malawi	ERS168643	Malawi	ERS188076	Malawi	ERS188079	Thailand	ERS010526
Malawi	ERS168644	Malawi	ERS188090	Malawi	ERS188086	Thailand	ERS010528
Malawi	ERS168645	Malawi	ERS188097	Malawi	ERS188100	Thailand	ERS010530
Malawi	ERS168646	Malawi	ERS188111	Malawi	ERS188135	Thailand	ERS010531
Malawi	ERS168647	Malawi	ERS188118	Malawi	ERS188149	Thailand	ERS010532
Malawi	ERS168648	Malawi	ERS188125	Malawi	ERS188121	Thailand	ERS010600
Malawi	ERS168649	Malawi	ERS188132	Malawi	ERS188128	Thailand	ERS010601

Malawi	ERS168650	Malawi	ERS188139	Malawi	ERS188071	Thailand	ERS010602
Malawi	ERS168651	Malawi	ERS188146	Malawi	ERS188078	Thailand	ERS010605
Malawi	ERS168652	Malawi	ERS188066	Malawi	ERS188085	Thailand	ERS010606
Malawi	ERS168653	Malawi	ERS188073	Malawi	ERS188099	Thailand	ERS010622
Malawi	ERS175807	Malawi	ERS188080	Malawi	ERS188106	Thailand	ERS010626
Malawi	ERS175808	Malawi	ERS188087	Malawi	ERS188113	Thailand	ERS010634
Malawi	ERS175809	Malawi	ERS188094	Malawi	ERS188127	Thailand	ERS010648
Malawi	ERS168606	Malawi	ERS188101	Malawi	ERS188134	Thailand	ERS010649
Thailand	ERS010650	Thailand	ERS142879	Thailand	ERS174649	Vietnam	ERS143454
Thailand	ERS017464	Thailand	ERS142881	Thailand	ERS174651	Vietnam	ERS143462
Thailand	ERS017465	Thailand	ERS142883	Thailand	ERS174652	Vietnam	ERS143463
Thailand	ERS017466	Thailand	ERS142884	Thailand	ERS174653	Vietnam	ERS143464
Thailand	ERS017467	Thailand	ERS154522	Thailand	ERS174654	Vietnam	ERS143465
Thailand	ERS017468	Thailand	ERS164621	Thailand	ERS174655	Vietnam	ERS143467
Thailand	ERS017469	Thailand	ERS174521	Thailand	ERS174657	Vietnam	ERS143469
Thailand	ERS017470	Thailand	ERS174522	Thailand	ERS174658	Vietnam	ERS143471
Thailand	ERS017471	Thailand	ERS174523	Thailand	ERS174659	Vietnam	ERS143472
Thailand	ERS017472	Thailand	ERS174524	Thailand	ERS174661	Vietnam	ERS143473
Thailand	ERS023565	Thailand	ERS174525	Thailand	ERS174662	Vietnam	ERS143474
Thailand	ERS023566	Thailand	ERS174526	Thailand	ERS174663	Vietnam	ERS143475
Thailand	ERS023568	Thailand	ERS174527	Thailand	ERS174664	Vietnam	ERS143476
Thailand	ERS023569	Thailand	ERS174528	Thailand	ERS174665	Vietnam	ERS143477
Thailand	ERS023570	Thailand	ERS174529	Thailand	ERS174666	Vietnam	ERS143481
Thailand	ERS023572	Thailand	ERS174530	Thailand	ERS174667	Vietnam	ERS143484
Thailand	ERS023575	Thailand	ERS174531	Thailand	ERS174668	Vietnam	ERS143485
Thailand	ERS023576	Thailand	ERS174532	Thailand	ERS174669	Vietnam	ERS143493
Thailand	ERS142818	Thailand	ERS174533	Thailand	ERS347448	Vietnam	ERS143494
Thailand	ERS142819	Thailand	ERS174534	Thailand	ERS347472	Vietnam	ERS143495
Thailand	ERS142821	Thailand	ERS174535	Thailand	ERS347504	Vietnam	ERS143496
Thailand	ERS142822	Thailand	ERS174536	Thailand	ERS347512	Vietnam	ERS143497
Thailand	ERS142824	Thailand	ERS174537	Thailand	ERS010783	Vietnam	ERS143498
Thailand	ERS142825	Thailand	ERS174538	Thailand	ERS010784	Vietnam	ERS143499
Thailand	ERS142827	Thailand	ERS174539	Vietnam	ERS010034	Vietnam	ERS143501
Thailand	ERS142830	Thailand	ERS174540	Vietnam	ERS010035	Vietnam	ERS143502
Thailand	ERS142831	Thailand	ERS174541	Vietnam	ERS010036	Vietnam	ERS143505
Thailand	ERS142833	Thailand	ERS174631	Vietnam	ERS013083	Vietnam	ERS143506
Thailand	ERS142834	Thailand	ERS174632	Vietnam	ERS013084	Vietnam	ERS143508
Thailand	ERS142836	Thailand	ERS174633	Vietnam	ERS013085	Vietnam	ERS143509
Thailand	ERS142842	Thailand	ERS174634	Vietnam	ERS013086	Vietnam	ERS143511
Thailand	ERS142843	Thailand	ERS174635	Vietnam	ERS013087	Vietnam	ERS143514
Thailand	ERS142845	Thailand	ERS174636	Vietnam	ERS013088	Vietnam	ERS143516
Thailand	ERS142848	Thailand	ERS174637	Vietnam	ERS013089	Vietnam	ERS143518
Thailand	ERS142849	Thailand	ERS174638	Vietnam	ERS013102	Vietnam	ERS143519
Thailand	ERS142854	Thailand	ERS174639	Vietnam	ERS013103	Vietnam	ERS143520
Thailand	ERS142855	Thailand	ERS174640	Vietnam	ERS086846	Vietnam	ERS154466
Thailand	ERS142857	Thailand	ERS174641	Vietnam	ERS142875	Vietnam	ERS154483
Thailand	ERS142858	Thailand	ERS174642	Vietnam	ERS143419	Vietnam	ERS174506
Thailand	ERS142861	Thailand	ERS174643	Vietnam	ERS143421	Vietnam	ERS174542
Thailand	ERS142863	Thailand	ERS174644	Vietnam	ERS143424	Vietnam	ERS174543
Thailand	ERS142867	Thailand	ERS174645	Vietnam	ERS143428	Vietnam	ERS174544

Thailand	ERS142869	Thailand	ERS174646	Vietnam	ERS143429	Vietnam	ERS174545
Thailand	ERS142872	Thailand	ERS174647	Vietnam	ERS143433	Vietnam	ERS174546
Thailand	ERS142878	Thailand	ERS174648	Vietnam	ERS143437	Vietnam	ERS174547
Vietnam	ERS174548	Vietnam	ERS086810	Vietnam	ERS088710	Vietnam	ERS086997
Vietnam	ERS174550	Vietnam	ERS086811	Vietnam	ERS088712	Vietnam	ERS086998
Vietnam	ERS174551	Vietnam	ERS086812	Vietnam	ERS088713	Vietnam	ERS086999
Vietnam	ERS174552	Vietnam	ERS086814	Vietnam	ERS086797	Vietnam	ERS087000
Vietnam	ERS174553	Vietnam	ERS086815	Vietnam	ERS086798	Vietnam	ERS087028
Vietnam	ERS174554	Vietnam	ERS086816	Vietnam	ERS086799	Vietnam	ERS087029
Vietnam	ERS174555	Vietnam	ERS086817	Vietnam	ERS086800	Vietnam	ERS087030
Vietnam	ERS174556	Vietnam	ERS086819	Vietnam	ERS086801	Vietnam	ERS087032
Vietnam	ERS174557	Vietnam	ERS086820	Vietnam	ERS086802	Vietnam	ERS087033
Vietnam	ERS174558	Vietnam	ERS086821	Vietnam	ERS086803	Vietnam	ERS087034
Vietnam	ERS174560	Vietnam	ERS086822	Vietnam	ERS086804	Vietnam	ERS087035
Vietnam	ERS174670	Vietnam	ERS086823	Vietnam	ERS086806	Vietnam	ERS087036
Vietnam	ERS174671	Vietnam	ERS086824	Vietnam	ERS086807	Vietnam	ERS086886
Vietnam	ERS174672	Vietnam	ERS086825	Vietnam	ERS086808	Vietnam	ERS086887
Vietnam	ERS174673	Vietnam	ERS086827	Vietnam	ERS086809	Vietnam	ERS086888
Vietnam	ERS174674	Vietnam	ERS086828	Vietnam	ERS347506	Vietnam	ERS086909
Vietnam	ERS174675	Vietnam	ERS086833	Vietnam	ERS347521	Vietnam	ERS086913
Vietnam	ERS174676	Vietnam	ERS086834	Vietnam	ERS347529	Vietnam	ERS086934
Vietnam	ERS174677	Vietnam	ERS086836	Vietnam	ERS347537	Vietnam	ERS086950
Vietnam	ERS224919	Vietnam	ERS086837	Vietnam	ERS085458	Vietnam	ERS086981
Vietnam	ERS224924	Vietnam	ERS086839	Vietnam	ERS085459	Vietnam	ERS086983
Vietnam	ERS336368	Vietnam	ERS086840	Vietnam	ERS085460	Vietnam	ERS086984
Vietnam	ERS336375	Vietnam	ERS086847	Vietnam	ERS085461	Vietnam	ERS086985
Vietnam	ERS336380	Vietnam	ERS086848	Vietnam	ERS085462	Vietnam	ERS086986
Vietnam	ERS336381	Vietnam	ERS086864	Vietnam	ERS085463	Vietnam	ERS086987
Vietnam	ERS336386	Vietnam	ERS086865	Vietnam	ERS085464	Vietnam	ERS086991
Vietnam	ERS336392	Vietnam	ERS086872	Vietnam	ERS085465	Vietnam	ERS086994
Vietnam	ERS347474	Vietnam	ERS086873	Vietnam	ERS085466	Vietnam	ERS086995
Vietnam	ERS347475	Vietnam	ERS086874	Vietnam	ERS085467		
Vietnam	ERS347490	Vietnam	ERS086876	Vietnam	ERS085468		
Vietnam	ERS347499	Vietnam	ERS086884	Vietnam	ERS086796		

S5 Table. The 368 high-quality *Plasmodium vivax* isolates used in the study

Country	Identifier	Country	Identifier	Country	Identifier	Country	Identifier
Cambodia	**ERR020103**	Colombia	SRR1568159	Malaysia	ERR1138869	Mexico	SRR1568201
Cambodia	**ERR023039**	Colombia	SRR1568160	Malaysia	ERR1138870	Mexico	SRR1568218
Cambodia	**ERR023040**	Colombia	SRR1568169	Malaysia	ERR1138871	Mexico	SRR1568219
Cambodia	**ERR023041**	Colombia	SRR1568171	Malaysia	ERR1138872	Mexico	SRR1568223
Cambodia	**ERR023042**	Colombia	SRR1568207	Malaysia	ERR1138873	Mexico	SRR1568225
Cambodia	**ERR027119**	Colombia	SRR1568213	Malaysia	ERR1138875	Mexico	SRR1568231
Cambodia	**ERR039234**	Colombia	SRR1568221	Malaysia	ERR1138876	PNG	SRR1562605
Cambodia	**ERR054080**	Colombia	SRR1568227	Malaysia	ERR1138879	PNG	SRR1562669
Cambodia	**ERR054082**	Colombia	SRR1568230	Malaysia	ERR1138881	PNG	SRR1562672
Cambodia	**ERR111729**	Colombia	SRR1568235	Malaysia	ERR1138882	PNG	SRR1562960
Cambodia	**ERR123849**	Colombia	SRR1568236	Malaysia	ERR1138883	PNG	SRR1562963
Cambodia	**ERR152408**	Colombia	SRR1573226	Malaysia	ERR1138884	PNG	SRR1568105
Cambodia	**ERR152410**	Ethiopia	ERR925441	Malaysia	ERR1138885	PNG	SRR1568147
Cambodia	**ERR152413**	Ethiopia	ERR925440	Malaysia	ERR1475395	PNG	SRR1568177
Cambodia	**ERR211549**	Ethiopia	ERR925439	Malaysia	ERR1475396	PNG	SRR1568185
Cambodia	**ERR211557**	Ethiopia	ERR925438	Malaysia	ERR1475397	PNG	SRR1568189
Cambodia	**ERR211561**	Ethiopia	ERR925437	Malaysia	ERR1475398	PNG	SRR1568214
Cambodia	**ERR216477**	Ethiopia	ERR925436	Malaysia	ERR1475399	PNG	SRR1759411
Cambodia	**ERR216554**	Ethiopia	ERR925435	Malaysia	ERR1475418	PNG	SRR1759522
Cambodia	**ERR337538**	Ethiopia	ERR925434	Malaysia	ERR1475419	PNG	SRR1759523
Cambodia	**ERR386533**	Ethiopia	ERR925433	Malaysia	ERR1475420	PNG	SRR1759592
Cambodia	**ERR386534**	Ethiopia	ERR925431	Malaysia	ERR1475425	PNG	SRR1759594
Cambodia	**ERR386535**	Ethiopia	ERR925430	Malaysia	ERR1475427	PNG	ERR022864
Cambodia	**ERR386536**	Ethiopia	ERR925424	Malaysia	ERR1475429	PNG	ERR175552
Cambodia	**ERR386537**	Ethiopia	ERR925421	Malaysia	ERR1475430	PNG	ERR175555
Cambodia	**ERR386538**	Ethiopia	ERR925420	Malaysia	ERR1475434	PNG	ERR175557
Cambodia	**ERR386539**	Ethiopia	ERR925417	Malaysia	ERR1475439	PNG	ERR216469
Cambodia	**ERR386541**	Ethiopia	ERR925416	Malaysia	ERR1475441	PNG	ERR216474
Cambodia	**ERR386542**	Ethiopia	ERR925412	Malaysia	ERR1475451	PNG	ERR527450
Cambodia	**ERR386543**	Ethiopia	ERR925411	Malaysia	ERR1475456	PNG	ERR527453
Cambodia	**ERR386546**	Ethiopia	ERR925410	Malaysia	ERR1475457	PNG	ERR527467
Cambodia	**ERR388742**	Ethiopia	ERR925409	Malaysia	ERR054089	PNG	ERR527468
Colombia	SRR1562518	Ethiopia	ERR775192	Malaysia	ERR152414	Peru	SRR1562512
Colombia	SRR1562524	Ethiopia	ERR775191	Malaysia	ERR152415	Peru	SRR1562513
Colombia	SRR1562555	Ethiopia	ERR775190	Malaysia	ERR527337	Peru	SRR1562519
Colombia	SRR1562818	Ethiopia	ERR775189	Malaysia	ERR527363	Peru	SRR1562521
Colombia	SRR1562870	Malaysia	ERR1106842	Mexico	SRR1562522	Peru	SRR1562525
Colombia	SRR1562965	Malaysia	ERR1106843	Mexico	SRR1562526	Peru	SRR1562534
Colombia	SRR1562967	Malaysia	ERR1106846	Mexico	SRR1562839	Peru	SRR1562535
Colombia	SRR1562971	Malaysia	ERR1138855	Mexico	SRR1562840	Peru	SRR1562538
Colombia	SRR1562975	Malaysia	ERR1138856	Mexico	SRR1562968	Peru	SRR1562567
Colombia	SRR1564650	Malaysia	ERR1138857	Mexico	SRR1568077	Peru	SRR1562606
Colombia	SRR1564660	Malaysia	ERR1138858	Mexico	SRR1568110	Peru	SRR1562614
Colombia	SRR1564664	Malaysia	ERR1138861	Mexico	SRR1568126	Peru	SRR1562615
Colombia	SRR1564665	Malaysia	ERR1138862	Mexico	SRR1568127	Peru	SRR1562624

Colombia	SRR1564670	Malaysia	ERR1138864	Mexico	SRR1568150	Peru	SRR1562851
Colombia	SRR1568112	Malaysia	ERR1138865	Mexico	SRR1568153	Peru	SRR1562871
Colombia	SRR1568118	Malaysia	ERR1138866	Mexico	SRR1568158	Peru	SRR1562931
Colombia	SRR1568128	Malaysia	ERR1138867	Mexico	SRR1568181	Peru	SRR1562958
Colombia	SRR1568155	Malaysia	ERR1138868	Mexico	SRR1568190	Peru	SRR1562972
Peru	SRR1564630	Thailand	ERR527372	Thailand	ERR111714	Thailand	ERR337629
Peru	SRR1568107	Thailand	ERR527371	Thailand	ERR111715	Thailand	ERR527338
Peru	SRR1568113	Thailand	ERR527370	Thailand	ERR111716	Thailand	ERR527339
Peru	SRR1568117	Thailand	ERR527369	Thailand	ERR111717	Thailand	ERR527340
Peru	SRR1568122	Thailand	ERR527368	Thailand	ERR111718	Thailand	ERR527341
Peru	SRR1568123	Thailand	ERR527367	Thailand	ERR111719	Thailand	ERR527342
Peru	SRR1568149	Thailand	ERR527366	Thailand	ERR111720	Thailand	ERR527343
Peru	SRR1568157	Thailand	ERR527365	Thailand	ERR111721	Thailand	ERR527344
Peru	SRR1568162	Thailand	ERR527364	Thailand	ERR111722	Thailand	ERR527345
Peru	SRR1568163	Thailand	ERR426035	Thailand	ERR111723	Thailand	ERR527346
Peru	SRR1568165	Thailand	ERR426015	Thailand	ERR111724	Thailand	ERR527348
Peru	SRR1568166	Thailand	ERR404246	Thailand	ERR111725	Thailand	ERR527350
Peru	SRR1568168	Thailand	ERR164695	Thailand	ERR111726	Thailand	ERR527383
Peru	SRR1568172	Thailand	SRR1568229	Thailand	ERR111727	Thailand	ERR527382
Peru	SRR1568174	Thailand	SRR1568209	Thailand	ERR111728	Thailand	ERR527381
Peru	SRR1568175	Thailand	SRR1568208	Thailand	ERR1475350	Thailand	ERR527380
Peru	SRR1568178	Thailand	SRR1568205	Thailand	ERR1475351	Thailand	ERR527379
Peru	SRR1568179	Thailand	SRR1568186	Thailand	ERR1475352	Thailand	ERR527378
Peru	SRR1568182	Thailand	SRR1568180	Thailand	ERR1475353	Thailand	ERR527377
Peru	SRR1568183	Thailand	SRR1568161	Thailand	ERR1475354	Thailand	ERR527376
Peru	SRR1568184	Thailand	SRR1568154	Thailand	ERR1475355	Thailand	ERR527375
Peru	SRR1568187	Thailand	SRR1568152	Thailand	ERR337595	Thailand	ERR527374
Peru	SRR1568191	Thailand	SRR1568148	Thailand	ERR337596	Thailand	ERR527355
Peru	SRR1568195	Thailand	SRR1568109	Thailand	ERR337597	Thailand	ERR527354
Peru	SRR1568196	Thailand	SRR1568103	Thailand	ERR337599	Thailand	ERR527353
Peru	SRR1568198	Thailand	SRR1562974	Thailand	ERR337600	Thailand	ERR527352
Peru	SRR1568199	Thailand	SRR1562970	Thailand	ERR337601	Thailand	ERR527351
Peru	SRR1568202	Thailand	SRR1562962	Thailand	ERR337602	Thailand	ERR111709
Peru	SRR1568203	Thailand	SRR1562959	Thailand	ERR337603	Thailand	ERR111710
Peru	SRR1568206	Thailand	SRR1562845	Thailand	ERR337604	Thailand	ERR111711
Peru	SRR1568210	Thailand	SRR1562671	Thailand	ERR337605	Thailand	ERR111712
Peru	SRR1568211	Thailand	SRR1562616	Thailand	ERR337606	Thailand	ERR111713
Peru	SRR1568216	Thailand	SRR1562520	Thailand	ERR337607	Thailand	ERR337617
Peru	SRR1568232	Thailand	ERR773748	Thailand	ERR337608	Thailand	ERR337618
Peru	SRR1568234	Thailand	ERR773747	Thailand	ERR337609	Thailand	ERR337619
Peru	SRR1568787	Thailand	ERR773746	Thailand	ERR337610	Thailand	ERR337620
Peru	SRR1759047	Thailand	ERR773745	Thailand	ERR337611	Thailand	ERR337621
Peru	SRR1759122	Thailand	ERR713941	Thailand	ERR337612	Thailand	ERR337622
Peru	SRR1759307	Thailand	ERR527362	Thailand	ERR337613	Thailand	ERR337623
Peru	SRR1759336	Thailand	ERR527361	Thailand	ERR337614	Thailand	ERR337625
Thailand	ERR527385	Thailand	ERR527358	Thailand	ERR337615	Thailand	ERR337626
Thailand	ERR527384	Thailand	ERR527356	Thailand	ERR337616	Thailand	ERR337628

S6 Table

Plasmodium falciparum **loci identified by *DeepSweep* (DS; with >3 SNPs)**

Chr.	Location	Gene ID (*PF3D7_)*	Gene Function	DS hits	iHS hits	Rsb hits
4	765952	0417400	conserved protein *(close to pfdhfr)*	8		37
5	852924	0520800	conserved protein	4		
5	921557	0522400	conserved protein *(close to pfmdr1)*	13		8
5	951346	0522900	zinc finger protein *(close to pfmdr1)*	8		
6	496916	0611800	conserved protein	6		
6	1109895	0627700	transportin	4		3
6	1115827	0627800	acetyl-CoA synthetase	20	11	39
6	1139634	0628100	HECT-domain (ubiquitin-transferase)	18	1	30
6	1163355	0628200	EIF2AK (PK4)	5		2
6	1292572	0630900	ATP-dependent RNA helicase HAS1	4		6
7	333558	0707200	conserved protein *(close to pfcrt)*	5		12
7	370246	0708000	cytoskeleton associated protein (close to *pfcrt)*	4		1
7	375694	0708200	conserved protein *(close to pfcrt)*	9		14
7	409992	0709100	Cg1 protein *(close to pfcrt)*	11		38
7	417927	0709300	Cg2 protein *(close to pfcrt)*	7		46
7	467220	0710200	conserved protein *(close to pfcrt)*	5	2	
8	488913	0809600	peptidase family C50 *(close to pfdhps)*	9	3	29
8	542388	0810600	ATP-dependent RNA helicase DBP1 (close to *pfdhps*)	5		8
8	563088	0811200	ER membrane protein complex subunit 1 (*close to pfdhps*)	7		11
8	585494	0811600	conserved protein *(close to pfdhps)*	4		
8	598114	0811900	RNA-binding protein *(close to pfdhps)*	6		4
12	750432	1219000	formin 2	4		1
12	943344	1223400	phospholipid-transporting ATPase	6		3
12	954302	1223500	conserved protein	8		11
14	2508460	1461800	conserved protein	5		
14	2536662	1462400	conserved protein	4		39

Chr, Chromosome; iHS and Rsb counts defined as the number of SNPs in a gene that have an |iHS| or |Rsb|score with a p-value < 0.0001; *pfdhfr* (Chr. 4: 749001, *pfmdr1* (Chr. 5: 960020), *pfcrt* (Chr. 7: 404770), *pfdhps* (Chr. 8: 549408)

S7 Table

***Plasmodium vivax* loci identified by *DeepSweep* (DS; with >3 SNPs)**

Chr.	Location	Gene ID (*PVP01_*)	Gene Function	DS Hits	iHS hits	Rsb hits
1	904054	0119600	Plasmodium exported protein	4		5
2	100527	0202000	hypothetical protein	11		4
2	156981	0203000	multidrug resistance-associated protein 1	7		33
2	745122	0217200	Plasmodium exported protein	5		
3	620559	0313900	exported serine/threonine protein kinase	4		
5	945918	0523400	Plasmodium exported protein (PHIST)	4		6
5	1041740	0525700	DNA helicase MCM9 *(close to pvdhfr)*	6		3
5	1047865	0525800	histone acetyltransferase *(close to pvdhfr)*	5	2	8
5	1064836	0526300	conserved protein *(close to pvdhfr)*	9	2	4
5	1070542	0526400	conserved protein *(close to pvdhfr)*	10	1	12
5	1093253	0526800	conserved protein *(close to pvdhfr)*	19	4	12
6	1011569	0624300	PIR protein	4		
7	64704	0701100	reticulocyte binding protein 1b	10		8
7	500160	0709800	cysteine repeat modular protein 1	11		0
7	1462407	0735200	Plasmodium exported protein	5		6
9	972542	0922400	peptidase M16	4		
9	1735229	0939900	RNA-binding protein	7	2	1
9	1752568	0940100	AP2 domain transcription factor	5		
9	2150596	0948800	tryptophan-rich protein	10		5
10	490336	1011000	zinc finger protein *(close to pvmdr1)*	11		1
10	1443984	1033900	tryptophan-rich protein	7		
10	1470845	1034400	Plasmodium exported protein	9		
11	61701	1101300	Plasmodium exported protein	14		2
12	3026696	1271500	lysophospholipase (PST-A)	8		
13	814038	1317300	conserved protein	7		15
14	43063	1401100	Plasmodium exported protein	5		
14	798622	1418100	AP2 domain transcription factor AP2-G3	7		1
14	1227470	1428700	conserved protein *(close to pvdhps)*	13	1	5
14	1232652	1428800	histone-arginine methyltransferase CARM1 *(close to pvdhps)*	4	1	
14	1245928	1429000	CCR4-associated factor 1 *(close to pvdhps)*	5		39
14	1300426	1430100	ABC1 family *(close to pvdhps)*	4		
14	1312634	1430400	JmjC domain-containing protein *(close to pvdhps)*	5	1	25
14	1320290	1430500	conserved protein *(close to pvdhps)*	4		
14	1336114	1430700	peptidase family C50	21		
14	1431856	1432900	GPI ethanolamine phosphate transferase 3	11	1	1

Chr, Chromosome; iHS and Rsb Counts defined as the number of SNPs in a gene that have an iHS or Rsb score with a p-value < 0.0001. *pvdhfr* (Chr. 5: 1078299), *pvmdr1* (Chr. 10: 480936, *pvdhps* (Chr. 14: 1271030)

S8 Table

***Plasmodium falciparum* loci with the most iHS and Rsb hits**

Chrom	Location	Gene ID (*PF3D7_*)	Function	iHS	Rsb	Deep Sweep
7	452987	0710000	conserved protein		49	2
7	417927	0709300	Cg2 protein		46	7
13	2116999	1352900	Plasmodium exported protein	1	41	
14	2536662	1462400	conserved protein		39	4
6	1115827	0627800	acetyl-CoA synthetase	11	39	20
7	409992	0709100	Cg1 protein		38	11
4	765952	0417400	conserved protein		37	8
5	1107081	0526600	conserved protein		37	
6	1139634	0628100	HECT-domain (ubiquitin-transferase)	1	30	18
8	488913	0809600	peptidase family C50	3	29	9
13	756296	1318300	conserved protein		25	
7	1379445	0731800	alpha/beta hydrolase		25	
14	3125133	1475900	KELT protein	4	24	
4	989562	0421700	conserved protein	20	23	
8	608343	0812100	proteasome activator complex subunit 4		20	1
11	1294496	1133400	apical membrane antigen 1	19	8	
10	1395940	1035200	S-antigen	18	8	
14	2792063	1468100	conserved protein		17	3
10	217522	1004600	conserved Plasmodium membrane protein		17	
5	1288394	0531500	unspecified product	1	17	
13	1011360	1324300	conserved Plasmodium membrane protein		17	
13	1421390	1335100	merozoite surface protein 7	2	16	
13	1466337	1335900	thrombospondin-related anonymous protein	16	12	
7	340301	0707300	rhoptry-associated membrane antigen		16	1
5	1036670	0525000	zinc finger protein		16	
12	785829	1219600	aminophospholipid-transporting P-ATPase	15		
9	285998	0905700	autophagy-related protein 3		15	
7	929713	0721500	conserved Plasmodium membrane protein		15	
8	427868	0808500	Plasmodium RNA of unknown function RUF6	2	15	
8	549345	0810800	HPPK-DHPS		15	4
8	1313279	0830800	SURFIN 8.2	15	4	
7	375693	0708200	conserved protein		14	9
5	482369	0511400	conserved protein	2	14	
7	951490	0722300	ubiquitin carboxyl-terminal hydrolase		13	1
8	919102	0820300	conserved protein		13	
14	2527372	1462300	GTP-binding protein		13	2
7	333557	0707200	conserved protein		12	5
3	221968	0304600	circumsporozoite (CS) protein	1	12	
2	618524	0215000	acyl-CoA synthetase		12	
6	1265574	0630300	DNA polymerase epsilon catalytic subunit A	1	11	1
13	108319	1301900	Plasmodium exported protein	1	11	
3	653725	0316200	conserved protein		11	
13	479886	1311100	meiosis-specific nuclear structural protein 1		11	
4	695218	0415800	RING zinc finger protein		11	

6	852589	0620400	merozoite surface protein 10		11	
7	382171	0708400	heat shock protein 90		11	
6	590882	0614100	conserved protein		11	
8	563087	0811200	ER membrane protein complex subunit 1		11	7
12	2119088	1252100	rhoptry neck protein 3		11	
12	954302	1223500	conserved protein		11	8
5	64553	0501200	parasite-infected erythrocyte surface protein		11	
9	281899	0905600	WD repeat-containing protein 66		11	
10	1400721	1035300	glutamate-rich protein GLURP	10	1	
5	1042329	0525100	acyl-CoA synthetase		10	
8	846031	0818600	BEM46-like protein	2	10	
5	1184595	0528900	conserved protein		10	
6	665485	0615900	protein phosphatase		10	
7	432013	0709600	ribonucleases P/MRP protein subunit POP1		9	
12	175563	1203300	conserved protein		9	
6	571507	0613800	AP2 domain transcription factor		9	
7	883783	0720400	apoptosis-inducing factor		9	
14	411037	1410300	WD repeat-containing protein		9	
10	1573998	1039000	serine/threonine protein kinase, FIKK family		9	
11	138792	1103000	conserved protein		9	
13	2341158	1359000	conserved protein		9	
10	61898	1001000	Plasmodium exported protein (hyp12)	2	9	
13	2386189	1359700	conserved protein		9	
12	712688	1218200	conserved protein		9	
11	1376532	1135100	protein phosphatase PPM8		9	
6	241806	0605800	DNA repair protein RAD50		9	
13	100779	1301700	CX3CL1-binding protein 2	9	2	
7	404151	0709000	chloroquine resistance transporter		8	4
10	1420620	1035800	probable protein		8	
14	1990103	1448500	conserved protein		8	
5	921557	0522400	conserved protein		8	13
14	3121488	1475800	conserved protein	8	5	
7	1089266	0726000	28S ribosomal RNA		8	
12	660272	1216600	CelTOS	6	8	
13	780617	1318900	conserved protein		8	
10	497315	1012900	autophagy-related protein 18		8	
6	1272955	0630400	conserved protein		8	
8	542387	0810600	ATP-dependent RNA helicase DBP1		8	5
7	505906	0711500	regulator of chromosome condensation	7	8	3
7	729631	0716700	conserved protein		8	
8	555489	0811000	cullin-1		8	2
14	46645	1401200	Plasmodium exported protein		8	
9	1437390	0936300	ring-exported protein 3		8	
2	501713	0212400	conserved Plasmodium membrane protein		7	
7	932493	0721600	40S ribosomal protein S5		7	
9	1416021	0935800	cytoadherence linked asexual protein 9		7	
2	855072	0221200	Plasmodium exported protein (hyp15)		7	
14	1226553	1431200	OST-HTH associated domain protein		7	
2	109622	0202100	liver stage associated protein 2		7	

8	331678	0806100	conserved protein	7	
14	2183066	1453200	conserved protein	7	1
7	957542	0722500	pre-mRNA-splicing factor CWC15	7	
14	304487	1408200	AP2 domain transcription factor AP2-G2	7	
9	110619	0902500	serine/threonine protein kinase, FIKK family	6	
9	747006	0918100	cytochrome b5-like heme/steroid binding protein	6	

iHS and Rsb Counts defined as the number of SNPs in a gene that have an iHS or Rsb score with a p-value < 0.0001

S9 Table
***Plasmodium vivax* loci with the most iHS and Rsb hits**

Chr.	Location	Gene ID (*PVP01_*)	Function	iHS hits	Rsb hits	Deep Sweep
2	148495	0202900	18S ribosomal RNA	2	62	
10	189242	1003700	phosphoenolpyruvate/phosphate translocator	2	45	
14	1245928	1429000	CCR4-associated factor 1		39	5
2	156981	0203000	multidrug resistance-associated protein 1		33	7
14	1284723	1429800	protein phosphatase PPM7		29	
10	146870	1002700	conserved protein	2	28	3
10	153038	1002800	SURF1 domain-containing protein		26	
14	1312633	1430400	JmjC domain-containing protein	1	25	5
10	184519	1003600	conserved protein		24	
1	883223	0119200	Plasmodium exported protein (PHISTc)		21	
10	318571	1007200	conserved protein	5	18	
10	205509	1004100	conserved protein		18	
3	123931	0302600	conserved protein	13	16	
10	1376089	1032000	50S ribosomal protein L28, apicoplast	6	16	
13	814038	1317300	conserved protein		15	7
14	2257490	1451700	asparagine and aspartate rich protein 1		13	
14	1263931	1429300	cullin-1		13	2
5	210653	0504700	18S ribosomal RNA		12	
11	1483167	1134800	major facilitator superfamily-related transporter		12	
14	823643	1418900	conserved protein		12	
10	1324328	1030700	hypothetical protein		12	2
12	788337	1219200	hypothetical protein	8	12	2
7	551240	0711200	tRNA (adenine(58)-N(1))-methyltransferase non- catalytic subunit TRM6	1	12	
14	1255147	1429100	ER membrane protein complex subunit 1		12	3
5	1070542	0526400	conserved protein	1	12	10
2	105832	0202100	Plasmodium exported protein	2	12	
14	1267716	1429400	conserved protein		12	1
2	175717	0203400	eukaryotic translation initiation factor 4E	2	12	2
5	1093253	0526800	conserved protein	4	12	19
8	108099	0802000	5.8S ribosomal RNA		11	
9	328701	0905600	RNA polymerase subunit	1	11	
10	168725	1003200	conserved protein		11	
10	1387953	1032500	conserved protein	1	11	
13	2014020	1346200	ribosomal protein S27a	2	11	
4	212211	0404700	Plasmodium exported protein		10	
13	2017523	1346400	zinc finger protein	9	10	
4	584464	0414300	conserved protein		10	
5	1258182	0529800	AP2 domain transcription factor		10	
14	2696171	1462600	conserved protein	1	10	
7	1217570	0728900	merozoite surface protein 1	10	5	1
14	2207640	1450700	CG2-related protein		10	
14	1278613	1429700	ATP-dependent RNA helicase DBP1		10	

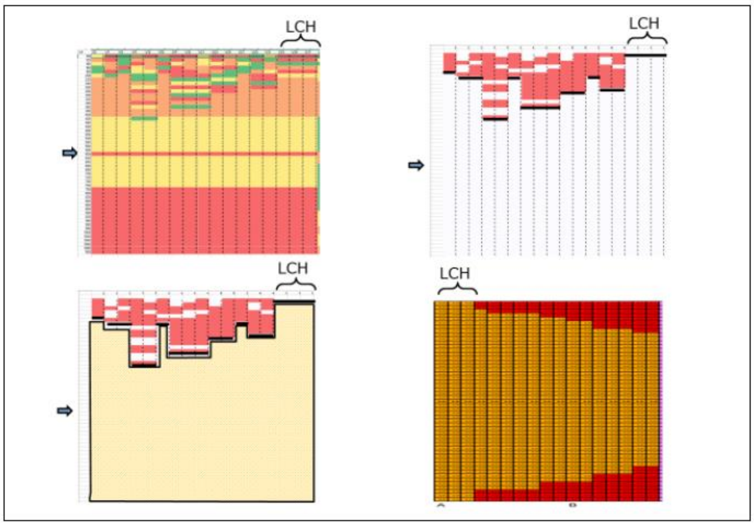
10	159076	1002900	formin 1		10	
14	2035601	1447000	WD repeat-containing protein	1	10	
12	2248858	1255000	rhoptry neck protein 2		9	2
6	456433	0610300	glutamine--fructose-6-phosphate aminotransferase [isomerizing]		9	
11	1742196	1141100	conserved protein		9	
5	1073823	0526500	mRNA-binding protein PUF2		9	2
11	1675817	1139500	cardiolipin synthetase		9	
10	521068	1011500	conserved protein		9	3
12	2987482	1270700	conserved protein	1	9	
4	680716	0416700	conserved protein		9	
4	747321	0418000	serine-repeat antigen (SERA)		9	
5	190906	0504400	sporozoite invasion-associated protein 2		9	
11	698237	1116200	NLI interacting factor-like phosphatase		9	
5	855549	0520900	conserved protein	1	9	
14	2710458	1462900	ankyrin-repeat protein		8	
11	1423420	1133300	conserved oligomeric Golgi complex subunit 4		8	
5	1077709	0526600	bifunctional dihydrofolate reductase-thymidylate synthase		8	5
6	579569	0613600	partial CSTF domain-containing protein		8	
8	1611934	0838000	Plasmodium exported protein		8	
7	64704	0701100	reticulocyte binding protein 1b		8	10
7	692112	0715200	conserved protein		8	
13	567027	1312400	conserved protein		8	
5	1368574	0532400	cysteine-rich protective antigen	8	2	
12	324002	1208000	6-cysteine protein	1	8	
12	153028	1203700	conserved protein		8	
5	1047865	0525800	histone acetyltransferase	2	8	5
2	391828	0209400	conserved protein		8	
3	496884	0311200	conserved protein		7	
1	772847	0117000	filament assembling protein		7	
14	1260196	1429200	mitochondrial carrier protein		7	1
14	1270993	1429500	hydroxymethyldihydropterin pyrophosphokinase-dihydropteroate synthase		7	4
14	2701487	1462700	kinesin		7	
9	274077	0904400	actin-like protein	2	7	
8	1504916	0835500	conserved protein		7	
11	1247714	1128700	tRNA Alanine		7	
6	668362	0616000	ookinete surface protein P28	1	7	
10	449901	1010100	divalent metal transporter		7	1
10	1327562	1030800	hypothetical protein		7	1
10	1306926	1030300	myosin B		7	
10	374884	1008500	conserved protein		7	
14	815587	1418600	conserved protein		7	
10	1407703	1033100	conserved protein	7	0	
14	817482	1418700	40S ribosomal protein S19		7	
10	1303077	1030200	60S ribosomal protein L31		7	
6	283136	0606900	conserved protein		6	
11	141431	1103200	DNA repair endonuclease XPF		6	

7	1140161	0726700	conserved protein		6	
11	678373	1115800	conserved protein		6	
12	768965	1218700	thrombospondin-related anonymous protein		6	
7	76205	0701200	reticulocyte binding protein 1a	5	6	
11	1629987	1138400	transketolase		6	
13	1291399	1330800	liver specific protein 1		6	1
7	1462407	0735200	Plasmodium exported protein		6	5
6	427598	0609500	ATP synthase-associated protein		6	
13	2015443	1346300	small nucleolar RNA snoR22		6	
5	722009	0517200	zinc finger protein		6	
13	1139543	1326200	actin-related protein	2	6	
5	945918	0523400	Plasmodium exported protein (PHIST)		6	4

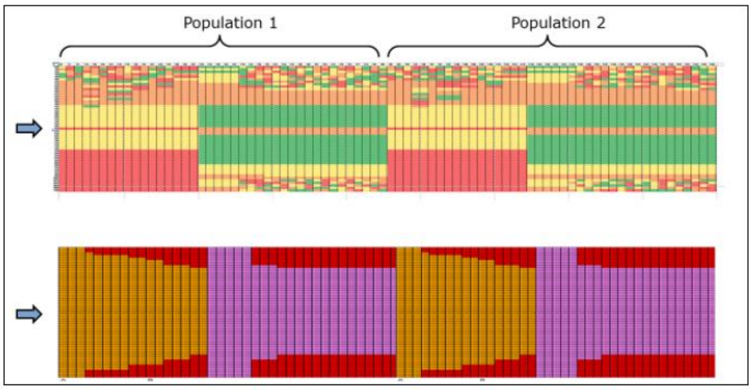
iHS and Rsb Counts defined as the number of SNPs in a gene that have an iHS or Rsb score with a p-value < 0.0001

S1 Figure
The creation of haplo-images

Panel 1)



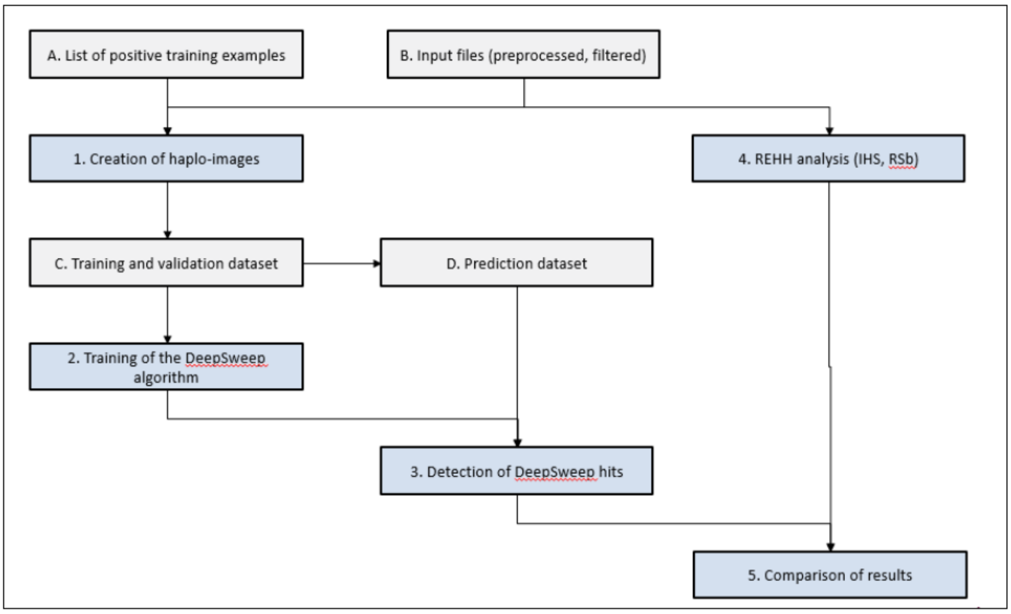
Panel 2)



Panel 1 Top-left: The image shows a small hypothetical genomic dataset of 51 SNPs (rows) and 17 samples (columns) with nucleotides re-coded as 1,2,3 and 4. All samples come from one population and have the same allele in the mid-position (row 25, highlighted with arrow). The creation of a haplo-image for this 25th SNP would start with determining the longest common haplotype (LCH). In this example, the last three columns share the LCH. **Panel 1 Top-right:** The differences between these three samples that make up the LCH and the other samples are shown in red. Panel 1 Bottom-left: The overlaps between the LCH samples and the other samples is shown in yellow. **Panel 1 Bottom-right:** A re-ordering based on shared overlap gives a haplo-image (for one population and one allele). **Panel 2 Top**: A hypothetical dataset with two alleles and two populations. **Panel 2 Bottom:** The resulting haplo-image. It should be noted that the actual process involves haplo-images each comprised of 1,401 SNPs. These SNPs were however not adjacent to one another but were chosen to be a specific distance apart. This distance was 100 basepoints for *P. falciparum* and 50 basepoints for *P. vivax*. The resulting genomic haplotype matrices have a size equivalent to the range of SNPs (e.g. 1,401 SNPs) and the number of samples (e.g. 1,100 samples). However, to improve the computational speed of the *DeepSweep* algorithm, these genomic matrices were further shrunk to a size of a height of 40 pixels by a width of 200 pixels, using "nearest" interpolation in the Scipy image package (59).

S2 Figure

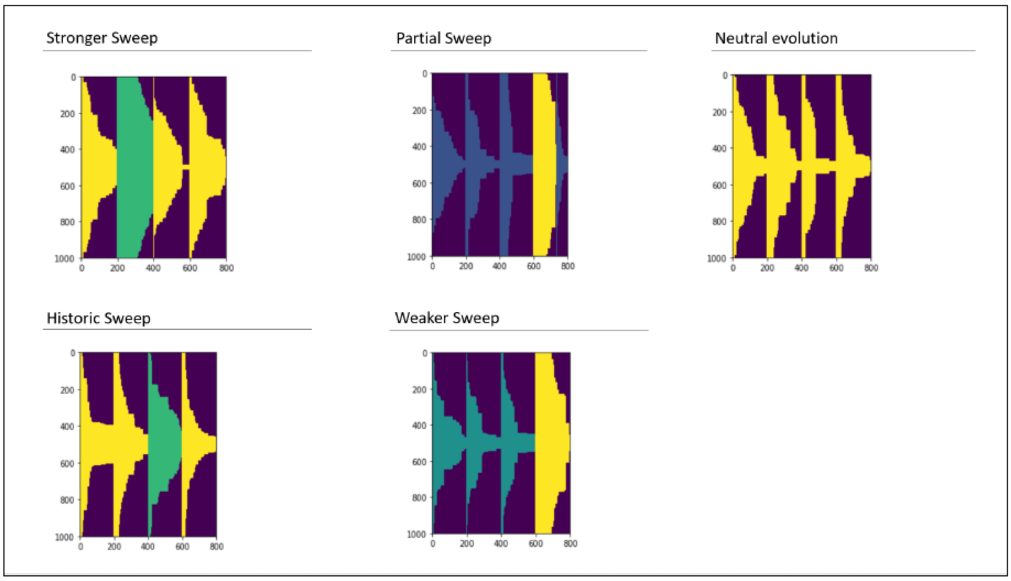
Workflow



Legend: Grey boxes are datasets, blue boxes are activity steps. Step 1 (creation of haplo-images) is further expanded upon in **S1 Figure**.

S3 Figure

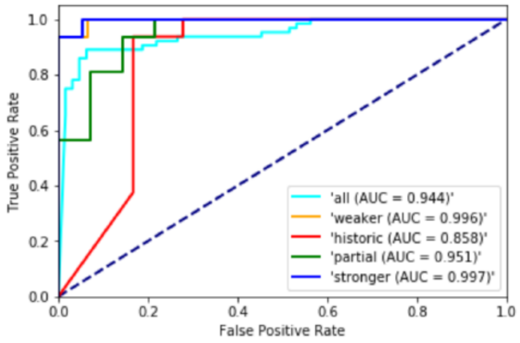
Exemplar images of simulated isolates undergoing different types of sweeps or neutral evolution



Each individual diagram is a haplo-image of a specific SNP for population of parasites, with the genomic information simulated following the settings as described in **S1 Table**. The haplo-images are created following the explanation in **S1 Figure**, with genomes of the individual parasites ordered on the horizontal axis and the overlap in haplotype for the SNP in focus shown on the vertical axis. The colour coding links to overlap in specific nucleotides (yellow, green, dark blue, light blue) with the purple background indicating no overlap. The sweeps in these illustrative examples are different nucleotides/alleles than the ancestral nucleotide/allele. Weaker sweep refers to simulation with relatively low selection coefficient; historic refers to a simulated sweep that occurred further in the past; partial refers to a partial sweep that is not fully fixed; stronger refers to a sweep with a relatively high selection coefficient.

S4 Figure

Model performance on simulated datasets

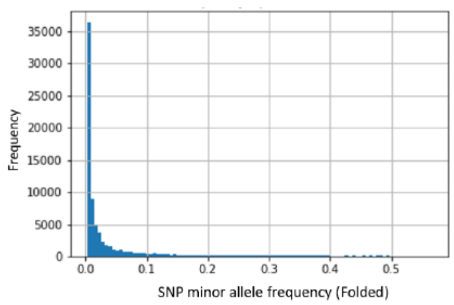


"weaker" refers to the simulation with a low selection coefficient; "historic" refers to a simulated sweep that occurred further in the past; "partial" refers to a partial sweep that is not fully fixed; "stronger" refers to a sweep with a high selection coefficient; "all" refers to all sweeps combined; AUC Area under the ROC Curve.

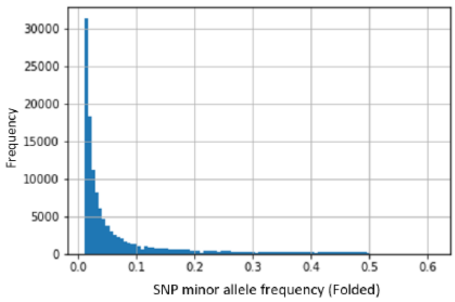
S5 Figure

Distribution of the minor allele frequencies across the SNPs

a) ***P. falciparum*** **(N=750k SNPs)**

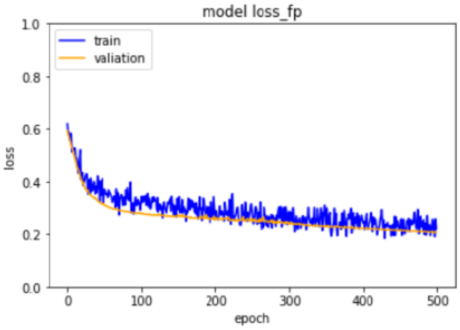
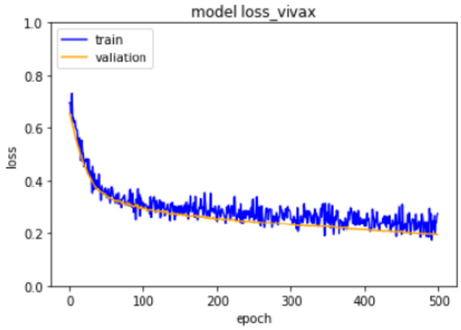
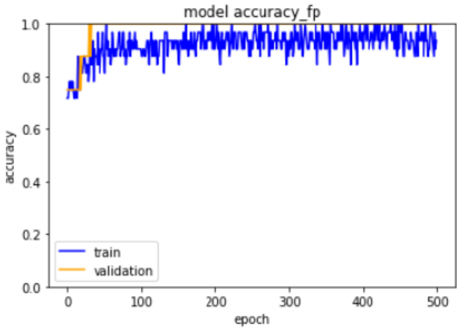
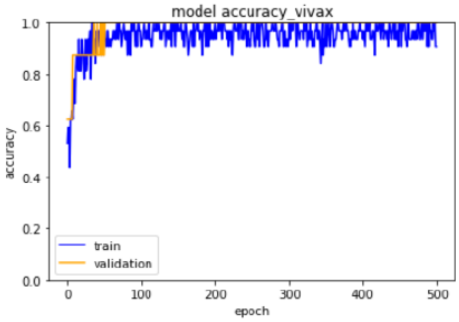


b) ***P. vivax*** **(N=588k SNPs)**



S6 Figure

Model performance for *Plasmodium falciparum* and *P. vivax* on training and validation datasets

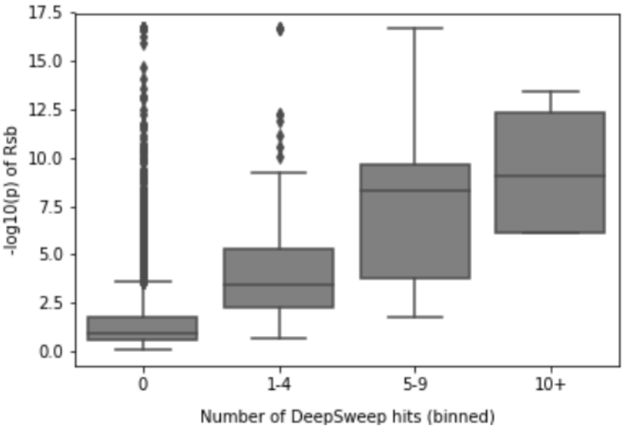
Plasmodium falciparum	*Plasmodium vivax*
	
	

The left panel shows the performance of the model in the *P. falciparum* parasite data, and the right panel shows the performance of the model in the *P. vivax* parasite data. The top panel shows model loss (measured as binomial loss) and the bottom panel shows model accuracy (measured as correct classification). The blue lines show the statistics for the training datasets and the orange lines show the statistics for the validation datasets. A negative slope in the top panel indicates a decrease in loss as the model trains over more epochs. An increasing slope in the bottom panel indicates an increase in accuracy and a reduction in misclassification as the model trains over more epochs.

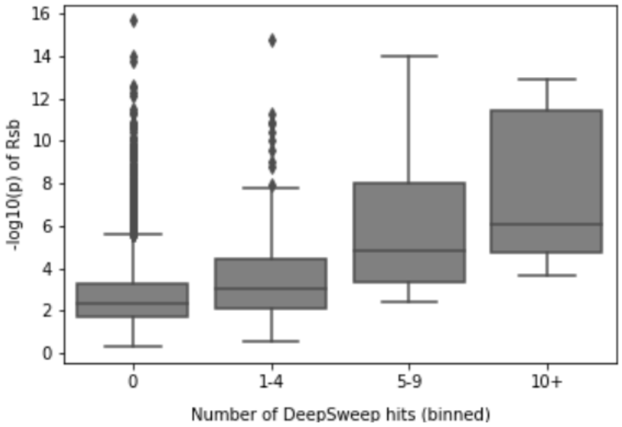
S7 Figure

Relationship between -log10 p-value of Rsb hits and number of *DeepSweep* hits

a) ***P. falciparum***



b) ***P. vivax***



Chapter 5

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1701929	**Title**	Mr.
First Name(s)	Wouter		
Surname/Family Name	Deelder		
Thesis Title	Machine learning methods for infectious diseases: applications for tuberculosis and malaria.		
Primary Supervisor	Prof. Taane Clark, Dr. Luigi Palla		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Nature Scientific Reports		
When was the work published?	December 2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	N/A
Please list the paper's authors in the intended authorship order:	N/A
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I helped to conceive and design the study. I assisted in data processing and performing the population genetic analysis. I implemented the statistical and machine learning algorithms and analysed the data. I wrote the first draft of the manuscript, and finalised it for submission after receiving revisions from co-authors.

SECTION E

Student Signature	
Date	December 9, 2022

Supervisor Signature	
Date	December 9, 2022

scientific reports

OPEN

Geographical classification of malaria parasites through applying machine learning to whole genome sequence data

Wouter Deelder[1,2], Emilia Manko[1], Jody E. Phelan[1], Susana Campino[1,4], Luigi Palla[1,3,4] & Taane G. Clark[1,4]✉

Malaria, caused by Plasmodium parasites, is a major global health challenge. Whole genome sequencing (WGS) of *Plasmodium falciparum* and *Plasmodium vivax* genomes is providing insights into parasite genetic diversity, transmission patterns, and can inform decision making for clinical and surveillance purposes. Advances in sequencing technologies are helping to generate timely and big genomic datasets, with the prospect of applying Artificial Intelligence analytical techniques (e.g., machine learning) to support programmatic malaria control and elimination. Here, we assess the potential of applying deep learning convolutional neural network approaches to predict the geographic origin of infections (continents, countries, GPS locations) using WGS data of *P. falciparum* (n = 5957; 27 countries) and *P. vivax* (n = 659; 13 countries) isolates. Using identified high-quality genome-wide single nucleotide polymorphisms (SNPs) (*P. falciparum*: 750 k, *P. vivax*: 588 k), an analysis of population structure and ancestry revealed clustering at the country-level. When predicting locations for both species, classification (compared to regression) methods had the lowest distance errors, and > 90% accuracy at a country level. Our work demonstrates the utility of machine learning approaches for geo-classification of malaria parasites. With timelier WGS data generation across more malaria-affected regions, the performance of machine learning approaches for geo-classification will improve, thereby supporting disease control activities.

Malaria, caused by *Plasmodium* parasites and transmitted by Anopheles mosquitoes, remains a pressing global health problem, with a mortality and morbidity burden heavily concentrated among children less than five years old. The morbidity and mortality impacts of *Plasmodium falciparum* malaria are predominantly concentrated in Sub-Saharan Africa, whereas the burdens of *Plasmodium vivax* are most heavily felt in Asia and South America[1]. The complex co-evolutionary history between *Plasmodium* parasites, humans, and Anopheles mosquitoes is contained within the genome of each organism, and genomic tools and data are of key importance for understanding the fundamental genetic underpinning of malaria, its geo-spatial distribution and control strategies to eliminate it. There is a rapidly growing number of *P. falciparum* and *P. vivax* isolate DNA that have undergone whole genome sequencing (WGS), with continued advances in genomic technologies likely to accelerate the timely generation of datasets from clinical and surveillance blood samples to inform disease epidemiology and control.

The rich information contained in WGS data can be used to infer transmission patterns, detect drug resistance, and support wider malaria control initiatives and elimination strategies[2,3]. WGS data in combination with population genomic methods can detect selective sweeps associated with drug resistance and infer the geographic origin of infections, including if infections are found to be imported or drug resistant and whether treatment should be adapted accordingly. It is known that malaria parasites have a population structure primarily based on geography[4,5]. Several informative molecular barcodes for speciation and geography have been developed[2,3], but typically these barcodes have not used the whole genome due to the high-dimensionality of the data and the associated computational cost[3]. However, machine learning (a subfield of Artificial Intelligence) with its ability to incorporate and analyse very large and high-dimensional datasets in an efficient manner, seems potentially well

[1]London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Dalberg Advisors, 7 Rue de Chantepoulet, 1201 Geneva, Switzerland. [3]Department of Public Health and Infectious Diseases, University of Rome La Sapienza, Rome, Italy. [4]These authors contributed equally: Susana Campino, Luigi Palla and Taane G. Clark. ✉email: Taane.Clark@lshtm.ac.uk

suited for geo-predicting using WGS data. Machine learning can be applied for classification, which concerns predicting a label (e.g., country, continental region), and regression, which involves predicting a quantity (e.g., longitude or latitude).

Machine learning has been applied effectively across a variety of problems in malaria research, including the detection of evolutionary selection associated with drug resistance[6,7], the classification and detection of parasites in red blood cells[8–11], and antimalarial drug discovery[12]. Deep learning is a subset of machine learning where algorithms aim to extract and learn series of hierarchical representations, often leveraging large amounts of data. The application of deep learning, and especially neural networks, has been explored within population genetics[13,14], including for other pathogens[15,16]. Pioneering work has also shown that machine learning, including deep learning convolutional neural networks (CNNs), can be used to predict geographic locations from human, mosquito and *P. falciparum* genetic variation[17], building on methods and the use of large genotyping chips or WGS for population structure assessment[18,19]. Here, we aim to further expand on the application of geo-prediction for malaria parasites by using a very large dataset of isolates sourced globally, (*P. falciparum*, n = 5957, 27 countries; *P. vivax,* n = 659, 13 countries) across 11 regions (South East Asia (SEA), Southern SEA (SSEA), South Asia, South America, West Africa, Central Africa, South Central Africa, East Africa, Horn of Africa, Southern Africa, Oceania). We explore the potential of both regular machine learning approaches that aim to learn representations from sequence and geographical data, as well as deep learning approaches that aim to learn and extract layers of hierarchical representations of SNP combinations linked to geography. We compare four commonly applied approaches, including classification methods that predict locations and subsequently interpolate to specific coordinates, as well as compare the performance across geographies (countries) both including the observations within those and excluding them from the training sets used to develop the models.

Materials and methods

Processing of raw sequencing data. Publicly available raw Illumina (> 150 bp paired end) sequence data from previously published studies of *P. falciparum* and *P. vivax* was downloaded from the ENA repository (see S1 Table and S2 Table for accession numbers), and accompanied by meta-data including locations of sampling (see S1 Table and S2 Table for latitude and longitude coordinates). The data included public raw sequence and GPS data from MalariaGEN projects (www.malariagen.net). Raw WGS data for *P. falciparum* (n = 5957) and *P. vivax* (n = 659) were aligned with the *Pf3D7* (v3) and *PvP01* (v1) reference genomes, respectively, using bwa-mem software (v0.7.12) using default parameter settings (e.g., concerning mismatch and sequence read clipping penalties; see http://bio-bwa.sourceforge.net/bwa.shtml). The samtools (v1.9) functions fixmate and markdup were applied to the resulting BAM files to call a set of potential variants[20]. For variant quality control, calibration assessments were performed using the GATK's BaseRecalibrator and ApplyBQSR functions, benchmarking off known high quality variants from genetic crosses for *P. falciparum*[5,21] and previously curated datasets for *P. vivax*[20]. A revised set of SNPs and insertions/deletions (indels) was called with GATK's HaplotypeCaller (version 4.1.4.1) using the option -ERC GVCF[5,22]. Variants were then assigned a quality score using GATK's Variant Quality Score Recalibration (VQSR), and those with a VQSLOD score < 0, representing variants more likely to be false than true, were filtered out[7,22]. Additionally, SNPs were removed if they had more than 10% missing alleles[7,22].The resulting dataset comprised of parasite genomes of *P. falciparum* (5,957 isolates, 750 k SNPs) and of *P. vivax* (659 isolates, 588 k SNPs). The population structure was assessed using a principal component analysis (PCA) of between isolate SNP differences. In parallel, ADMIXTURE analysis[23] was performed to understand the composition of ancestral groups across geography, where the optimal number of groups (K) was established using cross validation with values ranging between 1 and 20. This cross validation analysis led to 10 ancestral groups for both *P. falciparum* and *P. vivax* (K = 10).

Statistical models and performance. Using machine learning (ML) and deep learning (DL) statistical models, the goal was to use SNPs to predict geographical source at a location (GPS), country, and regional resolution. We applied two standard models for classification at a country and region level: (1) penalized multinomial logistic regression classifier (LOG-C; ML); (2) CNN (CNN-C; DL). Subsequently, we used the predictive probabilities placed on different locations to perform a weighted interpolation between these locations and make predictions at the GPS coordinate level.

In particular, the final prediction location (longitude and latitude) was determined by a weighted average of classifier predictions, where weights are the probabilities placed by the model on each location.

We also applied two regression models for GPS coordinate prediction: (iii) penalised linear regression model (LIN-R; ML); (iv) CNN (CNN-R; DL). The LOG-C and LIN-R models were tuned on the regularization strength C for the L1 penalty (LASSO) and implemented in the sklearn Python package (https://scikit-learn.org). The penalty parameters were tuned using cross-validation (see below, S3 Table). The deep learning CNN architecture was implemented using the Keras library (version 2.2.4)[24] in Python. Our CNN models had an architecture with a soft-max prediction layer and regularization through dropout[25] to prevent overfitting and support transferability. The main model had one convolutional layer with 4 filters, with respective filter size of (40, 9) followed by two drop-out and dense layers with ReLu activation (similar to[17]), and applied the Stochastic Gradient Descent algorithm for optimisation. We trained and validated the models for 1000 epochs. The parameterisation of the models is summarised (S3 Table). We created a stratified three-fold split in the dataset (80% training, 10% validation, 10% test) for all models, and used the validation dataset to cross-validate parameters (S3 Table). The LOG-C and LIN-R models were cross-validated (stratified, four-fold) on the regularization strength C for the L1 penalty. The reported scores (accuracy, mean weighted distance error) were calculated by making predictions on the hold-out test set (see S3 Table for the final parameter set). In addition, we conducted a "leave-one-geography-out", where

Region	Country	*Pf.* SNP Diversity	*Pf.* N*	*Pf.* %	*Pv.* SNP Diversity	*Pv.* N**	*Pv.* %
West Africa	Benin	0.040	76	1.3	–	–	–
	Burkina Faso	0.028	86	1.4	–	–	–
	Gambia	0.035	164	2.8	–	–	–
	Ghana	0.033	928	15.6	–	–	–
	Guinea	0.040	161	2.7	–	–	–
	Ivory Coast	0.034	70	1.2	–	–	–
	Mali	0.034	378	6.3	–	–	–
	Mauritania	0.035	77	1.3	–	–	–
	Nigeria	0.050	18	0.3	–	–	–
	Senegal	0.039	84	1.4	–	–	–
East Africa	Kenya	0.035	116	1.9	–	–	–
	Tanzania	0.035	320	5.4	–	–	–
	Uganda	0.053	15	0.3	–	–	–
Horn of Africa	Ethiopia	0.048	25	0.4	0.060	44	6.7
Central Africa	Cameroon	0.033	237	4.0	–	–	–
South Central Africa	DRC	0.032	339	5.7	–	–	–
Southern Africa	Madagascar	0.040	24	0.4	–	–	–
	Malawi	0.027	29	0.5	–	–	–
South Asia	India	–	–	–	0.062	40	6.1
	Bangladesh	0.037	83	1.4	–	–	–
South East Asia (SEA)	Cambodia	0.040	1118	18.8	0.049	70	10.6
	Laos	0.039	126	2.1	–	–	–
	Myanmar	0.039	246	4.1	0.061	27	4.1
	Thailand	0.038	928	15.6	0.056	160	24.3
	Vietnam	0.036	147	2.5	0.048	13	2.0
	China	–	–	–	0.066	12	1.8
Southern SEA (SSEA)	Malaysia	–	–	–	0.040	48	7.3
South America	Colombia	0.046	16	0.3	0.055	30	4.6
	Peru	0.037	24	0.4	0.059	88	13.4
	Brazil	–	–	–	0.061	82	12.5
	Mexico	–	–	–	0.039	20	3.0
Oceania	PNG	0.040	120	2.0	0.037	24	3.6
Total	–	–	*5955*	*100*	–	*658*	*100*

Table 1. Sample origin and SNP Diversity by geographic location. Pf *P. falciparum*, Pv *P. vivax*; PNG Papua New Guinea; DRC Democratic Republic of Congo.

each single geography in the training dataset was omitted in turn, with the model trained on the remaining geographies, to understand generalizability towards previously unseen locations[26].

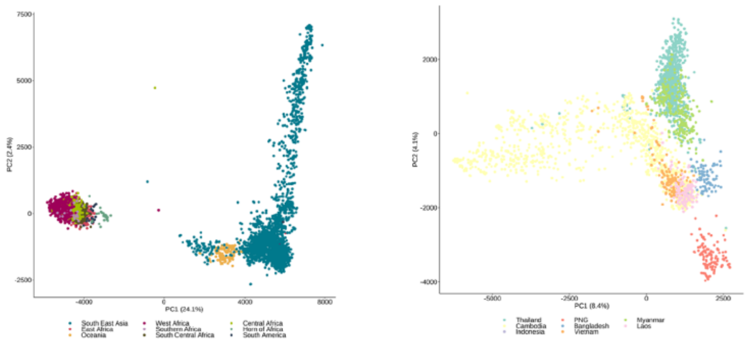
Classification accuracy was determined after assigning predicted latitude and longitude pairs to individual countries. For the classification models, a mean (weighted) distance error was calculated using the Haversine method to allow for (angular) distance calculations along a sphere, based on the difference of the actual and estimated location. The latter was determined by a weighted average of classifier predictions, where weights are the probabilities placed by the model on each location. The accuracy was calculated based on the labels of the prediction versus the test data. In particular, the baseline accuracy using a naive prediction based on the most common country would be 18.8% for *P. falciparum* (Cambodia) and 24.3% for *P. vivax* (Thailand). For the regression models, the error was calculated using the Haversine method based on the difference between the predicted and actual latitude and longitude using angular distance.

Results

Malaria isolate sequence data and population structure.

Raw WGS data with accompanying geographic origin information was available in the public domain for *P. falciparum* (n = 5957, 27 countries) and *P. vivax* (n = 659, 13 countries) (Table 1), which represent the global distributions for each parasite. Most *P. falciparum* isolates were sourced from SEA (2,648, 44.5%) followed by West Africa (2,042, 34.3%) and East Africa (451, 7.6%). Whilst, for *P. vivax,* most isolates were sourced from SEA (282, 42.9%) followed by South America (220, 33.4%) and SSEA (48) (Table 1). By analysing each species separately, high quality genome-wide SNPs were identified across the isolates (*P. falciparum* 750 k SNPs, *P. vivax* 588 k SNPs). Most SNPs have low minor allele frequencies (SNPs with MAF < 1%: *P. falciparum* 94.6%, *P. vivax* 77.6%) (S1 Figure). Most SNPs were in genic regions (*P. falciparum* 76.5%, *P. vivax* 54.3%), with a high proportion of non-synonymous (NS) amino acid

(a) ***P. falciparum*** **(n = 5,957; 750k SNPs)**



(b) ***P. vivax*** **(n = 658; 588k SNPs)**

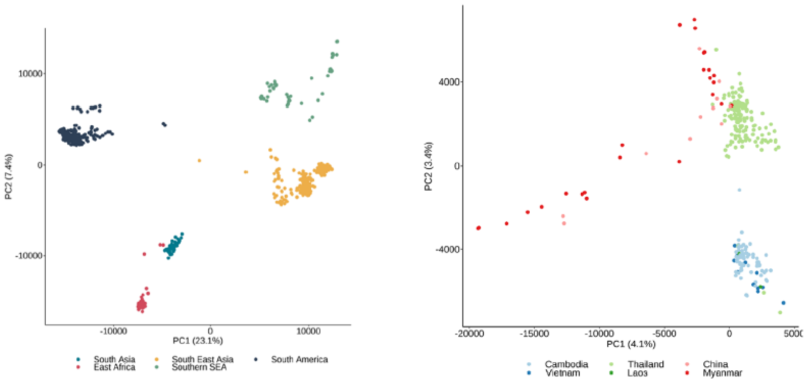


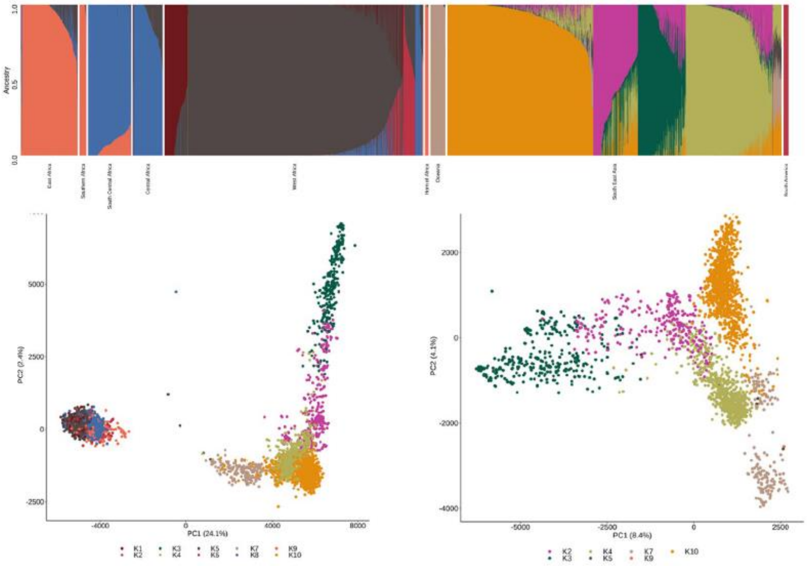
Figure 1. Population structure using principal component analysis based on all high-quality SNPs. Axes show percentage of variation explained by each principal component (PC).

changes (*P. falciparum* 63.0%, *P. vivax* 42.5%). The genetic diversity amongst *P. falciparum* isolates was relatively homogeneous across the 27 countries (SNP π: median 0.037, range 0.027–0.053), and lower in magnitude than *P. vivax*, whose data was sourced from 13 countries (SNP π: median 0.056, range 0.037–0.066) (Table 1).

Unsupervised clustering methods were applied to the genome-wide SNPs of each species to reveal the extent of their population structure and linked (pseudo-)ancestral patterns. Principal component analysis (PCA) of *P. falciparum* and *P. vivax* isolates revealed the expected separation by continent, and clear evidence of population structure at both the regional and country level (Fig. 1). An analysis of population structure and ancestry using ADMIXTURE software[23] determined the number of ancestral groups (*P. falciparum* K = 10, *P. vivax* K = 10), and their relative abundance for each isolate was estimated (Fig. 2). For *P. falciparum*, there were dominant ancestral groups across region and continent (Africa 4, SEA 4, Oceania 1, South America 1), with some evidence of mixture of ancestries (e.g., SEA isolates with 3 ancestral populations), but a general consistency within country. For *P. vivax*, the numbers of dominant ancestral groups by region differed from *P. falciparum* (South America 4, SEA 2, SSEA 2, East Africa 1, South Asia 1), due to sampling and Plasmodium species endemicity differences, such as the near absence of *P. vivax* in Africa. Overall, there was more homogeneity of ancestral groups within *P. vivax* isolates, with some groups broadly linked to neighbouring countries (comparison with Fig. 1). These analyses confirmed that spatial-genomic clustering and classification is possible using WGS data.

Application of geo-classification models. For *P. falciparum*, the predictive performance of the classification methods (LOG-C, CNN-C) was stronger than for the regression models (LIN-R, CNN-R) in regional (Table 2) and country-wide (Table 3) analyses (mean distance error (km): LIN-R 470, LOG-C 93, CNN-R 245, CNN-C 77). For locations included in the training dataset, the performance of the classification models was close to 100% at the regional level, and close to 90% at the country level (S4 Table, S5 Table). The poorest per-

(a) ***P. falciparum***



(b) ***P. vivax***

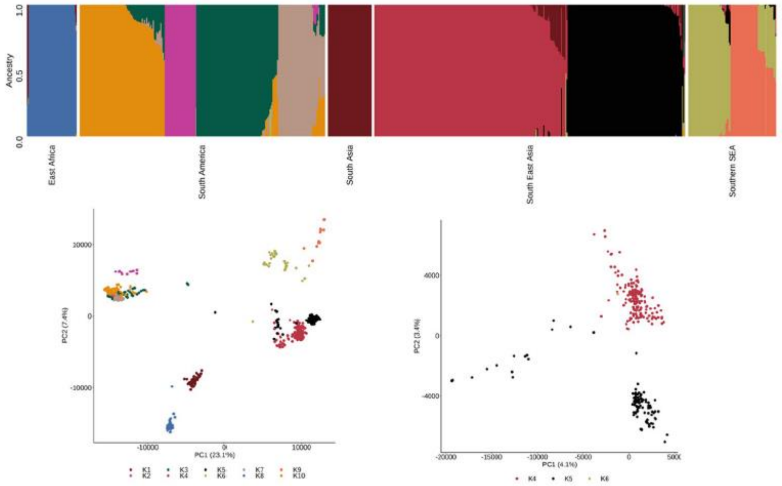


Figure 2. ADMIXTURE analysis involving 10 inferred ancestral populations (denoted as K1 to K10).

formance of the models was for African populations, for example, the mean distance error for CNN-C was high in West African (267 km) and East African countries (117 km, especially Kenya and Uganda), as well as Malawi (530 km) (Table 3), compared to other regions. This observation is consistent with the complex ancestries in African populations (Fig. 2), as well as another deep learning analysis[17]. As expected, where we predicted countries absent in data used by the training models, the distance errors (km) were at least ~ five-fold larger (LIN-R 2246, LOG-C 1848, CNN-R 1983, CNN-C 1540), with the poorest predictions for Peru (Table 4). The best performing model in this setting was the CNN-C classifier (Fig. 3).

Parasite	Region	N	LIN-R*	LOG-C*	CNN–R	CNN –C*
Pf	West Africa	2042	665 [375–1354]	302 [5–681]	368 [161–1169]	267 [45–728]
	East Africa	451	708 [693–1198]	200 [3–1581]	297 [289–856]	117 [0–1856]
	Horn of Africa	25	569 [569–569]	0 [0–0]	124 [124–124]	0 [0–0]
	Central Africa	237	635 [635–635]	29 [29–29]	184 [184–184]	0 [0–0]
	SC Africa	339	478 [478–478]	3 [3–3]	34 [34–34]	0 [0–0]
	Southern Africa	53	490 [490–968]	7 [7–433]	1543 [1018–1543]	0 [0–530]
	SEA	2648	312 [247–744]	19 [8–121]	152 [39–559]	7 [0–53]
	South America	40	1936 [1820–2053]	3 [0–7]	3683 [2535–4832]	0 [0–0]
	Oceania	120	488 [488–488]	0 [0–0]	697 [697–697]	0 [0–0]
Pv	Horn of Africa	44	334 [334–334]	0 [0–0]	142 [142- 142]	0 [0–0]
	South Asia	40	500 [500–500]	0 [0–0]	517 [517–517]	0 [0–0]
	South East Asia	282	616 [156–2751]	25 [0–1033]	578 [288–704]	0 [0–1463]
	Southern SEA	48	213 [213–213]	0 [0–0]	957 [957–957]	0 [0–0]
	South America	220	906 [134–3080]	0 [0–0]	667 [574–2773]	0 [0–0]
	Oceania	24	175 [175–175]	0 [0–0]	1103 [1103–1103]	0 [0–0]

Table 2. Mean distance Error (km) per model by region using geographies included in the training data. Pf *P. falciparum*, Pv *P. vivax,* * mean [range], CNN Convolutional Neural Network, SC South Central, SEA South East Asia; LOG-C multinomial logistic regression classifier; CNN-C CNN classifier; LIN-R penalised linear regression model; CNN-R CNN regression model.

For *P. vivax*, the predictive performance of the classification methods (LOG-C, CNN-C) was also superior compared to regression models (LIN-R, CNN-R) across regional (Table 2) and country-wide (Table 3) analyses (mean distance error (km): LIN-R 890, LOG-C 33, CNN-R 819, CNN-C 36) (Table 3). For locations included in the training dataset, the performance of the classification models was close to 100% at both the regional and country level, with the poorest performance in neighbouring China and Myanmar (S4 Table, S5 Table). The (mean) distance error for the countries not used in the development of the model is distinctively larger (km: LIN-R 1481, LOG-C 2508, CNN-R 2512, CNN-C 2405), with the poorest predictions for Ethiopia and Peru (Table 4). The best performing model in this setting was a LIN-R regression (Fig. 3).

Discussion

WGS data of *Plasmodium* parasites can detect imported infections, drug resistance, and transmission patterns, thereby assisting decision making in clinical and malaria control settings. With the implementation of WGS gaining traction across health systems, there is an opportunity to implement statistical learning methodologies to assist surveillance activities. A clear use-case includes the determination of the geographical origin of isolates, building on insights from previous work which shows that genomic data can be used to cluster parasites by geography[2–5]. Our work reveals that machine learning approaches, particularly those focusing on classification (e.g., deep learning CNNs), have the potential to accurately predict geographic locations at a GPS and country-level resolution. As expected, the performance was much stronger for isolates of which the geographic origin was already represented at the country level in the dataset, demonstrating the need for WGS to be implemented more widely to fill country gaps in genetic diversity. The weakest predictions were for *P. falciparum* in West and East Africa, where common ancestries, mixed infections, movement of people, drug resistance and malaria endemicities can complicate genetic diversity analysis. The distance errors are similar to a previous machine learning analysis of *P. falciparum* (median < 20 km), which implemented a single deep learning approach on a smaller dataset[17]. Our CNN for classification approach appeared to perform well across parasite species, was implemented with measures to minimise the effects of over-fitting, and its performance is likely to improve with greater isolate sampling and WGS data.

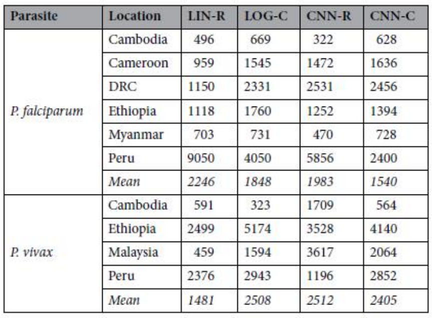
Whilst we have implemented a limited set of machine learning methods, there is scope to test alternative approaches (e.g., gradient boosted trees, support vector machines)[16] or further optimise our model parametrisations (beyond the default settings) to improve performance. For example, while L1-penalized regression approaches are generally quite competitive, stability selection on top of the LASSO leads generally to improvements[27]. Moreover, the resulting model is white box and leads to a set of interpretable SNPs. CNNs are the most utilised deep learning network type, and known to outperform alternative approaches[28]. However, one limitation of CNN models is their "black box" nature, with a complex architecture consisting of several layers, and in our context (and others[17]) making it difficult to establish which (combinations of) SNPs are informative for the geographical profiling. Other studies have used population genomic approaches to determine informative SNPs, with a focus on applying genotyping assays or amplicon sequencing for resource poor settings[2,3]. We provide computer code to implement the models, to assist future assessments in simulation or empirical studies. Future work should focus on the development of an online "geo-locator" tool that reveals a prediction of location, which can be assessed for its plausibility against the actual position, if known, and feedback into the model building and learning process. Such a framework could also be extended to integrate explicit drug resistance markers[29], as well as genomic data for malaria vectors[17], and use sequences generated on portable

Parasite	Region	Location	LIN-R	LOG-C	CNN-R	CNN-C
P. falciparum	West Africa	Benin	700	4	354	45
		Burkina Faso	374	96	161	88
		Gambia	775	132	317	107
		Ghana	401	48	193	52
		Guinea	751	515	459	402
		Ivory Coast	630	681	695	728
		Mali	563	345	208	271
		Mauritania	615	676	382	410
		Nigeria	1039	329	1169	329
		Senegal	1354	274	565	263
	East Africa	Kenya	693	200	297	117
		Tanzania	707	3	289	0
		Uganda	1198	1581	856	1856
	Horn of Africa	Ethiopia	568	0	124	0
	Central Africa	Cameroon	635	28	184	0
	SC Africa	DRC	477	2	34	0
	Southern Africa	Madagascar	490	6	1543	0
		Malawi	968	432	1018	530
	SEA	Bangladesh	743	9	159	0
		Cambodia	312	18	112	21
		Laos	276	121	152	53
		Myanmar	360	10	559	0
		Thailand	247	7	39	7
		Vietnam	356	90	199	0
	South America	Colombia	2052	0	4832	0
		Peru	1820	7	2535	0
	Oceania	PNG	488	0	697	0
	Mean		*470*	*93*	*245*	*77*
P. vivax	Horn of Africa	Ethiopia	334	0	142	0
	South Asia	India	500	0	517	0
	SEA	Cambodia	638	25	648	0
		China	2751	1033	704	1463
		Myanmar	616	311	350	311
		Thailand	604	0	288	0
		Vietnam	156	0	578	0
	SSEA	Malaysia	213	0	957	0
	South America	Brazil	3080	0	2773	6
		Colombia	1057	0	667	0
		Mexico	134	0	1502	0
		Peru	755	0	574	0
	Oceania	PNG	175	0	1103	0
	Mean		*890*	*33*	*819*	*36*

Table 3. Mean distance error (km) per model on test data using those countries included in the training data. DRC Democratic Republic of Congo; PNG Papua New Guinea; CNN Convolutional Neural Network; LOG-C multinomial logistic regression classifier; CNN-C CNN deep learnerclassifier; LIN-R penalised linear regression model; CNN-R Penalised CNN regression model; SC South Central; SEA South East Asia; SSEA Southern SEA.

and field deployable sequencing platforms (e.g., Oxford Nanopore Technology MinION). Such tools would be of immediate value to malaria control programs in endemic countries, including those that are implementing elimination activities who wish to differentiate between locally acquired or imported infections. It would also assist those countries with low malaria burden, including through the detection of imported parasites that could threaten malaria elimination targets.

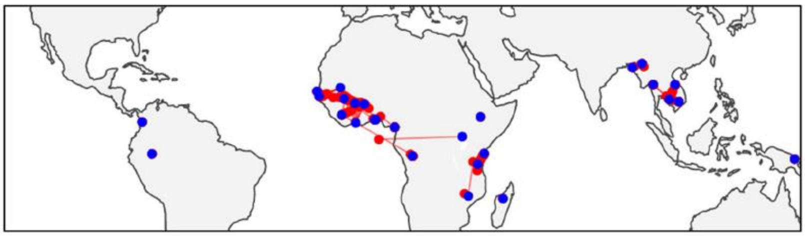
In summary, our study has demonstrated that machine learning methods can play an informative role in determining the geographic origin of WGS isolates, thereby providing important insights for both control and surveillance activities. Further, such approaches will be scalable when WGS becomes routine and cost effective, resulting in a setting with increasingly "big data" being available for decision making. The utility of this "learning"



Parasite	Location	LIN-R	LOG-C	CNN-R	CNN-C
P. falciparum	Cambodia	496	669	322	628
	Cameroon	959	1545	1472	1636
	DRC	1150	2331	2531	2456
	Ethiopia	1118	1760	1252	1394
	Myanmar	703	731	470	728
	Peru	9050	4050	5856	2400
	Mean	*2246*	*1848*	*1983*	*1540*
P. vivax	Cambodia	591	323	1709	564
	Ethiopia	2499	5174	3528	4140
	Malaysia	459	1594	3617	2064
	Peru	2376	2943	1196	2852
	Mean	*1481*	*2508*	*2512*	*2405*

Table 4. Mean distance error (km) per model on test data for unseen geographies. CNN Convolutional Neural Network; DRC Democratic Republic of Congo; LOG-C multinomial logistic regression classifier; CNN-C CNN deep learning classifier; LIN-R penalised linear regression model; CNN-R Penalised CNN regression model.

(a) ***P. falciparum*** **(CNN-C)**



(b) ***P. vivax*** **(LOG-C)**

Figure 3. Maps with predicted vs. actual locations for the best predictive models. Blue points are the actual locations in the dataset, red points are the predicted locations (where different to actual), with red lines link the actual and the predicted locations. CNN-C deep learning Convolutional Neural Network classifier. LOG-C penalised multinomial logistic regression classifier.

system will improve with time, as underlying methodologies and model performances improve with more data becoming available, and they are implemented within informatic tools to assist surveillance and clinical decision making. This utility underscores the benefit of making sequencing data and linked geographical information publicly available to global databases in a more-timely fashion to understand infection dynamics, the advantages of which have also been demonstrated by the COVID-19 crisis.

Conclusion

Advances in sequencing technologies are making real time genomics-informed surveillance and clinical management a reality. With the resulting big genomic datasets, our study has shown that machine learning methods, a subset of Artificial Intelligence, can accurately predict the geographical source of malaria parasites from sequence data. With greater geographical coverage and informatics infrastructure, such approaches will improve in performance and assist malaria control and elimination activities.

Data availability

The raw WGS data is available from the European Nucleotide Archive (ENA) (see S1 Table and S2 Table for project accession numbers). Computing code and machine learning models are available from https://github.com/WDee/GeoComparison.

Received: 31 December 2021; Accepted: 1 December 2022

References

1. World Health Organization. World Malaria Report (2020).
2. Preston, M. D. *et al.* A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. *Nat. Commun.* **5**, 1–7 (2014).
3. DiezBenavente, E. *et al.* A molecular barcode to inform the geographical origin and transmission dynamics of Plasmodium vivax malaria. *PloS Genet.* **16**, e1008576 (2020).
4. Diez Benavente, E. *et al.* Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. *Nat. Commun.* **12**, 1–11 (2021).
5. Samad, H. *et al.* Imputation-based population genetics analysis of plasmodium falciparum malaria parasites. *PLOS Genet.* **11**, e1005131 (2015).
6. Pybus, M. *et al.* Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* **31**, 493 (2015).
7. Deelder, W. *et al.* Using deep learning to identify recent positive selection in malaria parasite sequence data. *Malar. J.* **20**, 1–9 (2021).
8. Quan, Q., Wang, J. & Liu, L. An effective convolutional neural network for classifying red blood cells in malaria diseases. *Interdiscip. Sci. Comput. Life Sci.* **12**, 217–225 (2020).
9. Liang, Z. *et al.* CNN-based image analysis for malaria diagnosis. In: *Proc. - 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016* 493–496 (2017). https://doi.org/10.1109/BIBM.2016.7822567.
10. Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S. & Thoma, G. Image analysis and machine learning for detecting malaria. *Transl. Res.* **194**, 36–55 (2018).
11. Fuhad, K. M. F. *et al.* Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. *Diagnostics* **10**, 329 (2020).
12. Neves, B. J. *et al.* Deep Learning-driven research for drug discovery: Tackling malaria. *PLoS Comput. Biol.* **16**, e1007025 (2020).
13. Flagel, L., Brandvain, Y. & Schrider, D. R. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol. Biol. Evol.* **36**, 220–238 (2019).
14. Sanchez, T., Cury, J., Charpiat, G. & Jay, F. Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. *bioRxiv* https://doi.org/10.1101/2020.01.20.910539 (2020).
15. Deelder, W. *et al.* Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data. *Front. Genet.* **10**, 922 (2019).
16. Libiseller-Egger, J., Phelan, J., Campino, S., Mohareb, F. & Clark, T. G. Robust detection of point mutations involved in multidrug-resistant mycobacterium tuberculosis in the presence of co-occurrent resistance markers. *PLoS Comput. Biol.* **16**, e1008518 (2020).
17. Battey, C. J., Ralph, P. L. & Kern, A. D. Predicting geographic location from genetic variation with deep neural networks. *Elife* **9**, 1–22 (2020).
18. Guillot, G., Jónsson, H., Hinge, A., Manchih, N. & Orlando, L. Accurate continuous geographic assignment from low- to high-density SNP data. *Bioinformatics* **32**, 1106–1108 (2016).
19. Bhaskar, A., Javanmard, A., Courtade, T. A., Tse, D. & Valencia, A. Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies. *Bioinformatics* **33**, 879–885 (2017).
20. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).
21. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Res.* **26**, 1288–1299 (2016).
22. Benavente, E. D. *et al.* Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. *PLoS ONE* **12**, e0177134 (2017).
23. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
24. Chollet, F., & others. Keras. GitHub. Retrieved from https://github.com/fchollet/keras (2015).
25. Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
26. Mordelet, F. & Vert, J. P. ProDiGe: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinform.* **12**, 1–15 (2011).
27. Mahé, P. & Tournoud, M. Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection. *BMC Bioinform.* **19**, 1–11 (2018).
28. Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **8**, 1–74 (2021).
29. Turkiewicz, A. *et al.* Genetic diversity of the Plasmodium falciparum GTP-cyclohydrolase 1, dihydrofolate reductase and dihydropteroate synthetase genes reveals new insights into sulfadoxine-pyrimethamine antimalarial drug resistance. *PLoS Genet.* **16**, e1009268 (2020).

Acknowledgements

TGC was funded by Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, MR/R020973/1 and MR/X005895/1) grants. SC was funded by BloomsburySET and Medical Research Council

UK grants (MR/M01360X/1, MR/R025576/1, MR/R020973/1 and MR/X005895/1). We thank Aleksei Ponomarev for providing support on Python coding.

Author contributions

W.D., S.C., L.P., and T.G.C. conceived and designed the study. E.M. and J.E.P. performed the bioinformatic processing of the raw sequencing data. W.D. and E.M. performed the population genetic and statistical analysis, under the supervision of S.C., L.P. and T.G.C. W.D. wrote the first draft of the manuscript. All authors commented on and edited the manuscript and approved the final version. W.D. and T.G.C. compiled the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-25568-6.

Correspondence and requests for materials should be addressed to T.G.C.

Reprints and permissions information is available at www.nature.com/reprints.

S1 Table

Please note the separate file S1_Table.xlsx – available online upon publication

S2 Table

Please note the separate file S2_Table.xlsx – available online upon publication

S3 Table

The machine learning parameter settings for the models

Name	Classifier	Predicts	Fixed Parameters	Cross-validated parameters *
LOG-C**	Penalised Multinomial Logistic Regression – classification	Region, Country, and GPS	Penalty type = ”L1” Tolerance=0.001 Maximum iterations=1000	Penalty = 0.01/0.1
LIN-R**	Penalised Linear Regression - regression	GPS	Penalty type = ”L1” Tolerance=0.001 Maximum iterations=1000	Penalty = 0.003/0.003
CNN-C	Convolutional Neural Network - classification	Region, Country, and GPS	Epochs=1000 Early stopping with patience of 900	-
CNN-R	Convolutional Neural Network – regression	GPS	Epochs=1000 Early stopping with patience of 900	-

GPS Global Positioning System; LOG-C penalised multinomial logistic regression classifier; CNN-C CNN classifier; LIN-R penalised linear regression model; CNN-R CNN regression model.

* Performed on *P. falciparum* and *P. vivax* data separately, across a cross-validation range of parameter values of 0.001, 0.0031, 0.01, 0.031, 0.1, 0.31 and 1, resulting in this case in the same penalty values for *P. falciparum* and *P. vivax*.
** There are two penalty parameters due to latitude/longitude

S4 Table

Classification accuracy at a country level for *P. falciparum* (Pf) and *P. vivax* (Pv)

Region	Country	Pf LOG-C	Pf CNN-C	Pv LOG-C	Pv CNN-C
West Africa	Benin	100	100	-	-
	Burkina Faso	75.0	75.0	-	-
	Gambia	93.8	100.0	-	-
	Ghana	95.7	90.3	-	-
	Guinea	62.5	62.5	-	-
	Mali	63.2	63.2	-	-
	Mauritania	-	50.0	-	-
	Nigeria	50.0	50.0	-	-
	Senegal	62.5	50.0	-	-
East Africa	Kenya	63.6	72.7	-	-
	Tanzania	100	100	-	-
	Uganda	50.0	50.0	-	-
Horn of Africa	Ethiopia	100	100	100	100
Central Africa	Cameroon	95.7	100	-	-
South Central Africa	DRC	100	100	-	-
Southern Africa	Madagascar	100	100	-	-
	Malawi	66.7	33.3	-	-
South Asia	India	-	-	100	100
	Bangladesh	100	100	-	-
South East Asia (SEA)	Cambodia	98.2	97.3	100	100
	Laos	83.3	83.3	-	-
	Myanmar	100.0	95.8	66.7	66.7
	Thailand	98.9	98.9	100	100
	Vietnam	71.4	92.9	100	100
	China	-	-	0	100
Southern SEA	Malaysia	-	-	100	100
South America	Colombia	100	100	100	100
	Peru	100	100	100	100
	Brazil	-	-	100	100
	Mexico	-	-	100	100
Oceania	Papua New Guinea	100	100	100	100

CNN Convolutional Neural Network, DRC Democratic Republic of Congo; LOG-C multinomial logistic regression classifier; CNN-C CNN deep learning classifier; LIN-R penalised linear regression model; CNN-R Penalised CNN regression model.

S5 Table

Confusion matrices for the best predictive classification models

(a) *P. falciparum* (CNN-C, regional level)
Please note the separate file S5_Table.xlsx – available online upon publication

(b) *P. vivax* (LOG-C, regional level)
Please note the separate file S5_Table.xlsx – available online upon publication

(c) *P. falciparum* (CNN-C, country level)
Please note the separate file S5_Table.xlsx – available online upon publication

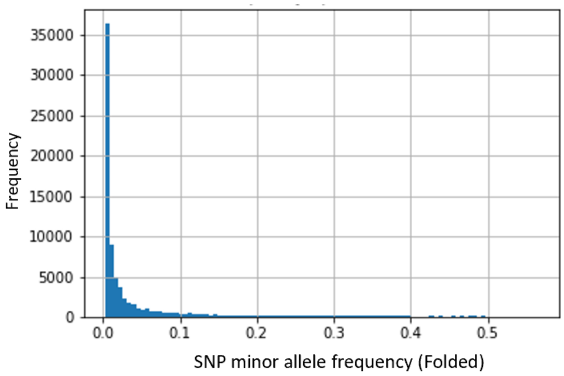
(d) *P. vivax* (LOG-C, country level)
Please note the separate file S5_Table.xlsx – available online upon publication

CNN Convolutional Neural Network, LOG-C multinomial logistic regression classifier; CNN-C CNN deep learning classifier

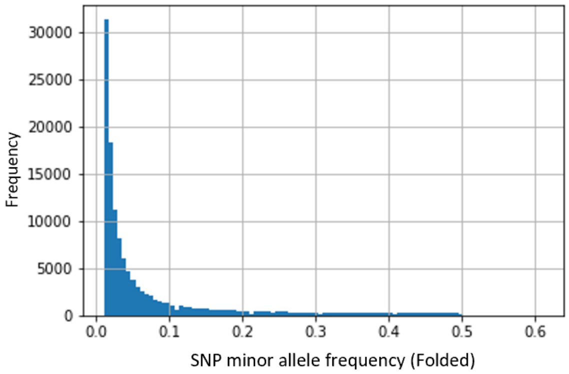
S1 Figure

Distribution of the minor allele frequencies across the SNPs

a) ***P. falciparum*** **(N=750k SNPs)**



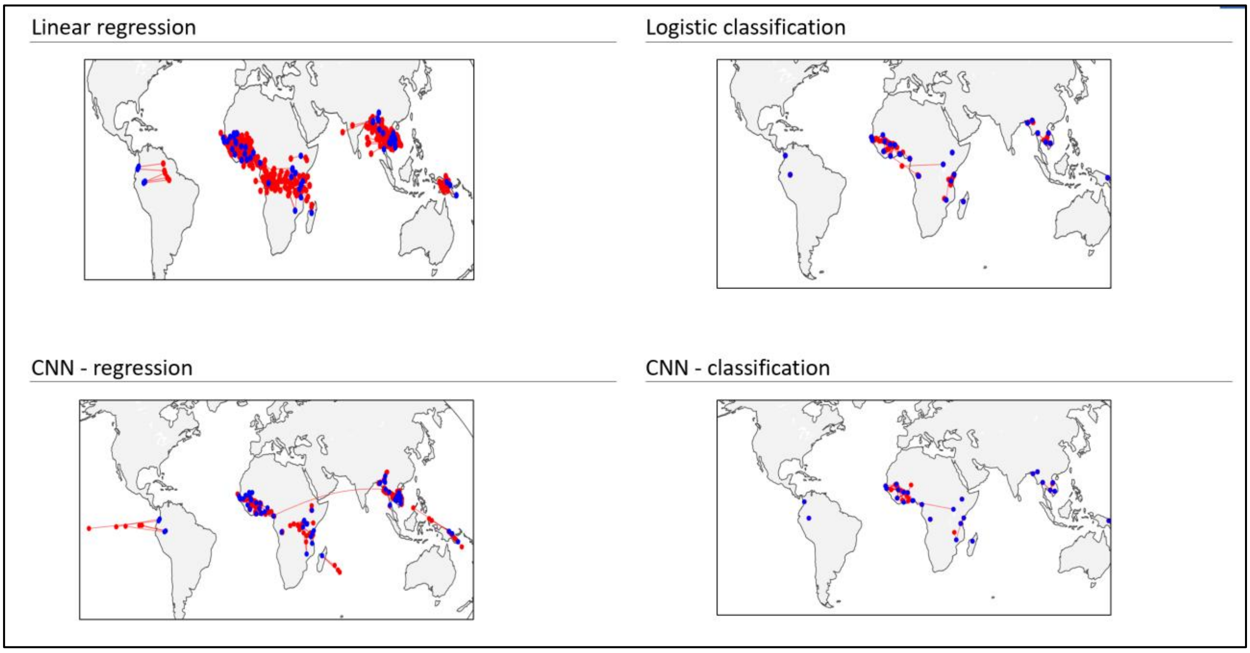
b) ***P. vivax*** **(N=588k SNPs)**



S2 Figure

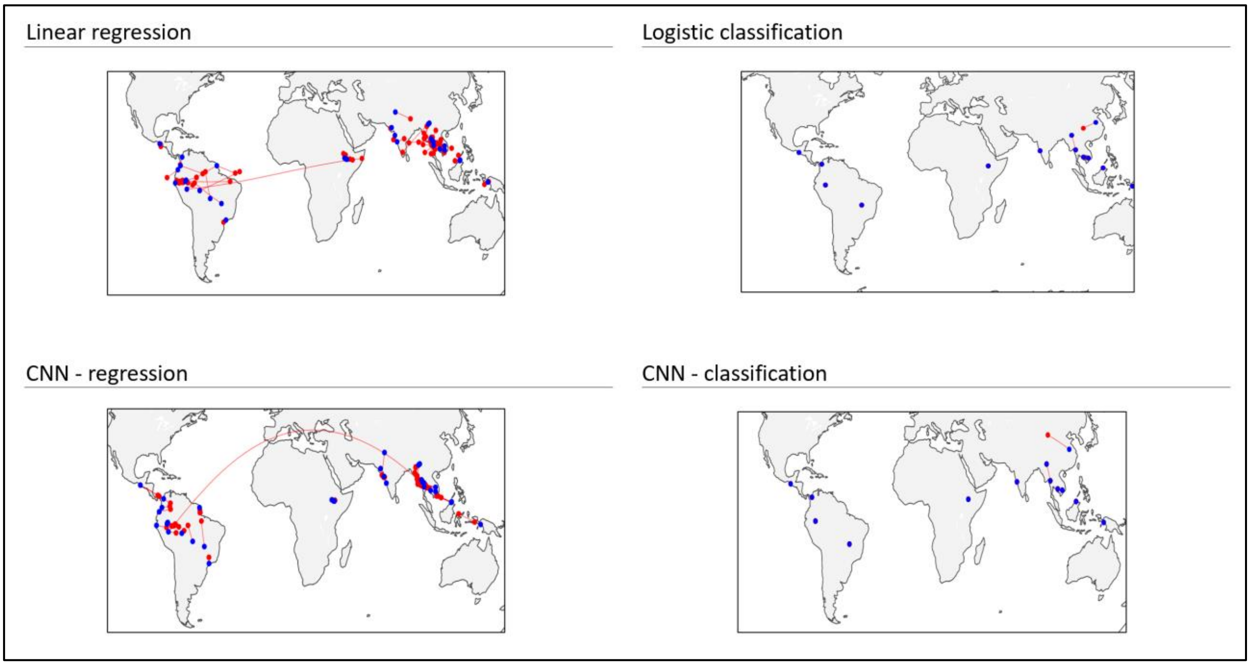
Maps with predicted vs. actual locations for all models

(a) P. falciparum



Legend: Blue points are the actual locations in the dataset, red points are the predicted locations, with red lines linking the actual and the predicted locations. Logistic classification refers to a multinomial logistic model.

(b) P. vivax



Legend: Blue points are the actual locations in the dataset, red points are the predicted locations, with red lines linking the actual and the predicted locations; Logistic classification refers to a multinomial logistic model.

Discussion

In this thesis, I explored the application of ML methods to whole-genome sequenced datasets for *P. Falciparum* and *P. Vivax* parasites and *M. tuberculosis* bacterial isolates. In Chapters 2 and 3, I applied (customized) ML methods to *M. tuberculosis* isolates to predict drug-resistance. In Chapter 4, I described the application of novel deep learning method (*Deepsweep)* to identify loci under putative selection pressure in the genomes of *P. falciparum* and *P. vivax*. Finally, in Chapter 5, I aimed to resolve the challenge of accurate geo-classification of the origin of *P. falciparum* and *P. vivax* infections by applying ML methods. There are several cross-cutting observations can be drawn from this work.

Machine learning methods can improve on traditional statistical methods for analysing WGS datasets. For example, the ML methods applied in Chapter 2 had a higher predictive accuracy for some drugs than traditional GWAS-based methods. In Chapter 3, similarly, the Treesist-TB algorithm was shown to outperform the TB-Profiler method for some drugs. In these applications, the ability of the ML methods to include interactions between features (i.e., epistatic effects) was likely a contributor to the superior performance relative to traditional methods. In all applications, the risks of overfitting were minimised. In Chapter 4, it was demonstrated that the novel application of a deep learning (DeepSweep) approach can be used for selective sweep detection and prediction, thereby identifying loci that are putatively subject to selective pressure. The DeepSweep algorithm can simultaneously analyse and identify population-genomic features encoded in different parts of the haplo-images, with great flexibility for training on other population genetic signatures and data from other organisms or simulated models. In Chapter 5, the analysis revealed a strong performance of ML and deep learning methods to determine the geographic origin of malaria isolates. Overall, our applications and ML methods appear well-suited for the analysis of high-dimensional WGS datasets.

It is critical to understand big data at a granular level. The datasets that we use, including both the genomic features and the training labels, often contain errors and idiosyncrasies that are artefacts of the way they were collected and compiled. For example, in Chapter 2 we observed and discussed that the drug-sensitivity labels contain inaccuracies due to the complexity and sensitivity of phenotypic drug sensitivity testing. Of course, these inaccuracies are one of the main reasons given to consider genotyping instead of phenotypic testing. Training an algorithm on the phenotypic labels could result in an extreme situation to a model that perfectly predicts the erroneous labels. When combined with other potential sources of bias (such as a non-representative structure of the dataset), this can significantly impair the performance of the models when applied outside the training set. For example, the known inaccuracies in the DST process for pyrazinamide (PZA) combined with sequential testing and the structure of our datasets (which contain insufficient isolates that are solely resistant to PZA and an overrepresentation of MDR and XDR cases) lead to the inclusion of non-causally linked resistance markers in ML models, which very likely would not translate into optimal performance in a real-life clinical setting. The sole way to prevent this generalization problem is careful analysis and understanding of the process through which the data and the phenotypic labels were generated, including conducting descriptive analyses to understand the structure of the datasets, and where needed assessing model performance on simulated artificial data. This process can be slow and time-consuming, and it is at perennial risk of being ignored when ML models show seemingly impressive results when applied “out of the box”.

It is important to ensure the transferability of a model from the training environment to a clinical or programmatic setting. The datasets used to train ML models in the infectious disease setting are often not an exhaustive or representative sample of the population of interest as encountered in clinical practice or disease control settings. In the balance between optimization and generalization, the models might as a result be too optimized to these specific datasets (or mis-specified) at the cost of generalizability to unseen data. An example is given in chapter 5, where I demonstrate that the country-level predictive accuracy of the geo-classification ML models drop sharply for isolates from countries held out of the training set. Another example can be found in Chapter 2, where I discuss whether cross-resistance markers should be included in predictive models. Due to the interaction between DST errors, sequential testing, and the structure of the genomic training datasets, the inclusion of such markers is likely to decrease the generalizability of the model and lead to lower actual performance in clinical settings. The inclusion of co-occurrent resistance markers might lead to overoptimism in the estimated performance that may not translate optimally into clinical practice. In Chapter 3, I showed that it is possible to improve generalizability through model specification decisions such as restricting the inclusion of highly unlikely sub-structures in the decision-tree models. Although this exclusion slightly diminishes the performance in the training environment (lower optimization), I believe that it is likely to increase the performance of the model in clinical and programmatic settings (higher generalization). As shown in Chapter 5, in some cases, the generalizability of the ML models can be tested straight-forwardly, for example, through leave-out cross-validation. In other cases, potential generalization challenges can only be progressed by careful scrutiny of the covariates or features and the manner in which these features are combined in the ML models. The difficulty of performing this task tends to be inversely proportional to the complexity of the model; this is an important consideration for the application of deep learning models.

There is a need to develop and implement methods that are able to estimate predictive uncertainty. Many ML methods generate predictions without providing estimates of the confidence in these predictions. In clinical and programmatic settings, the degree of confidence in the prediction can however be of great value. For example, for the decision-tree models in Chapters 2 and 3, a clinician or diagnostician making a diagnostic decision might decide to perform additional phenotypic testing if the genomic prediction has a relatively high degree of uncertainty. For the geoclassification models described in Chapter 5, a decision maker will likely benefit from understanding the uncertainty in the prediction at different levels of geographic granularity. For the selective sweep model, users might want to prioritise laboratory-based confirmation of putative loci to those with the highest degree of predictive confidence. Historically, the estimation of predictive uncertainty has received little attention in the ML community and it has not routinely been incorporated into software packages. However, this approach is slowly changing, including through the pioneering work of using drop-out in deep-learning models to estimate predictive uncertainty (118,123).

It is important to understand the loss in performance caused by imposing constraints that serve to increase the interpretability and simplicity of ML models in the aim to stimulate adoption. Clinicians and diagnosticians may be more able and willing to adopt new tools if the underlying predictive models are interpretable and easily understandable. Thus, it is relevant to understand what the loss in performance might be if an ML algorithm is constrained to develop a model that meets predetermined objectives of simplicity (e.g., the final decision tree follows a simple set of rules and can fit on a single page). In some cases, the trade-off between slightly lower performance but higher likelihood to adopt might be worthwhile.

It is important to remain prudent and cautious about making inferential statements from predictive ML models. For several applications discussed in this thesis, I aimed to make predictions, for example, concerning whether an isolate is drug-resistant. Often, it is also of interest to understand which features (e.g., SNPs) drive this prediction. However, the estimates of accuracy that accompany predictions (e.g., as determined though cross-validation) do not apply to inferential observations. For example, within a 10-fold cross-validation, even though the variance in the predictive accuracy might be small across folds, the features included in the fitted models might shift drastically between folds.

The instability of feature inclusion will apply to the final fitted model as well, and therefore should be considered when making inferential statements.

There is a need to build on previous studies and prior knowledge. ML studies often start "from scratch" and do not incorporate the findings and outcomes of other studies and prior research. This exclusion often comes to the detriment of model performance. This issue was demonstrated in Chapter 4, where the benefits of including the information encoded in the sub-study labelling were revealed. Using this information, rather than grouping all isolates together in one dataset, allowed for the partial compensation for specific DST errors that might disproportionally affect some data subsets. In Chapter 4, I pioneered an approach for using prior knowledge to inform the number and prioritization of genes included in the ML models and the sub-structures allowed in the trees.

Future directions of work

There are multiple directions for future research across a range of dimensions, namely to:

- **Develop and use larger and higher-quality datasets.** Bigger datasets will help with making more accurate predictions. There is especially a need for more (labelled) data for rarer events, such as drug-resistance for third-line TB drugs. Moreover, there is an opportunity to use higher-quality data and filter out subsets of data that have high rates of suspected error (in either genomic sequencing or labelling). It would also be valuable if more quantitative drug-sensitivity datasets would be available for machine learning purposes. This would potentially allow for both the assessment of effect sizes of individual mutations and the identification of new mutations.
- **Build on, and bring in, other sources of information.** In many cases, there is existing knowledge that is not included in the training process. Some of these data sources (e.g. gene function) de-facto can serve as priors and, in a Bayesian manner, could be used to inform the confidence in the predictions of our machine learning models.
- **Build bridges to other fields of statistical learning**. In order to incorporate prior knowledge and other sources of information, there is also a parallel need and opportunity to further pioneer and develop more Bayesian-oriented ML approaches and strengthen the connections to this important domain of statistical learning.
- **Build bridges to other population-genetic methods.** There is an opportunity to develop ML methods that can stretch across population-genetic domains. For example, it is likely possible to make more accurate predictions by integrating the outcomes of phylogenetic inferences into

machine learning predictions. However, for example, at the moment, there are no well-established methods to incorporate the information encoded in phylogenetic tree structures as covariates in machine learning models.

- **Ensure that tools can be adopted by clinicians and programme managers.** More work can be done to ensure that the applications and tools derived from ML models, and their predictions, are available and adopted by intended users. Availability can be advanced by ensuring that tools are accessible through online portals (e.g., where users can upload their isolates and obtain predictions), which also would benefit users in low-resource settings. Adoption can be advanced by ensuring that models are (where possible) interpretable, that efforts have been made to ensure generalizability, and that measures of predictive accuracy are provided to the user.

Conclusions

The negative trend in global disease outcomes for TB and malaria, driven in part by resistance against available drugs, diagnostics, and tools, create a renewed need for methods that can help guide the optimal usage of the resources and commodities at our disposal. The increasing adoption of whole genome sequencing is creating a new wealth of raw genomic “big” data. ML approaches offer great potential to analyse these datasets and make predictions to guide decision makers. However, it is still essential to customize and adapt these ML methods to the disease-specific context, and to resist the temptation to apply them “out of the box.” There is a lurking danger of over-optimistic predictions and impressive performance on training datasets that likely will fail to generalize in real-life settings. With the right caution and customization, this thesis has shown that ML methods and approaches have the potential to play a valuable role in the fight against the scourges of TB and malaria, and with adaption, other infectious diseases.

References

1. World Health Organization. Global health estimates: Leading causes of DALYs [Internet]. 2022 [cited 2022 Mar 23]. Available from: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/global-health-estimates-leading-causes-of-dalys

2. World Health Organization. World malaria report. In Geneva; 2021 [cited 2022 Mar 21]. Available from: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2021

3. World Health Organization. Global tuberculosis report 2021 [Internet]. 2021 [cited 2022 Mar 21]. Available from: https://www.who.int/publications/i/item/9789240037021

4. CDC. Malaria Factsheet [Internet]. 2020 [cited 2018 Nov 13]. Available from: https://www.cdc.gov/dpdx/malaria/index.html

5. Poinar G. Plasmodium dominicana n. sp. (Plasmodiidae: Haemospororida) from Tertiary Dominican amber. Syst Parasitol 2005 611 [Internet]. 2005 May [cited 2022 Mar 3];61(1):47–52. Available from: https://link.springer.com/article/10.1007/s11230-004-6354-6

6. Perkins SL. Malaria's Many Mates: Past, Present, and Future of the Systematics of the Order Haemosporida. https://doi.org/101645/13-3621 [Internet]. 2014 Feb 1 [cited 2022 Mar 3];100(1):11–25. Available from: https://bioone.org/journals/journal-of-parasitology/volume-100/issue-1/13-362.1/Malarias-Many-Mates--Past-Present-and-Future-of-the/10.1645/13-362.1.full

7. Phillips MA, Burrows JN, Manyando C, Van Huijsduijnen RH, Van Voorhis WC, Wells TNC. Malaria. Nat Rev Dis Prim 2017 31 [Internet]. 2017 Aug 3 [cited 2022 Mar 3];3(1):1–24. Available from: https://www.nature.com/articles/nrdp201750

8. Deelder A. Lecture to accept professorship - Het Aantonen van Parasieten. 1986.

9. Aggarwa P, De A, Dev N, Rehani V, Gadpayle A, Yadav S. Changing Trends in Malaria. J Trop Dis [Internet]. 2013 Nov 25 [cited 2018 Oct 25];01(04):1–3. Available from: http://www.esciencecentral.org/journals/changing-trends-in-malaria-2329-891X.1000124.php?aid=20910

10. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature [Internet]. 2002 Oct 3 [cited 2017 Jan 19];419(6906):498–511. Available from: http://www.nature.com/doifinder/10.1038/nature01097

11. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature [Internet]. 2008 Oct 9 [cited 2018 Oct 25];455(7214):757–63. Available from: http://www.nature.com/doifinder/10.1038/nature07327

12. Diez Benavente E, Manko E, Phelan J, Campos M, Nolder D, Fernandez D, et al. Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. Nat Commun. 2021;XX:In press.

13. Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG. Imputation-Based Population

Genetics Analysis of Plasmodium falciparum Malaria Parasites. Sirugo G, editor. PLOS Genet [Internet]. 2015 Apr 30 [cited 2019 Feb 17];11(4):e1005131. Available from: http://dx.plos.org/10.1371/journal.pgen.1005131

14. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nat Commun [Internet]. 2014 Jun 13 [cited 2017 Jan 20];5. Available from: http://www.nature.com/doifinder/10.1038/ncomms5052

15. Talisuna AO, Okello PE, Erhart A, Coosemans M, D'Alessandro U. Intensity of Malaria Transmission and the Spread of Plasmodium falciparum–Resistant Malaria: A Review of Epidemiologic Field Evidence. 2007 [cited 2018 Oct 25]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK1685/

16. Institute of Medicine (US) Committee on the Economics of Antimalarial Drugs. A Brief History of Malaria - Saving Lives, Buying Time [Internet]. Arrow, Kenneth; Panosian, Claire; Gelband H, editor. 2004 [cited 2022 Mar 11]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK215638/

17. Fairhurst RM, Dondorp AM. Artemisinin-Resistant Plasmodium falciparum Malaria. Microbiol Spectr. 2016 Jun 2;4(3).

18. Slivinski N. Are We Headed for a New Era of Malaria Drug Resistance? [Internet]. The Scientist Magazine. 2019 [cited 2022 Mar 21]. Available from: https://www.the-scientist.com/features/are-we-headed-for-a-new-era-of-malaria-drug-resistance--65496

19. Mu J, Ferdig MT, Feng X, Joy DA, Duan J, Furuya T, et al. Multiple transporters associated with malaria parasite responses to chloroquine and quinine. Mol Microbiol. 2003 Aug;49(4):977–89.

20. Babiker HA, Pringle SJ, Abdel-Muhsin A, Mackinnon M, Hunt P, Walliker D. High-level chloroquine resistance in Sudanese isolates of Plasmodium falciparum is associated with mutations in the chloroquine resistance transporter gene pfcrt and the multidrug resistance gene pfmdr1. J Infect Dis. 2001 May 15;183(10):1535–8.

21. Kublin JG, Dzinjalamala FK, Kamwendo DD, Malkin EM, Cortese JF, Martino LM, et al. Molecular markers for failure of sulfadoxine-pyrimethamine and chlorproguanil-dapsone treatment of Plasmodium falciparum malaria. J Infect Dis. 2002 Feb 1;185(3):380–8.

22. Preechapornkul P, Imwong M, Chotivanich K, Pongtavornpinyo W, Dondorp AM, Day NPJ, et al. Plasmodium falciparum pfmdrl amplification, mefloquine resistance, and parasite fitness. Antimicrob Agents Chemother [Internet]. 2009 Apr [cited 2022 Mar 14];53(4):1509–15. Available from: https://journals.asm.org/doi/abs/10.1128/AAC.00241-08

23. Ashley EA, Dhorda M, Fairhurst RM, Amaratunga C, Lim P, Suon S, et al. Spread of Artemisinin Resistance in Plasmodium falciparum Malaria . N Engl J Med. 2014 Jul 31;371(5):411–23.

24. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois AC, Khim N, et al. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature. 2014;505(7481):50–5.

25. Cowell AN, Winzeler EA. The genomic architecture of antimalarial drug resistance. Brief Funct Genomics [Internet]. 2019 Sep 1 [cited 2022 Mar 3];18(5):314. Available from:

/pmc/articles/PMC6859814/

26. Benavente ED, Ward Z, Chan W, Mohareb FR, Sutherland CJ, Roper C, et al. Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. PLoS One [Internet]. 2017 May 1 [cited 2020 Oct 22];12(5). Available from: https://pubmed.ncbi.nlm.nih.gov/28493919/

27. Bloland PB. WHO/CDS/CSR/DRS/2001.4 Drug resistance in malaria Drug resistance in malaria. 2001 [cited 2022 Apr 2]; Available from: http://www.who.int/emc

28. Nielsen R. Molecular Signatures of Natural Selection SNP: single nucleotide polymorphism. Annu Rev Genet [Internet]. 2005 [cited 2017 Jan 27];39:197–218. Available from: http://genet.annualreviews.org

29. Schaffner S, Sabeti PC. Evolutionary adaptation in the human lineage. Nat Educ 1(1)14. 2009;

30. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. Annu Rev Genet [Internet]. 2013 Nov 23 [cited 2019 Feb 17];47(1):97–120. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24274750

31. Ronen R, Udpa N, Halperin E, Bafna V. Learning Natural Selection from the Site Frequency Spectrum. Genetics [Internet]. 2013 Sep 1 [cited 2017 Jan 20];195(1):181–93. Available from: http://www.genetics.org/cgi/doi/10.1534/genetics.113.152587

32. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci U S A [Internet]. 2016 Jan 5 [cited 2018 Oct 25];113(1):152–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26699508

33. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature [Internet]. 2010 Mar 10 [cited 2018 Oct 18];464(7288):587–91. Available from: http://www.nature.com/articles/nature08832

34. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. Pritchard JK, editor. PLoS Genet [Internet]. 2014 Feb 27 [cited 2018 Oct 25];10(2):e1004148. Available from: https://dx.plos.org/10.1371/journal.pgen.1004148

35. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007;8(11):857.

36. Ocholla H, Preston MD, Mipando M, Jensen ATR, Campino S, MacInnis B, et al. Whole-genome scans provide evidence of adaptive evolution in Malawian Plasmodium falciparum isolates. J Infect Dis [Internet]. 2014 Dec 15 [cited 2019 Apr 5];210(12):1991–2000. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24948693

37. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. rehh 20 a reimplementation R Packag rehh to Detect Posit Sel from haplotype Struct. 2016;067629.

38. Pavlidis P, Živković D, Stamatakis A, Alachiotis N. SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. Mol Biol Evol [Internet]. 2013 Sep 1 [cited 2018

Oct 25];30(9):2224–34. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst112

39. Alachiotis N, Stamatakis A, Pavlidis P. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. Bioinformatics [Internet]. 2012 Sep 1 [cited 2018 Oct 25];28(17):2274–5. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts419

40. Hahn MW. Molecular population genetics. Oxford University Press (OUP); 2018.

41. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics [Internet]. 2015 Aug 26 [cited 2017 Jan 17];31(24):btv493. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv493

42. Sheehan S, Song YS, Jones N, Li J, Li H, Jakobsson M, et al. Deep Learning for Population Genetic Inference. Chen K, editor. PLOS Comput Biol [Internet]. 2016 Mar 28 [cited 2017 Jan 19];12(3):e1004845. Available from: http://dx.plos.org/10.1371/journal.pcbi.1004845

43. Goodfellow I, Bengio Y, Courville A. Deep learning [Internet]. MIT Press; 2016 [cited 2017 Jan 19]. 775 p. Available from: http://www.deeplearningbook.org

44. Niemann S, Richter E, Rüsch-Gerdes S. Differentiation among Members of the Mycobacterium tuberculosis Complex by Molecular and Biochemical Features: Evidence for Two Pyrazinamide-Susceptible Subtypes of M. bovis. J Clin Microbiol [Internet]. 2000 [cited 2022 Mar 3];38(1):152. Available from: /pmc/articles/PMC86043/

45. Dheda K, Gumbo T, Maartens G, Dooley KE, McNerney R, Murray M, et al. The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. Lancet Respir Med [Internet]. 2017 Apr 1 [cited 2018 Sep 4];5(4):291–360. Available from: https://www.sciencedirect.com/science/article/pii/S2213260017300796

46. World Health Organisation. What is multidrug-resistant tuberculosis (MDR-TB) and how do we control it? [Internet]. 2018 [cited 2018 Sep 4]. Available from: http://www.who.int/features/qa/79/en/

47. World Health Organization. Meeting report of the WHO expert consultation on drug-resistant tuberculosis treatment outcome definitions, 17-19 November 2020 [Internet]. 2020 [cited 2022 Apr 2]. Available from: https://www.who.int/publications/i/item/9789240022195

48. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nat 1998 3936685 [Internet]. 1998 Jun 11 [cited 2022 Mar 3];393(6685):537–44. Available from: https://www.nature.com/articles/31159

49. Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, et al. Tuberculosis. Nat Rev Dis Prim 2016 21 [Internet]. 2016 Oct 27 [cited 2022 Mar 3];2(1):1–23. Available from: https://www.nature.com/articles/nrdp201676

50. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical

studies. Genome Med [Internet]. 2020 Dec 1 [cited 2022 Mar 14];12(1):1–10. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00817-3

51. Lalor MK, Casali N, Walker TM, Anderson LF, Davidson JA, Ratna N, et al. The use of whole-genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. Eur Respir J [Internet]. 2018 Jun 1 [cited 2022 Apr 2];51(6):1702313. Available from: https://erj.ersjournals.com/content/51/6/1702313

52. Comin J, Chaure A, Cebollada A, Ibarz D, Viñuelas J, Vitoria MA, et al. Investigation of a rapidly spreading tuberculosis outbreak using whole-genome sequencing. Infect Genet Evol. 2020 Jul 1;81:104184.

53. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. PLOS Med [Internet]. 2016 Oct 1 [cited 2022 Apr 2];13(10):e1002137. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002137

54. World Health Organization. Tuberculosis Factsheet [Internet]. 2020 [cited 2020 Sep 6]. Available from: http://www.who.int/en/news-room/fact-sheets/detail/tuberculosis

55. Trauner A, Borrell S, Reither K, Gagneux S. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. Drugs [Internet]. 2014 Jul [cited 2018 Sep 6];74(10):1063–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24962424

56. Safi H, Lingaraju S, Amin A, Kim S, Jones M, Holmes M, et al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-Arabinose biosynthetic and utilization pathway genes. Nat Genet [Internet]. 2013 [cited 2018 Sep 10];45(10):1190–7. Available from: https://www.nature.com/ng/journal/v45/n10/abs/ng.2743.html

57. Gygli SM, Borrell S, Trauner A, Gagneux S. Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives. FEMS Microbiol Rev [Internet]. 2017 May 1 [cited 2018 Sep 22];41(3):354–73. Available from: https://academic.oup.com/femsre/article/41/3/354/3089982

58. Tuberculosis-drugs-and-actions.jpg [Internet]. Wikipedia. 2022. Available from: https://en.wikipedia.org/wiki/Tuberculosis_management#/media/File:Tuberculosis-drugs-and-actions.jpg

59. Koch A, Cox H. Preventing drug-resistant tuberculosis transmission. Lancet Infect Dis [Internet]. 2020 Feb 1 [cited 2022 Apr 6];20(2):157–8. Available from: http://www.thelancet.com/article/S1473309919306139/fulltext

60. Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic Determinants of Drug Resistance in *Mycobacterium tuberculosis* and Their Diagnostic Value. Am J Respir Crit Care Med [Internet]. 2016 Sep 1 [cited 2018 Oct 15];194(5):621–30. Available from: http://www.atsjournals.org/doi/10.1164/rccm.201510-2091OC

61. Nguyen QH, Contamin L, Nguyen TVA, Bañuls AL. Insights into the processes that drive the evolution of drug resistance in Mycobacterium tuberculosis. Evol Appl [Internet]. 2018 Oct 1 [cited 2022 Apr 2];11(9):1498–511. Available from: https://pubmed.ncbi.nlm.nih.gov/30344622/

62. Guzzetta G, Jurman G, Furlanello C. A machine learning pipeline for quantitative phenotype prediction from genotype data. BMC Bioinformatics [Internet]. 2010 Oct 26 [cited 2018 Oct 25];11(Suppl 8):S3. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-S8-S3

63. Farhat M, Shapiro B, Kieser K, … RS-N, 2013 U. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. nature.com [Internet]. [cited 2018 Sep 4]; Available from: https://www.nature.com/ng/journal/v45/n10/abs/ng.2747.html

64. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, et al. Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data. Front Genet. 2019;10(SEP).

65. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics [Internet]. 2016 Jan 1 [cited 2022 Mar 24];107(1):1. Available from: /pmc/articles/PMC4727787/

66. Mitchell TM (Tom M. Machine Learning. McGraw-Hill; 1997. 414 p.

67. Heidema A, Boer J, et al. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. bmcgenet.biomedcentral.com [Internet]. 2006 [cited 2018 Sep 4]; Available from: https://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-7-23

68. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet [Internet]. 2004 Dec 10 [cited 2018 Sep 4];5(1):32. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-5-32

69. Hastie T, Tibshirani R, Friedman J. High-Dimensional Problems: p N. In 2009 [cited 2017 Jan 19]. p. 649–98. Available from: http://link.springer.com/10.1007/978-0-387-84858-7_18

70. Murphy KP. Machine learning : a probabilistic perspective [Internet]. MIT Press; 2012 [cited 2018 Oct 25]. 1067 p. Available from: https://mitpress.mit.edu/books/machine-learning-1

71. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning [Internet]. New York, NY: Springer New York; 2009 [cited 2017 Jan 19]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-0-387-84858-7

72. Sutton R, Barto A. Reinforcement learning: An introduction. Second. MIT Press; 1998.

73. Efron B, Hastie T. Computer age statistical inference algorithms, evidence, and data science [Internet]. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. 2017 [cited 2018 Sep 11]. Available from: https://www.ams.org/journals/bull/0000-000-00/S0273-0979-2018-01611-X/

74. Weng JJ, Ahuja N, Huang TS. Learning recognition and segmentation of 3-D objects from 2-D images. In: 1993 IEEE 4th International Conference on Computer Vision. 1993. p. 121–7.

75. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929–58.

76. Nair V, Hinton GE. Rectified linear units improve Restricted Boltzmann machines. In: ICML

2010 - Proceedings, 27th International Conference on Machine Learning. 2010. p. 807–14.

77. Brahma PP, Wu D, She Y. Why Deep Learning Works: A Manifold Disentanglement Perspective. IEEE Trans neural networks Learn Syst [Internet]. 2016 Oct 1 [cited 2022 Sep 24];27(10):1997–2008. Available from: https://pubmed.ncbi.nlm.nih.gov/26672049/

78. Sanchez T, Cury J, Charpiat G, Jay F. Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. bioRxiv. 2020 May 19;2020.01.20.910539.

79. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nat 2019 5727767 [Internet]. 2019 Jul 31 [cited 2022 Apr 7];572(7767):116–9. Available from: https://www.nature.com/articles/s41586-019-1390-1

80. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature [Internet]. 2017 Feb 2 [cited 2022 Apr 6];542(7639):115–8. Available from: https://pubmed.ncbi.nlm.nih.gov/28117445/

81. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA [Internet]. 2016 Dec 13 [cited 2017 Feb 1];316(22):2402. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.17216

82. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nat 2021 5967873 [Internet]. 2021 Jul 15 [cited 2022 Apr 6];596(7873):583–9. Available from: https://www.nature.com/articles/s41586-021-03819-2

83. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Heal Informatics [Internet]. 2017 Jan [cited 2018 Oct 18];21(1):4–21. Available from: http://ieeexplore.ieee.org/document/7801947/

84. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology [Internet]. 2017 Aug 24 [cited 2018 Oct 15];284(2):574–82. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2017162326

85. Liang Z, Powell A, Ersoy I, Poostchi M, Silamut K, Palaniappan K, et al. CNN-based image analysis for malaria diagnosis. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [Internet]. IEEE; 2016 [cited 2018 Oct 17]. p. 493–6. Available from: http://ieeexplore.ieee.org/document/7822567/

86. Mas D, Ferrer B, Cojoc D, Finaurini S, Mico V, Garcia J, et al. Novel image processing approach to detect malaria. Opt Commun [Internet]. 2015 Sep 1 [cited 2018 Oct 17];350:13–8. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0030401815002655

87. Fuhad KMF, Tuba JF, Sarker MRA, Momen S, Mohammed N, Rahman T. Deep Learning Based Automatic Malaria Parasite Detection from Blood Smear and Its Smartphone Based Application. Diagnostics [Internet]. 2020 May 1 [cited 2021 Oct 27];10(5). Available from: /pmc/articles/PMC7277980/

88. Vijayalakshmi A, Rajesh Kanna B. Deep learning approach to detect malaria from microscopic

images. Multimed Tools Appl 2019 7921 [Internet]. 2019 Jan 11 [cited 2022 Apr 6];79(21):15297–317. Available from: https://link.springer.com/article/10.1007/s11042-019-7162-y

89. Rashidi HH, Dang LT, Albahra S, Ravindran R, Khan IH. Automated machine learning for endemic active tuberculosis prediction from multiplex serological data. Sci Reports 2021 111 [Internet]. 2021 Sep 9 [cited 2022 Apr 6];11(1):1–12. Available from: https://www.nature.com/articles/s41598-021-97453-7

90. Morang'a CM, Amenga–Etego L, Bah SY, Appiah V, Amuzu DSY, Amoako N, et al. Machine learning approaches classify clinical malaria outcomes based on haematological parameters. BMC Med [Internet]. 2020 Dec 1 [cited 2022 Apr 6];18(1):1–16. Available from: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01823-3

91. Periwal V, Rajappan JK, Jaleel AU, Scaria V. Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. BMC Res Notes [Internet]. 2011 Nov 18 [cited 2018 Sep 28];4(1):504. Available from: http://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-4-504

92. Arshadi AK, Salem M, Collins J, Yuan JS, Chakrabarti D. Deepmalaria: Artificial intelligence driven discovery of potent antiplasmodials. Front Pharmacol. 2020;10:1526.

93. Wan X, Yang C, Yang Q, Xue H, Tang NL, Yu W. MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. BMC Bioinformatics [Internet]. 2009 Jan 9 [cited 2018 Oct 22];10(1):13. Available from: http://www.biomedcentral.com/1471-2105/10/13

94. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, et al. Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease. Am J Hum Genet [Internet]. 2013 Jun 6 [cited 2018 Oct 23];92(6):1008–12. Available from: https://www.sciencedirect.com/science/article/pii/S0002929713002152

95. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature [Internet]. 2013 Apr 7 [cited 2018 Nov 15];496(7446):504–7. Available from: http://www.nature.com/doifinder/10.1038/nature12060

96. Han BA, Schmidt JP, Bowden SE, Drake JM. Rodent reservoirs of future zoonotic diseases. Proc Natl Acad Sci U S A [Internet]. 2015 Jun 2 [cited 2018 Nov 15];112(22):7039–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26038558

97. Breiman L, Leo. Bagging predictors. Mach Learn [Internet]. 1996 [cited 2018 Sep 29];24(2):123–40. Available from: http://link.springer.com/10.1023/A:1018054314350

98. Natekin A, Knoll A, Friedman JH. Greedy Function Approximation : A Gradient Boosting Machine 1 Function estimation 2 Numerical optimization in function space. Front Neurorobot [Internet]. 1999 [cited 2018 Oct 14];7(3):1–10. Available from: https://www.jstor.org/stable/2699986

99. J. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. Ann Stat [Internet]. 2000 [cited 2018 Nov 8];29:1189–232. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.9093

100. World Health Organization. World Malaria Report. 2020;

101. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nat Commun. 2014;5.

102. Diez Benavente E, Campos M, Phelan J, Nolder D, Dombrowski JG, Marinho CRF, et al. A molecular barcode to inform the geographical origin and transmission dynamics of Plasmodium vivax malaria. PloS Genet. 2020;(In press.).

103. Deelder W, Benavente ED, Phelan J, Manko E, Campino S, Palla L, et al. Using deep learning to identify recent positive selection in malaria parasite sequence data. Malar J. 2021 Dec 1;20(1).

104. Quan Q, Wang J, Liu L. An Effective Convolutional Neural Network for Classifying Red Blood Cells in Malaria Diseases. Interdiscip Sci Comput Life Sci. 2020 Jun;12(2):217–25.

105. Liang Z, Powell A, Ersoy I, Poostchi M, Silamut K, Palaniappan K, et al. CNN-based image analysis for malaria diagnosis. Proc - 2016 IEEE Int Conf Bioinforma Biomed BIBM 2016. 2017 Jan 17;493–6.

106. Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. Transl Res. 2018 Apr 1;194:36–55.

107. Neves BJ, Braga RC, Alves VM, Lima MNN, Cassiano GC, Muratov EN, et al. Deep Learning-driven research for drug discovery: Tackling malaria. PLoS Comput Biol. 2020;16(2).

108. Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol Biol Evol [Internet]. 2019 [cited 2020 Jun 17];36(2):220–38. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6367976/

109. Libiseller-Egger J, Phelan J, Campino S, Mohareb F, Clark TG. Robust detection of point mutations involved in multidrug-resistant Mycobacterium tuberculosis in the presence of co-occurrent resistance markers. PLoS Comput Biol. 2020 Dec;16(12 December).

110. Battey CJ, Ralph PL, Kern AD. Predicting geographic location from genetic variation with deep neural networks. Elife [Internet]. 2020 Jun 1 [cited 2021 Mar 9];9:1–22. Available from: https://pubmed.ncbi.nlm.nih.gov/32511092/

111. Guillot G, Jónsson H, Hinge A, Manchih N, Orlando L. Accurate continuous geographic assignment from low- to high-density SNP data. Bioinformatics. 2016 Apr;32(7):1106–8.

112. Bhaskar A, Javanmard A, Courtade TA, Tse D, Valencia A. Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies. Bioinformatics. 2017;33(6):879–85.

113. Li H. Improving SNP discovery by base alignment quality. Bioinformatics. 2011 Apr;27(8):1157–8.

114. Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. Genome Res [Internet]. 2016 Sep 1 [cited 2022 Sep 28];26(9):1288–99. Available from: https://pubmed.ncbi.nlm.nih.gov/27531718/

115. Benavente ED, Ward Z, Chan W, Mohareb FR, Sutherland CJ, Roper C, et al. Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. PLoS One [Internet]. 2017 May 1 [cited 2022 Mar 14];12(5). Available from: /pmc/articles/PMC5426636/

116. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009/07/31. 2009 Sep;19(9):1655–64.

117. Chollet F. Keras [Internet]. Github; 2015. Available from: https://github.com/fchollet/keras

118. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Vol. 15, Journal of Machine Learning Research. 2014.

119. Mordelet F, Vert JP. ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics [Internet]. 2011 Oct 6 [cited 2020 Jun 26];12(1):1–15. Available from: https://link.springer.com/articles/10.1186/1471-2105-12-389

120. Mahé P, Tournoud M. Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection. BMC Bioinformatics [Internet]. 2018 Oct 17 [cited 2022 Sep 21];19(1):1–11. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2403-z

121. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J big data. 2021 Dec;8(1).

122. Turkiewicz A, Manko E, Sutherland CJ, Benavente ED, 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 Dec;16(12 December).

123. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. 33rd Int Conf Mach Learn ICML 2016 [Internet]. 2015 Jun 6 [cited 2022 Jan 11];3:1651–60. Available from: https://arxiv.org/abs/1506.02142v6