Developing an RPA-based Molecular Barcoding Tool for Plasmodium Malaria.

Matthew Charles William Higgins

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

University of London
March 2023

Department of Infection Biology

Faculty of Infectious and Tropical Disease

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by The Biotechnology and Biological Sciences Research Council

Research group affiliations:
Professor Susana Campino
& Professor Taane Clark

I, Matthew Higgins, 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 the thesis.

Signed:  Date: 27-06-2022

Abstract

In 2020, *Plasmodium spp* the causative agent of malaria, was associated with 627,000 deaths and 241 million cases. Diagnostics play an essential role in infectious disease control and this thesis highlights how Recombinase Polymerase Amplification (RPA) could underpin the next-generation of low-cost in-field diagnostics, empowering malaria eradication efforts. The thesis covers the development of a bioinformatics tool, PrimedRPA, to optimise RPA-assay design, which was subsequently validated in the detection of *P. vivax*, the most widespread *Plasmodium* parasite. The work explores adapting RPA for a one-step colorimetric assay to align diagnostic costs with existing malaria RDTs, in addition to making the assay more suitable for in-field use. Simultaneously, I outline the use of RPA in the detection of key biomarkers associated with artemisinin resistance, a critical component of existing malaria front-line therapies, through the deliberate introduction of primer-template mismatches. Building on this work, I characterise the impact of 315 primer-template mismatch combinations on RPA reaction kinetics, with the goal of developing a robust framework for RPA-based SNP genotyping. To understand the detrimental impact even a single mismatch can have upon an RPA reaction, I outline a new tool, PrimedInclusivity, which enables researchers to utilise existing whole genome sequencing surveillance data to assess assay performance *in-silico,* based on binding site diversity. Finally, to address the lack of whole genome sequence data for neglected *Plasmodium* parasites, *P. ovale walkeri* and *P. ovale curtisi*, I generate such data and develop two new and improved reference genomes, as well as perform a population genomic analysis with isolates sourced from the African continent. Overall, my thesis describes new tools for the development of RPA-based diagnostics and generation of sequence data to assist the elimination of malaria.

Additional Publications

During my PhD, I have contributed to other manuscripts, outlined below, which are not included in this thesis.

Ward, Daniel, Matthew Higgins, Jody E. Phelan, Martin L. Hibberd, Susana Campino, and Taane G. Clark. 2021. "An Integrated in Silico Immuno-Genetic Analytical Platform Provides Insights into COVID-19 Serological and Vaccine Targets." *Genome Medicine* 13 (1): 4.

Ibrahim, Amy, Ernest Diez Benavente, Debbie Nolder, Stephane Proux, Matthew Higgins, Julian Muwanguzi, Paula Josefina Gomez Gonzalez, et al. 2020. "Selective Whole Genome Amplification of Plasmodium Malariae DNA from Clinical Samples Reveals Insights into Population Structure." *Scientific Reports* 10 (1): 10832.

Acknowledgements

I would first like to thank my supervisors Susana Campino and Taane Clark; your time and support has been invaluable over the last 4 years. Thank you for welcoming me into your group and helping me grow into the scientist I am today. I would also like to thank everyone behind the BBSRC LIDO DTP program without whom I would not have been able to obtain the financial support to pursue my PhD.

To all current and past members of the Clark / Campino group, I cannot express in words how thankful I am for your support and friendship. Jody Phelan, Matt Ravenhall and Ben Sobkowiak, your initial mentorship helped me discover my passion for bioinformatics, thank you for your patience and support. Dan Ward, completing our PhDs together over the last 4 years has been an honour, thank you for being there when sharing the highs and pushing through the lows. I feel very lucky to work with many amazing scientists who I can call my friends. Emma Collins, Tansy Valentine, Holly Acford-Palmer, Leen Vanheer, Ashley Osborne, Sophie Moss, Emilia Manko, Anna Turkiewicz, Aline Freville, Amy Ibrahim, Pepi Gomez Gonzalez, Julian Libiseller-Egger, Anton Spader and Gary Napier, it has been wonderful working together these past few years, you are all incredible scientists and I hope you know how much you mean to me.

Several years ago, I made a promise to my secondary school science teachers - Jenny Smith, Kat Modlova and Mrs Huite - that I would remember to mention them if I went on to pursue a PhD. They first nurtured my interest in science, supporting a late bloomer to get into college and turning the possibility to go to university into a reality. Writing this I feel honoured to fulfil that promise, without you I would not be where I am today. To my partner Claire, your persistent encouragement over the last 6 years has made the world of difference and I am incredibly lucky to have you by my side. Finally, to family and parents, Tim and Ang, your sacrifices and support enabled me to pursue my education which gave me the opportunity to work in a field which I love. For this I cannot express how grateful I am and as such this thesis is dedicated to you both.

Table of Contents

Abbreviations

RPA	Recombinase Polymerase Amplification.
LAMP	Loop Mediated Isothermal Amplification
NAAT	Nucleic Acid Amplification Technology.
RDT	Rapid Diagnostic Test
LF	Lateral Flow
SWGA	Selective Whole Genome Amplification
PCR	Polymerase Chain Reaction.
MoA	Mechanism of Action
WHO	World Health Organisation
MBT	Molecular Barcoding Tool
LoD	Limit of Detection
ACT	Artemisinin Combination Therapy

Chapter 1. Introduction

Introduction

The Global Malaria Burden

The Human Cost of Malaria

Malaria is caused by a subset of protozoan *Plasmodium (P)* parasites that infect humans. To date, over 200 *Plasmodium* species have been discovered which infect a range of hosts including birds, reptiles and rodents [1]. Six *Plasmodium* species are known to commonly infect humans, *P. falciparum*, *P. vivax*, *P. malariae*, *P. knowlesi*, *P. ovale curtisi* and *P. ovale walkeri*. Although *P. knowlesi* is a zoonotic malaria parasite, this species causes many human cases in several regions of South East Asia [2]. Other zoonotic transmission events have been reported, including but not limited to *P. cynomolgi,* which typically infects macaques, but these events are rare [3].

Malaria is an ancient disease and considered to be one of the biggest drivers of human evolution [4] with evidence showing that *Plasmodium* plagued our *homo* ancestors [5]. Early records from ancient Greece and Rome highlight a cyclic fever, which shaped their civilizations over the centuries [6]. In the modern world, malaria is considered one of the big three infectious diseases, alongside HIV/AIDS and tuberculosis, which plague humanity [7]. Recent estimates highlight that there were 241 million cases of malaria in 2020, resulting in an estimated 627,000 deaths [8] with children under 5 years old in Sub-Saharan Africa the most at risk sub-group [9]. Compared to two decades ago this represents a ~25% decrease in annual mortality, which was estimated at 839,000 deaths annually [10]. However, whilst a net decrease in malaria mortality has been observed, recent trends indicate that progress is stalling with mortality estimates ranging from 429,000 to 469,000 between 2015 to 2019, prior to the SARS-CoV-2 pandemic [11]. The recent rise to 627,000 deaths is most likely attributed to intervention disruption due to the ongoing SARS-Cov-2 pandemic adding strain on public health services. [8]

Malaria Burden Distribution

Approximately half the world's population is at risk of malaria (**Figure 1)** [8]. However, disease burden is not evenly distributed, with 95% of all cases occurring on the African continent [8]. *P. falciparum* is responsible for the majority of these infections. Of the six human-infecting *Plasmodium* species, *P. vivax* is the most widely distributed and found in Europe, Asia, South America and Africa due to its adaptation for temperate climatic conditions, with outbreaks occurring as far north as Moscow, Russia [12]. The zoonotic *P. knowlesi* malaria parasite is found primarily in Southeast Asia due to the presence of the macaque population which acts as a reservoir for the parasite. The full geographic distribution is not well understood for the remaining neglected *Plasmodium* parasites, *P. malariae* and *P. ovale spp,* with cases predominantly reported in Africa. The geographic distribution of malaria is set to potentially change with increasing global temperatures, which could drive a resurgence in previously malaria free zones, including Europe, and increase infection rates in endemic countries [13,14]. In addition, climate-change mitigation strategies (e.g. solar geoengineering) are predicted to impact malaria distribution, including an increased disease burden across Southern Asia, if implemented [15].

Figure 1. Highlights the global malaria burden in 2020. Countries with zero indigenous cases for at least 3 consecutive years are considered to have eliminated malaria [8].

Economic Impact of Malaria

On top of the direct human-cost, countries with a high malaria burden suffer a proportionate economic penalty which has deep consequences for overall country development and public health. Typically, this exacerbates existing issues as most cases of malaria, when not imported, occur in developing countries with fragile public health infrastructures. A micro-economic study in 2017 revealed that a case of malaria in the city of Mopeia, Mozambique typically cost the household between US$ 7.80 - US$ 107.64 depending on the severity / complexity of the case [16]. However, with the Mozambican annual income at the time only being US$ 415 per capita, malaria represents a significant economic penalisation [16]. In addition, macro-economic studies have found that malaria can reduce a country's annual GDP growth by 1.3% and a 10% reduction in malaria burden can result in a 0.3% increase in GDP [17].

Aetiology of Malaria

Discovery of *Plasmodium* Parasite and Transmission Vector

In 1880 the causative agent behind malaria was hypothesised to be a protozoan parasite, by Alphonse Laveran, a French Military Doctor posted in Algeria [18]. By 1890, through the work of several prominent microbiologists, three species of *Plasmodium* had been discovered which are now respectively called *P. vivax*, *P. falciparum* and *P. malariae* [19]. *P. ovale spp* and *P. knowlesi* were discovered later by John Stephens in 1918 and Biraj Mohan Das Gupta in 1932, respectively [20,21]. In 1897 Ronald Ross, proved the role of the *Anopheles* mosquito in the transmission of the avian *Plasmodium* parasite, *P. relictum* [18]. In 1899, Ross went on to implicate the role of the *Anopheles* mosquito in the transmission of human infection *Plasmodium* parasite whilst working in Sierra Leone. Of the ~465 *Anopheles* species classified to date and distributed across the globe, ~70 are believed to have the capacity to transmit malaria and 41 are considered to be dominant vector species capable of transmission at a level which warrants major concern to public health [22]

Plasmodium life cycle

The *Plasmodium* life cycle is presented (**Figure 2)** [23]. Infection occurs when the female *Anopheles* mosquito bites to obtain a blood meal, injecting sporozoites into the bloodstream of the recipient. The sporozoites then circulate in the bloodstream until they reach the liver and subsequently invade the hepatocytes. Post-invasion, the sporozoites undergo the first phase of asexual multiplication, forming merozoites which eventually ruptures the hepatocytes, releasing the merozoites into the bloodstream, where they then go on to invade erythrocytes. Once inside the erythrocytes a second phase of asexual replication begins forming between 8 and 16 merozoites. This multiplication subsequently ruptures the erythrocytes releasing the merozoites back into the bloodstream where the cycle of erythrocyte invasion and rupture can

repeat indefinitely. During this time, some merozoites develop into male and female gametocytes, which are then taken back up by the female *Anopheles* mosquito when it feeds for another blood meal. Once inside the mosquito the gametocytes mature and form a ookinete within the lumen of the *Anopheles* gut. The ookinete penetrates the gut wall and forms oocysts which subsequently rupture and release sporozoites which travel to the salivary glands, going back to the start of the replication cycle. For *P. vivax* and *P. ovale spp* a deviation to the life cycle exists whereby both parasites have dormant liver stages referred to as hypnozoites. These hypnozoites can cause a relapse in infection weeks or even months or years after the initial infection date.

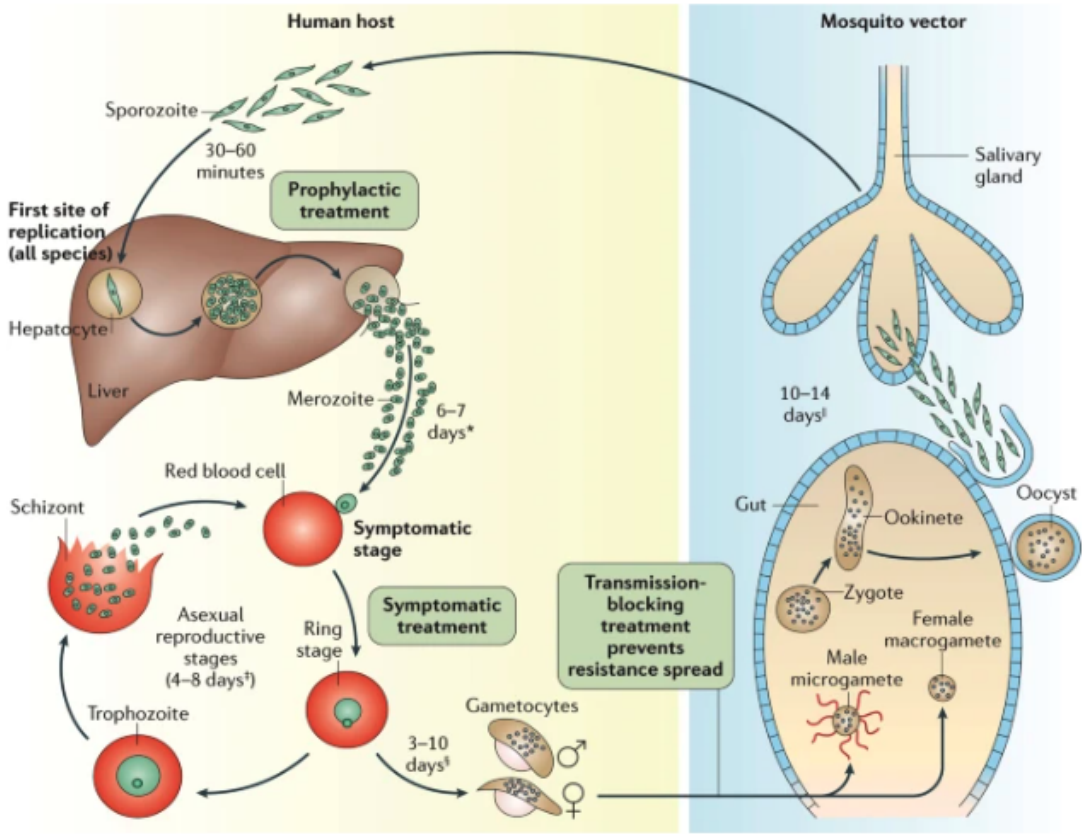


Figure 2. Life cycle of the *Plasmodium* parasite and an indication of what stages are targeted by antimalarial treatments[23].

Malaria pathogenesis

The cyclic invasion and rupturing of erythrocytes is responsible for the symptoms of malaria. In symptomatic individuals, parasite replication increases exponentially and may reach in excess of 10^{12} parasites per patient [24]. Periodicity fevers are a characteristic symptom of malaria, which change depending on the infecting *Plasmodium* parasite. For example, *P. falciparum, P. vivax* and *P. ovale spp* all have a 48-hour cycles whereas *P. malariae* has a 72-hour cycle due to the longer time required for merozoites asexual replication in an infected erythrocyte prior to rupturing. The manifestation of malaria infection can be categorised as either uncomplicated or severe [25]. Uncomplicated malaria is associated with cyclic fevers. Severe malaria is caused by deformed eryrothyctes, causing an obstruction of capillaries, which can result in organ failure, such as cerebral malaria where capillaries in the brain become obstructed leading to coma and subsequently death. Indicators of severe malaria include, acidosis (pH < 7.3), anaemia (Hb <5 g/dl) and high parasitaemia [26].

Several factors influence the risk of developing complex malaria. Pregnant women are three times more at risk of developing severe malaria than the standard population [27]. This is hypothesised to be due to the immunocompromised state of the mother during pregnancy or the sequestering of infected erythrocytes within the placenta resulting in complications. Other biological factors reduce an individual's risk to malaria. For example, heterozygous individuals who carry one β globin gene with the E6V mutation (glutamic acid amino acid is substituted by valine), which causes sickle cell anaemia in homozygous individuals, have a level of resistance to malaria severity[28]. In addition, individuals who lack Duffy antigens harbour natural resistance to *P. vivax*, which uses the receptor Duffy glycoprotein as a receptor during

erythrocyte invasion [29]. However, this has recently been refuted as Duffy-negative erythrocytes have been found to be infected with *P. vivax* [30] .

Malaria treatment

In 1820, quinine was the first compound to be chemically isolated and purified to treat malaria [31]. For centuries the bark of the *Cinchona officinalis*, the source of quinine, was known to have medicinal properties to combat fever. Once treatment with chemically isolated quinine became common practice, the arms race between the development of new antimalarials and emergence of antimalarial resistance began. During World War Two (WW2), lack of access to the *Cinchona* tree resulted in the creation of the first synthetic antimalarial, chloroquine, which subsequently became the dominant antimalarial treatment [32]. *P. falciparum* resistance to chloroquine was first reported in the late 1950s with widespread resistance established in the 1980s [31]. Other treatments developed in the 20th century have been rolled out and subsequently replaced as front-line therapies due to the emergence and spread of resistant parasites and the availability of antimalarials with less adverse effects. Nowadays, malaria treatment is guided by the knowledge of the infecting *Plasmodium* species and also where the infection has been clinically classified as uncomplicated or severe [33]. There are currently 14 medicines for treatment of malaria and 4 for preventative treatment listed by the World Health Organisation (WHO) as essential medicines [34]. Out of the treatments available today, Artemisinin combination therapies (ACTs) are the most effective [33]. Such combination therapies help avoid the emergence of resistance, by targeting the *Plasmodium* parasite through different mechanisms of action (MoA) pathways [35]. For example, artemisinin derivatives are commonly combined with piperaquine and used as a front-line treatment against *P. falciparum* [36]. The MoA for artemisinin derivatives is still widely debated but the most accepted theory to date, is that the compound interacts with haem, found in haemoglobin, to form free radicals which

increase oxidative stress in the parasite until it can no longer survive [37]. The MoA of piperaquine has also not been fully elucidated, but the most accepted hypothesis is that it accumulated in the *Plasmodium* digestive vacuole preventing haem detoxification which subsequently leads to the destruction of the parasite [38] . Reduced susceptibility to Artemisinin was first reported in 2013 for patients in Cambodia and has subsequently spread across Southeast Asia and emerged more recently in Africa [39,40]. The emergence of antimalarial resistance does not mean the therapy should be removed from the WHO essential medicines list. Temporary discontinuation of the antimalarial after the emergence of wide-spread resistance has been shown to lead to an increase of sensitivity in the population such that it can be re-introduced for therapeutic use [41]. In addition, whilst resistance may emerge in one *Plasmodium* species the treatment could still be effective in others. For example, chloroquine forms part of the front-line treatment to *P. vivax* in locations where resistance has not developed in the population [42]. For *P. vivax* and *P. ovale spp* infections, clinicians adjust their treatment to account for the parasite's dormant liver-stage, which if not eradicated, can result in a relapse in infection. Few antimalarials exist to target the liver stage, but primaquine is effective when treatment is adhered to [43,44]. However, individuals with glucose-6-phosphate dehydrogenase (G6PD) deficiency are at risk of severe haemolysis after treatment with primaquine and the drug is not recommended for pregnant women and infants.

Malaria vaccine technologies are beginning to show progress and aid in prevention. *Plasmodium* vaccine development has proven to be historically difficult due to the parasite's multiple immune evasion mechanisms [45]. The vaccine RTS,S/AS01 (RTS,S) is currently being rolled out through the malaria vaccine implementation programme and after 4 doses, given at 0, 1, 2, and 20 months has 36% efficacy against severe malaria, concluded as part of a phase 3 trial involving 15,460 children finishing in 2013 [46]. This falls far short of the 75% efficacy

goal established by the WHO. However, a new vaccine R21, in phase 2b trials, shows promise with efficacy levels of 77% when administered again across 4 doses, across 498 children in 2019 [47].

100 years of Malaria Eradication Strategies

Historic Global Malaria Programs

The first global malaria eradication strategy was undertaken by WHO between 1955–1969 [48]. At the time chloroquine and primaquine were widely available and effective in the treatment and prevention of *Plasmodium* infection, in addition, the insecticide dichlorodiphenyl - trichloroethane (DDT) had been successfully used in vector control and the downstream adverse human and wildlife consequences were unknown [49,50]. As a result, by the 1960s, malaria incidence declined dramatically across Asia and Latin America, with 15 countries eliminating malaria. However, Sub-Saharan Africa was excluded from the global program, due to perceived logistical challenges including poor local health infrastructure. It was noted that eradication in Africa was supposed to be completed at the end of the program, however this was never attempted [48,50]. By 1969 the campaign was stopped, due to the realisation that eradication was not feasible in certain countries and so the strategy was updated to focus on malaria control. This was driven by several factors including resurgence of malaria in Ceylon (now Sri Lanka) between 1968-1969, which was supposed to be a model country. In addition, there was an 85% reduction in funding when the USA stopped contributing to the WHO malaria program in 1963[48].

Across the 1970s and 1980s, malaria funding continued to shrink in real-terms and chloroquine resistance emerged across the globe. This emergence led to a dramatic increase in malaria incidence compared to the previous two decades and severe epidemics in countries including

Turkey and India. However, despite the general global increase in malaria incidence, seven countries declared malaria free status [48].

A new global approach to tackle malaria began in the 1990s with the establishment of a New Malaria Global Control Strategy in 1993 and the Roll Back Initiative in 1998. This strategic refocusing led to the development of new tools in the fight against malaria including, long-lasting insecticide-treated nets (LLINs), rapid diagnostic tests (RDTs), and the previously mentioned artemisinin-based combination therapies (ACTs). The successful deployment of these strategies ended the 1987-2007 hiatus where no countries obtained malaria free status.

Current Global Program

The latest global malaria strategy was outlined in 2015 and forms a 15-year blueprint for malaria control and elimination [51]. This strategy is based around 3 core pillars: (1) to ensure universal access to malaria prevention, diagnosis and treatment; (2) accelerate efforts towards elimination and attainment of malaria-free status; and (3) transform malaria surveillance into a core intervention. Alongside the 3 core pillars are stated milestones (**Table 1)**.

GOALS	MILESTONES		TARGETS
	2020	2025	2030
1. Reduce malaria mortality rates globally compared with 2015	At least 40%	At least 75%	At least 90%
2. Reduce malaria case incidence globally compared with 2015	At least 40%	At least 75%	At least 90%
3. Eliminate malaria from countries in which malaria was transmitted in 2015	At least 10 countries	At least 20 countries	At least 35 countries
4. Prevent re-establishment of malaria in all countries that are malaria-free	Re-establishment prevented	Re-establishment prevented	Re-establishment prevented

Table 1. Objectives of the current WHO malaria eradication program, established in 2015 [51].

The technical strategy also outlined the need for an increase in investment to meet the desired objects. This was to increase annual spending from US$ 2.7 billion to US$ 6.4 billion by 2020, and subsequently to US$ 7.7 billion by 2025, in line with achieving each milestone and striving to achieve the next. Sadly, according to the latest WHO 2020 Malaria Report, the funding goals are not being achieved. Funding in 2019 was estimated at $US 3.0 billion compared to $US 2.7 billion and US$ 3.2 billion for 2018 and 2017 respectively.

Technologies to Combat Malaria

Role of Gene-Target and Whole Genome Sequencing

Genomic sequencing has been pivotal in combating malaria, falling under pillars two and three of the existing WHO eradication strategy [52]. In 1996, the scientific community embarked on the task of sequencing the full genome of *P. falciparum* 3D7, a laboratory culturable strain obtained from a patient in the Netherlands [53]. This task was completed in 2002, revealing the parasite to have a 23 Mb nuclear genome consisting of 14 chromosomes encoding ~5300 genes and two non-nuclear organelles, the mitochondria and the apicoplast [53]. Since publication of the first genome, efforts have been made to capture the genomic diversity across *Plasmodium spp*. The MalariaGen *P. falciparum* community project is one such effort whose November 2020 release contained whole genome sequence (WGS) data for >7,000 samples across 28 countries [54]. This wealth of information has enabled the research community to gain a better understanding of the *Plasmodium* parasite, from guiding drug and vaccine development to the identification and surveillance of resistance markers. However, WGS coverage of each *Plasmodium* species is not even. Whilst *P. falciparum* and *P. vivax* have been well characterised with >7,000 and >1,100 isolates with WGS, <200 such isolates exist for *P. knowlesi* and *P. malariae* and <10 for *P. ovale spp* [55–58].

WGS studies have proved crucial in identifying genetic variations which are associated with resistance [59]. Such variations include, single nucleotide polymorphisms (SNPs), insertions / deletions (INDELS) or large structure variants. Once identified, a given variant can be monitored to increase global surveillance and inform clinicians to adjust their strategy if necessary. For example, in 2014 researchers [60] identified mutations associated with artemisinin resistance in the kelch propeller domain (*Pf.Kelch13*) through determining the genomic differences between resistant and susceptible *P. falciparum* lines [60]. Four non-synonymous SNPs, Y493H, R539T, I543T and C580Y, were identified and retrospective epidemiological studies have since shown the now dominant artemisinin-resistant *P. falciparum* C580Y lineage most likely arose in western Cambodia and then spread across Southeast Asia, outcompeting other parasites and acquiring piperaquine resistance [61]. Genomic surveillance can also be used to guide interventions such as the reintroduction of antimalarials. In Uganda, chloroquine was removed as a front-line therapy in 2006 after widespread resistance emerged and they switched to using ACTs as front-line therapy. However, in 2016, *Plasmodium spp* susceptibility to chloroquine was re-established in Eastern Uganda. A retrospective genomic analysis study revealed that the CVIET allele, associated with chloroquine resistance, decreased from 28.8% in 2013 to 1.1% in 2016 and was not detected in 2017, within Gulu, Northern Uganda [62]. This decrease demonstrates that localised surveillance of resistance markers can be used to optimise local malaria control interventions.

Alongside resistance surveillance, WGS data has enabled the development of molecular barcodes to determine the geographical origin of *Plasmodium* parasites. This is essential for locating the source of potential imported outbreaks, especially when attempting to prevent the reintroduction of *Plasmodium* to malaria free areas [63]. A 23 SNP barcode was developed for *P. falciparum,* which was 92% effective in identifying a parasite origin as South America, West

Africa, East Africa, South East Asia, or Oceania [63]. In addition, for the more widespread *P. vivax* parasite, a 71 SNP barcode was created, providing 91.4% accuracy in predictive ability for the geographic origin of infection [64]. As more genomic data becomes available it is hoped the geographical resolution of specific barcodes can be improved. Whilst the cost of genomics has decreased faster than Moore's law over the past decade it is still expensive, laborious and too specialised to be performed in a routine clinical or field setting [65]. Therefore, we have to rely on other techniques which fall under the umbrella term molecular barcoding tools (MBT) which allow us to detect biomarkers of interest. These tools include but are not limited to nucleic acid amplification techniques (NAATs). MBTs have the potential to not only detect the presence of *Plasmodium spp*, but specific biomarkers of interest such as those associated with antimalarial resistance. The detection of both can be used by clinicians in deciding the relevant treatment pathway or an epidemiologist for local surveillance efforts.

Existing Malaria Diagnostics

Delay in diagnosis and subsequently treatment, is reported as a leading cause of death in malaria patients [66]. The current gold-standard malaria diagnostic is microscopy whereby a trained practitioner determines what *Plasmodium* species is present, based on morphology, and quantifies the parasitemia [67]. The limit of detection (LoD) for this technique is ~50-200 parasites per µl of blood and the cost per test is $US 2.53 according to recent estimates [26,68]. Whilst microscopy is still considered the gold-standard by WHO it has several limitations. First, it is laborious and requires a trained specialist to conduct the diagnosis. *Plasmodium* species misclassification is common, with *P. ovale spp* infections commonly misclassified as *P. vivax* and vice-versa, due to similar morphological features [69]. To compound this, the prevalence of *P. malariae* and *P. ovale spp* infections are under-represented due to false-negative microscopy based-diagnosis as the infections typically manifest with a low

parasitemia[70]. Limitations with microscopy-based malaria diagnosis, led to the development of antigen-detecting rapid diagnostic tests (RDTs) in the early 1990s [71]. These immunochromatography-based diagnostics are easy-to-use and able to provide a result in <30 minutes in low resource field settings. The *Plasmodium* antigens targeted by commonly used RDTs include lactate dehydrogenase (pLDH) for pan-*Plasmodium* detection and histidine rich protein 2 (HRP2) for *P. falciparum* detection [72]. Between 2008 and 2018, a total of 332 RDTs have been assessed by the WHO, with 27 out of 34 products in the last round meeting all necessary criteria [71]. The limit of detection for RDTs is comparable to microscopy at ~50–200 parasites per µl of blood and the cost per test at approximately £0.30 (Chapter 3, Table 1) [68]. The sensitivity of pan-RDTs has been shown to fluctuate depending on the infecting *Plasmodium* species present, with the SD BIOLINE Malaria RDT having a detection ratio of 46.9% for *P. ovale spp* but 93% for *P. vivax* [73,74].

In countries where *P. falciparum* infections account for the majority of cases, most recommend the use of the pfHRP2-based RDTs, however, the efficacy of these tests is under threat [75]. The presence and growing prevalence of *Pf.HRP2/3* gene deletions, first reported in Peru, have now emerged independently across several countries including Uganda [76]. When looking at Uganda as a case study for the impact of this deletion, between 2017 and 2019, the pfHRP2/3 deletion accounted for 12.3% false negatives, which is significantly above the WHO diagnostic efficacy guidelines. This scenario highlights the importance of genomic surveillance as one tool to monitor diagnostic efficacy and in turn inform intervention measures [77]. In line with the objectives of the WHO malaria eradication strategy, as countries attempt to move towards elimination, more sensitive diagnostics are required to detect submicroscopic *Plasmodium* infections which are below the LoD of microscopy or RDTs. Such individuals are typically asymptomatic and can act as reservoirs leading to subsequent *Plasmodium* outbreaks [78].

Nucleic Acid Amplification Based Malaria Diagnostics

Nucleic acid amplification technology (NAAT) based diagnostics, such as polymerase chain reaction (PCR), may provide one such solution, overcoming sensitivity and specificity issues associated with microscopy and immunochromatic-based RDTs [79]. For example, across 1,724 samples in Equatorial Guinea, PCR was used to identify 19.4% and 13.3% false negatives in microscopy and RDT based diagnosis, respectively [79] In addition, a retrospective study based in Papua New Guinea revealed that across 300 participants, pfHRP2-based RDTs missed half of *P. falciparum* infections detected via PCR, including high gametocyte infections associated with high levels of transmission [80].

PCR was the first NAAT pioneered in 1983, and remains the dominant NAAT to date [81]. All NAATs rely on the same fundamental process, the design of primers which are short ssDNA oligonucleotides that bind to specific complementary target sequences following Watson-Crick nucleotide base pairing. Upon primer binding a DNA polymerase is recruited to the 3' terminus of the primer and facilitates DNA extension and amplification [82]. For most NAATs, primer sets are designed to facilitate exponential amplification of the desired target region. PCR was first used for *Plasmodium spp* detection and species-specific identification in the early 1990s [83,84,85,86]. Since then, numerous PCR assays have been developed seeking to improve the sensitivity and specificity of *Plasmodium* detection. One such approach was to target multi-copy regions within *Plasmodium* subtelomeres to enhance assay sensitivity. When benchmarked against a standard 18S rRNA qPCR *Plasmodium* assay this approach proved to be more sensitive, revealing an 8% underestimation in *Plasmodium* prevalence [87].

PCR can also be used for SNP genotyping, opening the door to high-throughput, low-cost detection and surveillance of antimalarial resistance associated genotypes, such as the K76T mutation in the *Pf.CRT* gene linked to chloroquine resistance [88]. In addition, PCR has been combined with sequencing platforms to not only detect known resistance biomarkers but identify novel markers for further investigation [89]. For example, the propeller domain of *Pf.Kelch13* which harbours genotypes associated with artemisinin resistance, can be amplified via PCR and subsequently capillary sequenced, to identify the presence of known resistance biomarkers [90,91]. However, PCR does have its limitations. It requires a thermocycle and so in a similar fashion to microscopy, samples are typically taken back to a laboratory to be tested. This process is laborious and requires a trained practitioner, so is not best suited for use in resource limited settings. Even though PCR has been shown to be more sensitive and specific than both microscopy and existing RDTs, it is not recommended as the gold-standard diagnostic by the WHO due to these limitations.

To improve the field utility of NAATs and move away from the limitations imposed by a thermocycler, isothermal NAATs have been developed which run at a single temperature. This enables the reaction to be performed in-field with a simple heating element. Loop mediated isothermal amplification (LAMP) and recombinase polymerase amplification (RPA) are the two most common isothermal NAATs to date [92,93]. LAMP is performed at 65ºC, the temperature at which dsDNA helixes begin to destabilise, allowing primer binding and subsequent extension via Bst polymerase [92]. Several commercial LAMP kits have been developed to combat malaria, including the Malaria kit LoopAMP® developed by Eiken Chemical. Numerous studies have shown, like PCR, LAMP outperforms typical microscopy and RDTs, however reaction time does have to be limited due to the potential false positives arising through primer interactions leading to non-specific amplification [94].

Recombinase Polymerase Amplification

RPA was pioneered by Neil Arms and Olaf Piepenburg in 2006 who went on to found TwistDx LTD [93]. RPA reactions are performed at 37-39ºC and can be powered using human body heat, removing the dependency on a heating-block in resource limited settings [95]. RPA can be performed at such a low temperature, compared to other NAATs, due to its reliance on three key T4 phage proteins, UvsX, UvsY and Gp32. A summary of the RPA reaction process is shown (**Figure 3).**

UvsX is an ortholog to the well characterised RecA recombinase protein and contains two DNA binding sites. Upon reaction initiation, UvsX proteins bind to ssDNA primers forming protein-nucleotide complexes, sometimes individually referred to as a presynaptic filament complex [96]. This complex then identifies and binds to the homologous region in dsDNA. ATP binding is required for this homologous pairing and subsequent ATP hydrolysis, drives strand exchange, enabling primer of the target site, to which the strand displacing Bsu polymerase is recruited too and subsequently facilitates amplification [97].

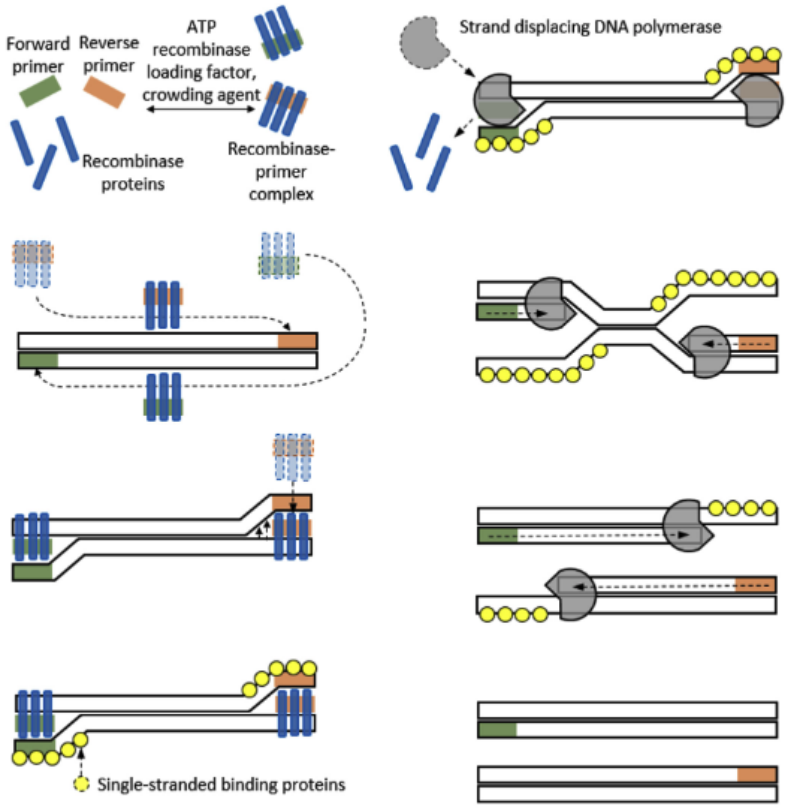


Figure 3. Schematic highlighting the stages of recombinase polymerase amplification (RPA).[98]

During amplification Gp32 helps to stabilise the ssDNA displaced during strand-exchange / D-loop formation, preventing the inserted ssDNA being displaced via branch-migration. As both UvsX and Gp32 are non-specific ssDNA binding proteins they compete to bind the primers on reaction initiation however, UvsY acts as a mediator ensuring that UvsX outcompetes Gp32. UvsY has been shown to co-occupy ssDNA binding sites simultaneously with UvsX and Gp32. When bound to Gp32-ssDNA complexes, UvsY destabilises the ssDNA interaction. In addition, studies have shown the UvsY stabilises / strengthens UvsX-ssDNA binding interactions when bound cooperatively. In summary, UvsX, UvsY and Gp32 together replace the denaturation and annealing steps of a typical PCR cycle. In vivo these proteins are essential

in maintaining genetic diversity of the T4-phage population by facilitating homologous recombination as well as the repair of dsDNA breaks [99].

Previous research has shown that RPA can match the sensitivity and specificity of ultra-sensitive PCR in the detection of *P. falciparum*, including samples from asymptomatic individuals [100]. In addition, RPA can be adapted for immunochromatic-based end point detection through the use of antigen-labelled primers / probes which subsequently create a dual-tagged amplicon. This approach has been applied in the detection of *P. knowlesi* and makes the technology more appropriate for use in-field settings. [101]. When combining these advantages, RPA holds the potential to revolutionise malaria diagnostics, not only with accurate in-field pathogen detection but also opening the door to in-field SNP genotyping. To date, no commercially available RPA-based diagnostic kits are available for malaria detection. In addition, the TwistDX Cambridge research site was closed after Alere, the parent company of TwistDX, was obtained by Abbot, and work to develop the technology from within industry has halted. As such the responsibility to enhance the RPA technology falls to academic labs.

Outline of Thesis.

In this thesis I outline my approach, integrating both wet and dry lab techniques, to enhance RPA with the goal of create a next-generation rapid malaria diagnostics test, in line with the existing WHO malaria eradication strategy. In **Chapter 2 (**Manuscript), I describe the construction of a RPA-specific bioinformatics tool, PrimedRPA, to assist with and enhance RPA assay design, overcoming limitations of manual assay design. With this tool at hand, I next sought to adapt RPA for low-cost, field-use, colorimetric-based reaction end point detection, **Chapter 3 (**Written Chapter), aligning unit economics with existing RDT

solutions. For RPA to truly fulfil its potential as a next generation malaria diagnostic tool, assays need to move beyond pathogen detection and enable the simultaneous detection of key biomarkers. RPAs genotyping ability is first explored in **Chapter 3** with the detection of biomarkers associated with antimalarial resistance, facilitated by the deliberate introduction of primer-template mismatches. This approach I explore fully in **Chapter 4** (Manuscript), through a systematic assessment of the impact primer-template mismatch combinations have on RPA reaction kinetics. After gauging the detrimental impact that even a single mismatch can have on RPA assay performance, **Chapter 5** (Manuscript), outlines the creation of the PrimedInclusivity software, enabling users to take advantage of existing whole genome sequence data to assess the conservation of the assay target site, in turn enhancing assay design. Not only does this tool allow users to quantify the presence of mismatches within a target population but predict the impact of a given mismatch on assay performance. Building on this, when reviewing the whole genome sequence data available for *Plasmodium* parasites, limited genomic characterisation of the neglected parasites *P. ovale curtisi* and *P. ovale walkeri* species has been completed. Recent reports indicate that these typically neglected parasites are increasing in prevalence and as such I sought to enhance their genomic characterisation, in line with the third pillar of the existing WHO malaria eradication strategy. **Chapter 6** (Manuscript) outlines my approach to achieve this in the design of a new SWGA primer set to enhance *P. ovale spp* whole genome sequencing efforts and the creation of two new reference genomes for *P. ovale curtisi* and *P. ovale walkeri* respectively. Finally, **Chapter 7**, discusses the overarching findings of my work and future perspectives for the role of RPA in the next generation of malaria diagnostics.

References

1. Sato, S. Plasmodium-a brief introduction to the parasites causing human malaria and their basic biology. *J. Physiol. Anthropol.* **40**, 1 (2021).
2. Singh, B. & Daneshvar, C. Human infections and detection of Plasmodium knowlesi. *Clin. Microbiol. Rev.* **26**, 165–184 (2013).
3. Lempang, M. E. P. *et al.* Primate malaria: An emerging challenge of zoonotic malaria in Indonesia. *One Health* **14**, 100389 (2022).
4. Sabbatani, S., Manfredi, R. & Fiorino, S. Malaria infection and human evolution. *Infez. Med.* **18**, 56–74 (2010).
5. Sallares, R., Bouwman, A. & Anderung, C. The spread of malaria to Southern Europe in antiquity: new approaches to old problems. *Med. Hist.* **48**, 311–328 (2004).
6. Carter, R. & Mendis, K. N. Evolutionary and historical aspects of the burden of malaria. *Clin. Microbiol. Rev.* **15**, 564–594 (2002).
7. Bourzac, K. Infectious disease: Beating the big three. *Nature* **507**, S4–7 (2014).
8. Organization, W. H. & Others. World malaria report 2021. (2021).
9. Roberts, D. & Matthews, G. Risk factors of malaria in children under the age of five years old in Uganda. *Malar. J.* **15**, 246 (2016).
10. Cibulskis, R. E. *et al.* Malaria: Global progress 2000 - 2015 and future challenges. *Infect Dis Poverty* **5**, 61 (2016).
11. White, N. J., Day, N. P. J., Ashley, E. A., Smithuis, F. M. & Nosten, F. H. Have we really failed to roll back malaria? *Lancet* **399**, 799–800 (2022).
12. Mironova, V. A. *et al.* Re-introduction of vivax malaria in a temperate area (Moscow region, Russia): a geographic investigation. *Malar. J.* **19**, 116 (2020).
13. Caminade, C. *et al.* Impact of climate change on global malaria distribution. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 3286–3291 (2014).

14. Fischer, L., Gültekin, N., Kaelin, M. B., Fehr, J. & Schlagenhauf, P. Rising temperature and its impact on receptivity to malaria transmission in Europe: A systematic review. *Travel Med. Infect. Dis.* **36**, 101815 (2020).
15. Carlson, C. J. *et al.* Solar geoengineering could redistribute malaria risk in developing countries. *Nat. Commun.* **13**, 2150 (2022).
16. Alonso, S. *et al.* The economic burden of malaria on households and the health system in a high transmission district of Mozambique. *Malar. J.* **18**, 360 (2019).
17. Gallup, J. L. & Sachs, J. D. *The Economic Burden of Malaria*. (American Society of Tropical Medicine and Hygiene, 2001).
18. Cox, F. E. History of the discovery of the malaria parasites and their vectors. *Parasit. Vectors* **3**, 5 (2010).
19. Grassi, B. *Studio di uno zoologo sulla malaria*. vol. 3 (R. Accademia dei lincei, 1900).
20. Stephens, J. W. W. A New Malaria Parasite of Man. *Ann. Trop. Med. Parasitol.* **16**, 383–388 (1922).
21. Knowles, R. & Gupta, B. M. D. A Study of Monkey-Malaria, and Its Experimental Transmission to Man. *Ind. Med. Gaz.* **67**, 301–320 (1932).
22. Sinka, M. E. Global distribution of the dominant vector species of malaria. in *Anopheles mosquitoes - New insights into malaria vectors* (InTech, 2013).
23. Phillips, M. A. *et al.* Malaria. *Nat Rev Dis Primers* **3**, 17050 (2017).
24. Trampuz, A., Jereb, M., Muzlovic, I. & Prabhu, R. M. Clinical review: Severe malaria. *Crit. Care* **7**, 315–323 (2003).
25. Bartoloni, A. & Zammarchi, L. Clinical aspects of uncomplicated and severe malaria. *Mediterr. J. Hematol. Infect. Dis.* **4**, e2012026 (2012).
26. Lalloo, D. G. *et al.* UK malaria treatment guidelines 2016. *J. Infect.* **72**, 635–649 (2016).
27. Schantz-Dunn, J. & Nour, N. M. Malaria and pregnancy: a global health perspective.

Rev. Obstet. Gynecol. **2**, 186–192 (2009).

28. Eleonore, N. L. E. *et al.* Malaria in patients with sickle cell anaemia: burden, risk factors and outcome at the Laquintinie hospital, Cameroon. *BMC Infect. Dis.* **20**, 40 (2020).
29. Dean, L. *The Duffy blood group*. (National Center for Biotechnology Information (US), 2005).
30. Golassa, L., Amenga-Etego, L., Lo, E. & Amambua-Ngwa, A. The biology of unconventional invasion of Duffy-negative reticulocytes by Plasmodium vivax and its implication in malaria epidemiology and public health. *Malar. J.* **19**, 299 (2020).
31. Achan, J. *et al.* Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malar. J.* **10**, 144 (2011).
32. Packard, R. M. The origins of antimalarial-drug resistance. *N. Engl. J. Med.* **371**, 397–399 (2014).
33. *WHO Guidelines for malaria*. (World Health Organization, 2022).
34. Tse, E. G., Korsik, M. & Todd, M. H. The past, present and future of anti-malarial medicines. *Malar. J.* **18**, 93 (2019).
35. Sinclair, D., Zani, B., Donegan, S., Olliaro, P. & Garner, P. Artemisinin-based combination therapy for treating uncomplicated malaria. *Cochrane Database Syst. Rev.* (2009) doi:10.1002/14651858.CD007483.pub2.
36. Wang, Q. *et al.* Efficacy and Safety of Artemisinin-Piperaquine for the Treatment of Uncomplicated Malaria: A Systematic Review. *Front. Pharmacol.* **11**, 562363 (2020).
37. Wang, J. *et al.* Artemisinin, the Magic Drug Discovered from Traditional Chinese Medicine. *Proc. Est. Acad. Sci. Eng.* **5**, 32–39 (2019).
38. Edgar, R. C. S., Counihan, N. A., McGowan, S. & de Koning-Ward, T. F. Methods Used to Investigate the Plasmodium falciparum Digestive Vacuole. *Front. Cell. Infect. Microbiol.* **11**, 829823 (2021).

39. Amato, R. *et al.* Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study. *Lancet Infect. Dis.* **18**, 337–345 (2018).

40. Rosenthal, P. J. Has artemisinin resistance emerged in Africa? *The Lancet infectious diseases* vol. 21 1056–1057 (2021).

41. Dagnogo, O. *et al.* Towards a re-emergence of chloroquine sensitivity in Côte d'Ivoire? *Malar. J.* **17**, 413 (2018).

42. Waqar, T., Khushdil, A. & Haque, K. Efficacy of Chloroquine as a first line agent in the treatment of uncomplicated malaria due to Plasmodium vivax in children and treatment practices in Pakistan: A Pilot study. *J. Pak. Med. Assoc.* **66**, 30–33 (2016).

43. Baird, J. K. & Hoffman, S. L. Primaquine therapy for malaria. *Clin. Infect. Dis.* **39**, 1336–1345 (2004).

44. Shimizu, S. *et al.* Optimal primaquine use for radical cure of Plasmodium vivax and Plasmodium ovale malaria in Japanese travelers--A retrospective analysis. *Travel Med. Infect. Dis.* **13**, 235–240 (2015).

45. Riley, E. M. & Stewart, V. A. Immune mechanisms in malaria: new insights in vaccine development. *Nat. Med.* **19**, 168–178 (2013).

46. RTS,S Clinical Trials Partnership. Efficacy and safety of the RTS,S/AS01 malaria vaccine during 18 months after vaccination: a phase 3 randomized, controlled trial in children and young infants at 11 African sites. *PLoS Med.* **11**, e1001685 (2014).

47. Datoo, M. S. *et al.* Efficacy of a low-dose candidate malaria vaccine, R21 in adjuvant Matrix-M, with seasonal administration to children in Burkina Faso: a randomised controlled trial. *Lancet* **397**, 1809–1818 (2021).

48. Nájera, J. A., González-Silva, M. & Alonso, P. L. Some lessons for the future from the Global Malaria Eradication Programme (1955-1969). *PLoS Med.* **8**, e1000412 (2011).

49. Bouwman, H., van den Berg, H. & Kylin, H. DDT and malaria prevention: addressing

the paradox. *Environmental health perspectives* vol. 119 744–747 (2011).

50. Breman, J. G. & Brandling-Bennett, A. D. The challenge of malaria eradication in the twenty-first century: research linked to operations is the key. *Vaccine* **29 Suppl 4**, D97–103 (2011).
51. Programme, G. M. Global technical strategy for malaria 2016-2030, 2021 update. https://www.who.int/publications/i/item/9789240031357 (2021).
52. Neafsey, D. E., Taylor, A. R. & MacInnis, B. L. Advances and opportunities in malaria population genomics. *Nat. Rev. Genet.* **22**, 502–517 (2021).
53. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).
54. MalariaGEN *et al.* An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. *Wellcome Open Res.* **6**, 42 (2021).
55. Benavente, E. D. *et al.* Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. *Nat. Commun.* **12**, 3160 (2021).
56. Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature* **542**, 101–104 (2017).
57. Benavente, E. D. *et al.* Whole genome sequencing of amplified Plasmodium knowlesi DNA from unprocessed blood reveals genetic exchange events between Malaysian Peninsular and Borneo subpopulations. *Sci. Rep.* **9**, 9873 (2019).
58. Hocking, S. E., Divis, P. C. S., Kadir, K. A., Singh, B. & Conway, D. J. Population Genomic Structure and Recent Evolution of Plasmodium knowlesi, Peninsular Malaysia. *Emerg. Infect. Dis.* **26**, 1749–1758 (2020).
59. Hupalo, D. N. *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat. Genet.* **48**, 953–958 (2016).
60. Ariey, F. *et al.* A molecular marker of artemisinin-resistant Plasmodium falciparum

malaria. *Nature* **505**, 50–55 (2014).

61. Imwong, M. *et al.* The spread of artemisinin-resistant Plasmodium falciparum in the Greater Mekong subregion: a molecular epidemiology observational study. *Lancet Infect. Dis.* **17**, 491–497 (2017).
62. Balikagala, B. *et al.* Recovery and stable persistence of chloroquine sensitivity in Plasmodium falciparum parasites after its discontinued use in Northern Uganda. *Malar. J.* **19**, 76 (2020).
63. Preston, M. D. *et al.* A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. *Nat. Commun.* **5**, 4052 (2014).
64. Diez Benavente, E. *et al.* A molecular barcode to inform the geographical origin and transmission dynamics of Plasmodium vivax malaria. *PLoS Genet.* **16**, e1008576 (2020).
65. November, J. More than Moore's Mores: Computers, Genomics, and the Embrace of Innovation. *J. Hist. Biol.* **51**, 807–840 (2018).
66. Mayor, A. & Bassat, Q. 'Resistance' to diagnostics: A serious biological challenge for malaria control and elimination. *EBioMedicine* **50**, 9–10 (2019).
67. Wongsrichanalai, C., Barcus, M. J., Muth, S., Sutamihardja, A. & Wernsdorfer, W. H. *A Review of Malaria Diagnostic Tools: Microscopy and Rapid Diagnostic Test (RDT)*. (American Society of Tropical Medicine and Hygiene, 2007).
68. Conteh, L. *et al.* Costs and Cost-Effectiveness of Malaria Control Interventions: A Systematic Literature Review. *Value Health* **24**, 1213–1222 (2021).
69. Kotepui, M., Masangkay, F. R., Kotepui, K. U. & De Jesus Milanez, G. Misidentification of Plasmodium ovale as Plasmodium vivax malaria by a microscopic method: a meta-analysis of confirmed P. ovale cases. *Sci. Rep.* **10**, 21807 (2020).
70. Mueller, I., Zimmerman, P. A. & Reeder, J. C. Plasmodium malariae and Plasmodium

ovale--the ‘bashful’ malaria parasites. *Trends Parasitol.* **23**, 278–283 (2007).

71. Cunningham, J. *et al.* A review of the WHO malaria rapid diagnostic test product testing programme (2008-2018): performance, procurement and policy. *Malar. J.* **18**, 387 (2019).
72. Coldiron, M. E. *et al.* Clinical diagnostic evaluation of HRP2 and pLDH-based rapid diagnostic tests for malaria in an area receiving seasonal malaria chemoprevention in Niger. *Malar. J.* **18**, 443 (2019).
73. Tang, J. *et al.* Assessment of false negative rates of lactate dehydrogenase-based malaria rapid diagnostic tests for Plasmodium ovale detection. *PLoS Negl. Trop. Dis.* **13**, e0007254 (2019).
74. Pujo, J. M. *et al.* Accuracy of SD Malaria Ag P.f/Pan® as a rapid diagnostic test in French Amazonia. *Malar. J.* **20**, 369 (2021).
75. Bosco, A. B. *et al.* Limitations of rapid diagnostic tests in malaria surveys in areas with varied transmission intensity in Uganda 2017-2019: Implications for selection and use of HRP2 RDTs. *PLoS One* **15**, e0244457 (2020).
76. Gamboa, D. *et al.* A large proportion of P. falciparum isolates in the Amazon region of Peru lack pfhrp2 and pfhrp3: implications for malaria rapid diagnostic tests. *PLoS One* **5**, e8091 (2010).
77. Organization, W. H. & Others. *False-negative RDT results and implications of new reports of P. falciparum histidine-rich protein 2/3 gene deletions*. https://apps.who.int/iris/bitstream/handle/10665/258972/WHO-HTM-GMP-2017.18-eng.pdf (2017).
78. Tadesse, F. G. *et al.* The Relative Contribution of Symptomatic and Asymptomatic Plasmodium vivax and Plasmodium falciparum Infections to the Infectious Reservoir in a Low-Endemic Setting in Ethiopia. *Clin. Infect. Dis.* **66**, 1883–1891 (2018).

79. Berzosa, P. *et al.* Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea. *Malar. J.* **17**, 333 (2018).

80. Hofmann, N. E. *et al.* Assessment of ultra-sensitive malaria diagnosis versus standard molecular diagnostics for malaria elimination: an in-depth molecular community cross-sectional study. *Lancet Infect. Dis.* **18**, 1108–1116 (2018).

81. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51 Pt 1**, 263–273 (1986).

82. Garibyan, L. & Avashia, N. Polymerase chain reaction. *J. Invest. Dermatol.* **133**, 1–4 (2013).

83. Jaureguiberry, G., Hatin, I., d'Auriol, L. & Galibert, G. PCR detection of Plasmodium falciparum by oligonucleotide probes. *Mol. Cell. Probes* **4**, 409–414 (1990).

84. Wataya, Y. *et al.* DNA diagnosis of falciparum malaria. *Nucleic Acids Symp. Ser.* 155–156 (1991).

85. Kimura, M. *et al.* Species-specific PCR detection of malaria parasites by microtiter plate hybridization: clinical study with malaria patients. *J. Clin. Microbiol.* **33**, 2342–2346 (1995).

86. Snounou, G., Viriyakosol, S., Jarra, W., Thaithong, S. & Brown, K. N. Identification of the four human malaria parasite species in field samples by the polymerase chain reaction and detection of a high prevalence of mixed infections. *Mol. Biochem. Parasitol.* **58**, 283–292 (1993).

87. Ballard, E. *et al.* A validation study of microscopy versus quantitative PCR for measuring Plasmodium falciparum parasitemia. *Trop. Med. Health* **47**, 49 (2019).

88. Vessière, A., Berry, A., Fabre, R., Benoit-Vical, F. & Magnaval, J.-F. Detection by real-time PCR of the Pfcrt T76 mutation, a molecular marker of chloroquine-resistant

Plasmodium falciparum strains. *Parasitol. Res.* **93**, 5–7 (2004).

89. Osborne, A. *et al.* Characterizing the genomic variation and population dynamics of Plasmodium falciparum malaria parasites in and around Lake Victoria, Kenya. *Sci. Rep.* **11**, 19809 (2021).
90. Gaye, A. *et al.* Amplicon deep sequencing of kelch13 in Plasmodium falciparum isolates from Senegal. *Malar. J.* **19**, 134 (2020).
91. Uwimana, A. *et al.* Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. *Nat. Med.* **26**, 1602–1608 (2020).
92. Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res.* **28**, E63 (2000).
93. Piepenburg, O., Williams, C. H., Stemple, D. L. & Armes, N. A. DNA detection using recombination proteins. *PLoS Biol.* **4**, e204 (2006).
94. Ponce, C. *et al.* Diagnostic accuracy of loop-mediated isothermal amplification (LAMP) for screening patients with imported malaria in a non-endemic setting. *Parasite* **24**, 53 (2017).
95. Kong, M. *et al.* A wearable microfluidic device for rapid detection of HIV-1 DNA using recombinase polymerase amplification. *Talanta* **205**, 120155 (2019).
96. Farb, J. N. & Morrical, S. W. Functional complementation of UvsX and UvsY mutations in the mediation of T4 homologous recombination. *Nucleic Acids Res.* **37**, 2336–2345 (2009).
97. Maher, R. L. & Morrical, S. W. Coordinated Binding of Single-Stranded and Double-Stranded DNA by UvsX Recombinase. *PLoS One* **8**, e66654 (2013).
98. Lobato, I. M. & O'Sullivan, C. K. Recombinase polymerase amplification: Basics, applications and recent advances. *Trends Analyt. Chem.* **98**, 19–35 (2018).

99. Liu, J. & Morrical, S. W. Assembly and dynamics of the bacteriophage T4 homologous recombination machinery. *Virol. J.* **7**, 357 (2010).

100. Lalremruata, A. *et al.* Recombinase Polymerase Amplification and Lateral Flow Assay for Ultrasensitive Detection of Low-Density Plasmodium falciparum Infection from Controlled Human Malaria Infection Studies and Naturally Acquired Infections. *J. Clin. Microbiol.* **58**, (2020).

101. Lai, M.-Y., Ooi, C.-H. & Lau, Y.-L. Recombinase Polymerase Amplification Combined with a Lateral Flow Strip for the Detection of Plasmodium knowlesi. *Am. J. Trop. Med. Hyg.* **98**, 700–703 (2018).

Chapter 2. PrimedRPA: primer design for recombinase polymerase amplification assays.

RESEARCH PAPER COVER SHEET

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

<u>SECTION A – Student Details</u>

Student ID Number	1702842	**Title**	Mr
First Name(s)	Matthew		
Surname/Family Name	Higgins		
Thesis Title	Developing an RPA-based Molecular Barcoding Tool for Plasmodium Malaria		
Primary Supervisor	Prof. Taane Clark		

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

<u>SECTION B – Paper already published</u>

Where was the work published?	Bioinformatics		
When was the work published?	2019		
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.

<u>SECTION C – Prepared for publication, but not yet published</u>

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

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 designed and built the the PrimedRPA software. I undertook the necessary laboratory work to validate the software. I wrote the first draft of the manuscript which was then circulated to supervisors and co-authors.

SECTION E

Student Signature	[redacted]
Date	26/06/2022

Supervisor Signature	[redacted]
Date	26/6/2022

Bioinformatics, 35(4), 2019, 682–684
doi: 10.1093/bioinformatics/bty701
Advance Access Publication Date: 8 August 2018
Applications Note

Sequence analysis

PrimedRPA: primer design for recombinase polymerase amplification assays

Matthew Higgins[1], Matt Ravenhall [1], Daniel Ward[1], Jody Phelan [1], Amy Ibrahim[1], Matthew S. Forrest[2], Taane G. Clark[1,3,*,†] and Susana Campino [1,*,†]

[1]Pathogen Molecular Biology Department, London School of Hygiene and Tropical Medicine (LSHTM), London WC1E 7HT, UK, [2]TwistDx, Coldhams Business Park, Cambridge CB1 3LH, UK and [3]Department of Infectious Disease Epidemiology, LSHTM, London WC1E 7HT, UK

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
Associate Editor: John Hancock

Received on May 24, 2018; revised on July 31, 2018; editorial decision on August 3, 2018; accepted on August 7, 2018

Abstract

Summary: Recombinase polymerase amplification (RPA), an isothermal nucleic acid amplification method, is enhancing our ability to detect a diverse array of pathogens, thereby assisting the diagnosis of infectious diseases and the detection of microorganisms in food and water. However, new bioinformatics tools are needed to automate and improve the design of the primers and probes sets to be used in RPA, particularly to account for the high genetic diversity of circulating pathogens and cross detection of genetically similar organisms. PrimedRPA is a python-based package that automates the creation and filtering of RPA primers and probe sets. It aligns several sequences to identify conserved targets, and filters regions that cross react with possible background organisms.
Availability and implementation: PrimedRPA was implemented in Python 3 and supported on Linux and MacOS and is freely available from http://pathogenseq.lshtm.ac.uk/PrimedRPA.html.
Contact: taane.clark@lshtm.ac.uk or susana.campino@lshtm.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The last decade has seen a prodigious increase in the development and adaptation of novel and existing isothermal amplification technologies for molecular diagnostics. Recombinase Polymerase Amplification (RPA) enables both sensitive and rapid isothermal DNA amplification (Piepenburg *et al.*, 2006). RPA is establishing itself as a robust alternative to PCR, and becoming a molecular tool of choice for the rapid, specific, and cost-effective identification of pathogens. Its minimal sample preparation requirements, low operation temperature (25–42°C), and commercial availability of freeze-dried reagents, mean this method has been applied in field laboratory settings and on-board automated sample-to-answer microfluidic devices. Further, this technique can be performed directly in non-processed samples, such as whole blood (Magro *et al.*, 2017).

There is no automated software for designing primer-probe sets for RPA. Identifying candidate regions for assay development can be difficult as regions need to be conserved, with little homology to potential background DNA. Also, the sequence for primers and a probe to bind should create as small an amplicon as possible. Any DNA in the reaction that is not the target can be considered as background. Existing primer design software such as *Primer3* (Untergasser *et al.*, 2012) and *RExPrimer* (Piriyapongsa *et al.*, 2009) cannot be used to automate TwistAmp® exo probe design, as they are typically longer than what these programs allow and specific requirements need to be met, including the positioning of two thymidine residues in the probe to which the fluorophore and quencher are attached. To overcome these issues, we developed *Primer design for RPA* (*PrimedRPA*), which automates the RPA primer and exo

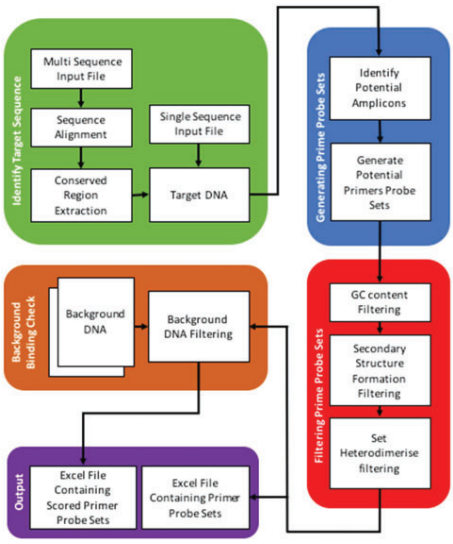


Fig. 1. The analytical pipeline of *PrimedRPA*

probe design process. In addition, as RPA is permissive to the presence of SNPs, the software can input and align several target sequences to account for the high genetic diversity of circulating pathogens. The software can also input several background sequences to avoid the design of primer/probes that can cross react with genetically similar organisms. Here we test the software against several pathogens and validate some of the resulting primers in the laboratory.

2 Materials and methods

The *PrimedRPA* package, developed in python, creates and filters RPA primer-probe sets specific for target DNA sequence(s). An overview of the package is presented in Figure 1. The user defines the input sequence(s), the parameters for filtering, and the sequence files for the background binding check, through altering the *PrimedRPA_Parameters.txt* file. The filtering parameters include primer, probe and amplicon lengths, GC content, ability to form a secondary structure and heterodimerise, and the tolerance of binding to background DNA (Fig. 1, Red and Brown). The user can input a single target sequence or multiple sequences. When the user inputs a sequence file ('fasta format') containing multiple sequences, an initial alignment is produced and conserved regions are extracted as target DNA. If a single sequence is inputted it will be taken as the target DNA (Fig. 1, Green). Candidate RPA primer-probe sets are then generated. These preliminary sets undergo filtering based on user-defined parameters. If a background binding check is required a filtered set is presented in ascending order of a score that reflects the primer-probe sets ability to bind to background DNA, where smaller scores reflect a lower probability of binding. The probes are exported as raw sequences allowing the user to choose where to insert the fluorophore, dSpacer and quencher. The script guarantees the presence of two thymidine residues in the middle region of the probe for the fluorophore and quencher to be attached to.

3 Results

To assess the performance of *PrimedRPA*, we attempted to identify primers-probe sets in pathogens that had previously been published. For *Streptococcus pneumoniae* we used the *lepA* gene as a target (3170 bp) and for the *Bovine ephemeral fever* virus (BEFV) a terminal region in the genome (460 bp). Within 6 s, we identified 71 and 138 primers-probes sets for S. *pneumoniae* and *BEF*, respectively, including some overlapping with previously published RPA sets (Hou *et al.*, 2017; Clancy *et al.*, 2015) (Supplementary Tables S1 and S2 for parameters and output examples). We also tested the software to identify primers-probes that could amplify Zika virus from any geographical region. By using 105 Zika sequences sourced globally, *PrimedRPA* identified 140 potential primer-probes sets that would bind to Zika independently of the genetic diversity. To demonstrate the utility in a setting where there is high inter-species similarities, the mitochondrial (mt) sequence (6 kb) of the *Plasmodium vivax* malaria parasite was processed with a background check using 495 mt sequences from the five other human infecting plasmodium species. Several potential sets were generated and we validated one set of primers in the laboratory (Supplementary Table S1 and File S1) that passed the background binding check. Sanger capillary sequencing confirmed that primers were specific for *P. vivax*, even in samples with mixed *P. vivax* and *P falciparum* DNA (Supplementary File S1).

4 Discussion

Automating the primer and exo probe design process for RPA will assist with implementing this technique and provide a stepping stone for its broader application in diagnostic tests. We have developed an *in silico* assay design tool, which provides multiple possible primers and probes that can be screened and optimized *in vitro* with the RPA technology. TwistAmp® exo fluorescent probes can be converted into lateral flow probes, and therefore the *PrimedRPA* package could be used to design such applications. Further, the software can be extended as nucleic acid amplification detection kits continue to evolve and their applications in biomedical settings increase.

Acknowledgements

The MRC eMedlab was used for computational work.

Funding

This work was supported by BBSRC LiDO PHD studentship (M.H. and M.R.), MRC LID PhD studentship (A.I.) and Bloomsbury Research Fund PhD studentship (D.W.). T.G.C. and S.C. are funded by MRC UK grants (MR/K000551/1, MR/M01360X/1, MR/N010469/1).

Conflict of Interest: MSF is an employee of TwistDx, the developer and manufacturer of RPA technology.

References

Clancy,E. *et al.* (2015) Development of a rapid recombinase polymerase amplification assay for the detection of Streptococcus pneumoniae in whole blood. *BMC Infect. Dis.*, **15**, 481.

Hou,P. *et al.* (2017) Development of a recombinase polymerase amplification combined with lateral-flow dipstick assay for detection of bovine ephemeral fever virus. *Mol. Cell. Probes*, **38**, 31–37.

Magro,L. *et al.* (2017) Paper-based RNA detection and multiplexed analysis for Ebola virus diagnostics. *Sci. Rep.*, **7**, 1347.

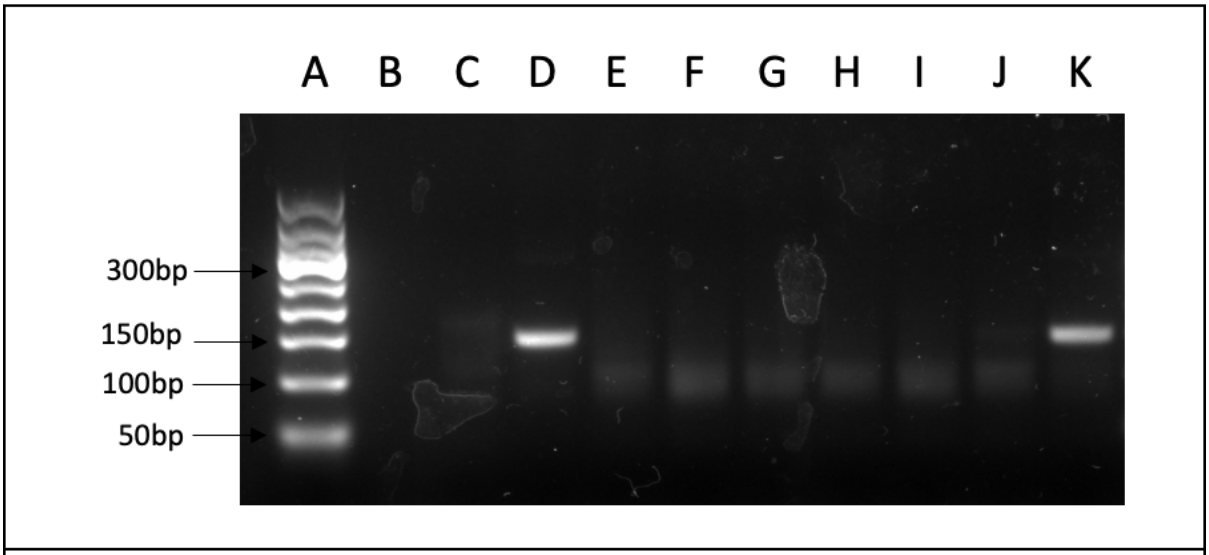
Piepenburg,O. *et al*. (2006) DNA detection using recombination proteins. *PLoS Biol*, **4**, e204.

Piriyapongsa,J. *et al*. (2009) RExPrimer: an integrated primer designing tool increases PCR effectiveness by avoiding 3' SNP-in-primer and mis-priming from structural variation. *BMC Genomics*, **10**, S4.

Untergasser,A. *et al*. (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res*., **40**, e115.

PrimedRPA Supplementary Information

A)



Supplementary Figure 1. 2% agarose gel showing outcome of testing the *Plasmodium vivax* specific primers on all human infecting *Plasmodium* species. The DNA present in each well is outlined as follows: A) Qiagen GelPilot 50bp Ladder. B) Negative control (no template & no primers). C) Negative control (no template) D) *P.vivax* E) *P. falciparum* F) *P. malariae* G) *P. knowlesi* H) *P. ovale spp* I) *Human* J) *P. falciparum* K) *P. vivax* & *P. falciparum.* RPA reaction was run for 30 minutes at 40°C. Clean up performed using SureClean Plus (Bioline).

B) Sanger Sequencing Results (*P. vivax* and *P. falciparum* mixed Sample K in gel)

TTACCTAGATACTATAGTTGAACAGGACATATACATATATTCATTATTCTGAATA
GAAAAAGAACTCTATAAATAACCATATAATTTCAACAAAATGCCAGTATAATAT
TGTAG

PlasmoDB Blast Results

BLASTN 2.8.0+
Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

Reference for database indexing: Aleksandr Morgulis, George Coulouris, Yan Raytselis, Thomas L. Madden, Richa Agarwala, Alejandro A. Schaffer (2008), "Database Indexing for Production MegaBLAST Searches", Bioinformatics 24:1757-1764.

RID: NHVZVZ2J014

Database: Nucleotide collection (nt)
48,886,869 sequences; 184,591,207,883 total letters
Query=
Length=114

Sequences producing significant alignments:	Score (Bits)	E Value
LT635627.1 Plasmodium vivax genome assembly, organelle: mitoc...	196	8e-47
KF668406.1 Plasmodium vivax isolate 54CDC cytochrome oxidase ...	196	8e-47
JQ240416.1 Plasmodium vivax isolate V08A32 mitochondrion, com...	196	8e-47
JQ240391.1 Plasmodium vivax isolate GX5 mitochondrion, comple...	196	8e-47
JQ240387.1 Plasmodium vivax isolate GX29 mitochondrion, compl...	196	8e-47
JQ240375.1 Plasmodium vivax isolate GX15 mitochondrion, compl...	196	8e-47
JQ240368.1 Plasmodium vivax isolate GW4 mitochondrion, comple...	196	8e-47
JQ240360.1 Plasmodium vivax isolate 200667 mitochondrion, com...	196	8e-47
JQ240353.1 Plasmodium vivax isolate 200647 mitochondrion, com...	196	8e-47
JQ240351.1 Plasmodium vivax isolate 200645 mitochondrion, com...	196	8e-47
JQ240348.1 Plasmodium vivax isolate 200633 mitochondrion, com...	196	8e-47
JQ240346.1 Plasmodium vivax isolate 200629 mitochondrion, com...	196	8e-47
JQ240345.1 Plasmodium vivax isolate 200627 mitochondrion, com...	196	8e-47
JQ240334.1 Plasmodium vivax isolate 200606 mitochondrion, com...	196	8e-47
JQ240333.1 Plasmodium vivax isolate 200604 mitochondrion, com...	196	8e-47
JQ240332.1 Plasmodium vivax isolate 200603 mitochondrion, com...	196	8e-47
JQ240331.1 Plasmodium vivax isolate 200601 mitochondrion, com...	196	8e-47
KC330557.1 Plasmodium vivax isolate Lo40B cytochrome c oxidas...	196	8e-47
KC330554.1 Plasmodium vivax isolate Lo23D cytochrome c oxidas...	196	8e-47
KC330553.1 Plasmodium vivax isolate Lo23B cytochrome c oxidas...	196	8e-47
KC330550.1 Plasmodium vivax isolate Lo48A cytochrome c oxidas...	196	8e-47
KC330551.1 Plasmodium vivax isolate Lo5A cytochrome c oxidase...	196	8e-47
KC330549.1 Plasmodium vivax isolate Lo40A cytochrome c oxidas...	196	8e-47
KC330548.1 Plasmodium vivax isolate Lo8A cytochrome c oxidase...	196	8e-47
KC330547.1 Plasmodium vivax isolate Lo13C cytochrome c oxidas...	196	8e-47
KC330546.1 Plasmodium vivax isolate Lo13B cytochrome c oxidas...	196	8e-47
KC330545.1 Plasmodium vivax isolate Lo13A cytochrome c oxidas...	196	8e-47
KC330543.1 Plasmodium vivax isolate Lo1A cytochrome c oxidase...	196	8e-47
KC330542.1 Plasmodium vivax isolate Lo29C cytochrome c oxidas...	196	8e-47
KC330538.1 Plasmodium vivax isolate Lo5B cytochrome c oxidase...	196	8e-47
KC330537.1 Plasmodium vivax isolate Lo17B cytochrome c oxidas...	196	8e-47
KC330536.1 Plasmodium vivax isolate Lo17A cytochrome c oxidas...	196	8e-47
KC330535.1 Plasmodium vivax isolate Lo72A cytochrome c oxidas...	196	8e-47
KC330533.1 Plasmodium vivax isolate Lo1C cytochrome c oxidase...	196	8e-47
KC330527.1 Plasmodium vivax isolate Ca66AA cytochrome c oxida...	196	8e-47
KC330515.1 Plasmodium vivax isolate Ca60B cytochrome c oxidas...	196	8e-47
KC330513.1 Plasmodium vivax isolate Ko37A cytochrome c oxidas...	196	8e-47
KC330512.1 Plasmodium vivax isolate Ko40A cytochrome c oxidas...	196	8e-47

KC330511.1 Plasmodium vivax isolate Ko2A cytochrome c oxidase... 196 8e-47
KC330509.1 Plasmodium vivax isolate Ko35B cytochrome c oxidas... 196 8e-47
KC330508.1 Plasmodium vivax isolate Ko35A cytochrome c oxidas... 196 8e-47
KC330507.1 Plasmodium vivax isolate Ko28A cytochrome c oxidas... 196 8e-47
KC330505.1 Plasmodium vivax isolate Ko28C cytochrome c oxidas... 196 8e-47
KC330504.1 Plasmodium vivax isolate Ko9A cytochrome c oxidase... 196 8e-47
AB550280.1 Plasmodium vivax mitochondrial DNA, complete genom... 196 8e-47
AB550276.1 Plasmodium vivax mitochondrial DNA, complete genom... 196 8e-47
DQ396549.1 Plasmodium vivax isolate T9605 mitochondrion, comp... 196 8e-47
DQ396547.1 Plasmodium vivax isolate IZ01052 mitochondrion, co... 196 8e-47
AY598136.1 Plasmodium vivax isolate CX9 mitochondrion, comple... 196 8e-47
AY598135.1 Plasmodium vivax isolate CX8 mitochondrion, comple... 196 8e-47
AY598134.1 Plasmodium vivax isolate CX7 mitochondrion, comple... 196 8e-47
AY598129.1 Plasmodium vivax isolate CX2 mitochondrion, comple... 196 8e-47
AY598128.1 Plasmodium vivax isolate CX1 mitochondrion, comple... 196 8e-47
AY598121.1 Plasmodium vivax isolate VX3 mitochondrion, comple... 196 8e-47
AY598108.1 Plasmodium vivax isolate IL48 mitochondrion, compl... 196 8e-47
AY598106.1 Plasmodium vivax isolate IL45 mitochondrion, compl... 196 8e-47
AY598103.1 Plasmodium vivax isolate IBM6 mitochondrion, compl... 196 8e-47
AY598102.1 Plasmodium vivax isolate IBM5 mitochondrion, compl... 196 8e-47
AY598101.1 Plasmodium vivax isolate IBM2 mitochondrion, compl... 196 8e-47
AY598100.1 Plasmodium vivax isolate IBY7 mitochondrion, compl... 196 8e-47
AY598099.1 Plasmodium vivax isolate IBY5 mitochondrion, compl... 196 8e-47
AY598063.1 Plasmodium vivax isolate TFF13 mitochondrion, comp... 196 8e-47
AY598050.1 Plasmodium vivax isolate TC28 mitochondrion, compl... 196 8e-47
AY598039.1 Plasmodium vivax isolate T124 mitochondrion, compl... 196 8e-47
AY791690.1 Plasmodium vivax isolate pvChesson cytochrome c ox... 196 8e-47
AY791666.1 Plasmodium vivax isolate pv20131 cytochrome c oxid... 196 8e-47
AY791631.1 Plasmodium vivax isolate pv01006 cytochrome c oxid... 196 8e-47
AY791612.1 Plasmodium vivax isolate india.01018 cytochrome c ... 196 8e-47
AY791604.1 Plasmodium vivax isolate CN96 cytochrome c oxidase... 196 8e-47
AY791602.1 Plasmodium vivax isolate CN9 cytochrome c oxidase ... 196 8e-47
AY791597.1 Plasmodium vivax isolate CN78 cytochrome c oxidase... 196 8e-47
AY791596.1 Plasmodium vivax isolate CN76 cytochrome c oxidase... 196 8e-47
AY791595.1 Plasmodium vivax isolate CN75 cytochrome c oxidase... 196 8e-47
AY791593.1 Plasmodium vivax isolate CN5 cytochrome c oxidase ... 196 8e-47
AY791592.1 Plasmodium vivax isolate CN3 cytochrome c oxidase ... 196 8e-47
AY791590.1 Plasmodium vivax isolate CN12 cytochrome c oxidase... 196 8e-47
AY791589.1 Plasmodium vivax isolate CN10 cytochrome c oxidase... 196 8e-47
AY791588.1 Plasmodium vivax isolate CN1 cytochrome c oxidase ... 196 8e-47
AY791587.1 Plasmodium vivax isolate D33c cytochrome c oxidase... 196 8e-47
AY791586.1 Plasmodium vivax isolate D33b cytochrome c oxidase... 196 8e-47
AY791585.1 Plasmodium vivax isolate D33a cytochrome c oxidase... 196 8e-47
AY791583.1 Plasmodium vivax isolate Thai3 cytochrome c oxidas... 196 8e-47
AY791582.1 Plasmodium vivax isolate pv02119 cytochrome c oxid... 196 8e-47
AY791581.1 Plasmodium vivax isolate pv02087 cytochrome c oxid... 196 8e-47
AY791580.1 Plasmodium vivax isolate pv99189 cytochrome c oxid... 196 8e-47
AY791579.1 Plasmodium vivax isolate pv99174 cytochrome c oxid... 196 8e-47
AY791578.1 Plasmodium vivax isolate pv99173 cytochrome c oxid... 196 8e-47
AY791573.1 Plasmodium vivax isolate pv20196 cytochrome c oxid... 196 8e-47
AY791572.1 Plasmodium vivax isolate pv20167 cytochrome c oxid... 196 8e-47
AY791567.1 Plasmodium vivax isolate pv20041 cytochrome c oxid... 196 8e-47
AY791563.1 Plasmodium vivax isolate pv02011 cytochrome c oxid... 196 8e-47
AY791562.1 Plasmodium vivax isolate pv01157 cytochrome c oxid... 196 8e-47
AY791556.1 Plasmodium vivax isolate pvONG cytochrome c oxidas... 196 8e-47
AY791554.1 Plasmodium vivax isolate pvNorth Korean cytochrome... 196 8e-47
AY791552.1 Plasmodium vivax isolate Indonesia 14 cytochrome c... 196 8e-47
AY791528.1 Plasmodium vivax isolate pv02012 cytochrome c oxid... 196 8e-47
AY791523.1 Plasmodium vivax isolate pv205 cytochrome c oxidas... 196 8e-47
AY791522.1 Plasmodium vivax isolate pv12123 cytochrome c oxid... 196 8e-47
KY923424.1 Plasmodium vivax isolate 1089PNG mitochondrion, co... 191 4e-45
KY923423.1 Plasmodium vivax isolate 8006PNG mitochondrion, co... 191 4e-45

Supplementary Table 1. Primers designed using *PrimedRPA*

Underlined are the homologous regions between the primers designed with *PrimedRPA* and the previously published and manually designed primers; FP = Forward Primer, RP = Reverse Primer.

Organism	**Sequence**
***S. Pneumonia* FP**	5′-ACAGCTCCGTCTGTTATTTACAAAGTTAATTTGAC-3′
***S. Pneumonia* RP**	5′-AGTCCCCACGCTTACGCTGAGCTAGCTCCATTACT-3′
***S. Pneumonia* FP**	5′-TCTGTTATTTACAAAGTTAATTTGACCGACGG-3′
***S. Pneumonia* RP**	5′-TAGTCACAAAGTCCCCACGCTTACGCTGAGCT-3′
BEFV FP	5′- AGAGCTTGGTGTGAATACAGACCTTTTGTTGAC – 3′
BEFV RP	5′- TCGAATTTGATCAATTTTTGATAATCCTCTATC - 3′
BEFV FP	5′- AGCTTGGTGTGAATACAGACCTTTTGTTGACAAGAA -3′
BEFV RP	5′- CCTCGAATTTGATCAATTTTTGATAATCCTCTATCC -3′
***P. vivax* FP**	5′- CCTTACGTACTCTAGCTTTTAACACAATATTATTGTC-3′
***P. vivax* RP**	5′- ACAATATTATACTGGCATTTTGTTGAAATTATATGGT- 3′

PrimedRPA Architecture

PrimedRPA is a python-based command line tool compatible with both Linux and Macintosh operating systems. Following successful installation, the user can begin designing Recombinase Polymerase Amplification (RPA) primers / probes, guided by the workflow and parameters outlined in the following sections. PrimedRPA incorporates the following 3rd-party software, Clustal Omega [1], Blast [2] and Samtools [3].

Standard Workflow

The following section outlines an example PrimedRPA workflow for the design of a pan *Plasmodium spp* RPA assay, specifically targeting a region in the mitochondria which is conserved across all 6 human infecting *Plasmodium* parasites. The assay is intended to be

used in the screening of human whole blood samples.

1. First, we need to specify the target sequence / sequences for PrimedRPA-based RPA assay design. The mitochondrial genomic sequences for all 6 human infecting *Plasmodium spp* parasites should be sourced and combined to create a single multi-sequence fasta file. The path to this file will be specified by the InputFile parameter (See PrimedRPA Parameter Overview).

2. As the assay will be used to screen human whole blood samples, ensuring it has no cross-reactivity with human DNA or the DNA of any other blood-borne pathogens is essential. To achieve this, the human reference genome along with the genomes for any other blood-borne pathogens which may be present should be combined into a separate multi-sequence fasta file. The path to this file will be specified by the BackgroundCheck parameter (See PrimedRPA Parameter Overview).

3. As PrimedRPA's default settings are in line with TwistDX LTD assay recommendations, the design process can begin by simply specifying the RunID, InputFile and BackgroundCheck parameters.

4. As the goal is to design a single RPA assay which can target all 6 plasmodium species and the InputFile specified contains multiple sequences, Clustal Omega will be utilised to first align the target sequences provided. The alignment generated will subsequently be converted into an alignment summary, containing the conservation of each position, hereafter referred to as position identity, as well as the most common nucleotide.

5. Following creation of the alignment summary, potential oligo binding sites are identified and screened. Oligo binding sites variants are defined according to user-specified parameters, PrimerLength (Default = 30), ProbeLength (Default = 50). The screening process involves filtering potential binding site according to several thresholds defined in the parameter file, including:

    i) A minimum identity threshold, to ensure the oligo-binding site is conserved across all target sequences, provided in the InputFile

    ii) Desired GC% content, to ensure the oligo-binding site is within the TwistDX recommended range of 30-70%.

    iii) The absence of homopolymer nucleotide repeats, which could hinder assay specificity and contribute to oligo secondary structure.

    iv) A maximum dimerization threshold, to exclude individual oligos and sets of oligos with high likelihood of forming dimer-complexes which would impede assay sensitivity and potentially contribute to false positives.

    v) A maximum cross reactivity threshold, to ensure the oligo binding site is conserved within the target species and not present within the background sequences specified. To run the cross-reactivity check, BLAST and Samtools are utilised to identify and extract similar oligo binding sites which could lead to off-target binding.

If an oligo binding site passes all the above filters it is stored for downstream use. All passed oligo binding sites are exported into the Oligo Binding Sites file, which can be re-used in other analyses.

6. After the identification of candidate oligo binding sites, PrimedRPA then begins the process to identify oligo binding site combinations which satisfy the AmpliconSizeLimit parameter (Default = 500bp). For each combination, the respective oligo sequences are derived for the respective forward and reverse RPA primers and the dimerization potential assessed. Primer combinations which pass the dimerization potential are exported as a potential candidate set.

A full tutorial for the PrimedRPA software can be found at https://github.com/Matthew Higgins2017/bioconda-PrimedRPA/wiki

PrimedRPA Parameter Overview

PrimedRPA enables users to parse the parameters highlighted in **Supplementary Table 2**, through either the Command Line Interface (CLI) or a parameter file (https://github.com/Matthew Higgins2017/bioconda-PrimedRPA/blob/master/PrimedRPA_Parameters.txt)

Supplementary Table 2: Parameters required for PrimedRPA.

Parameter	**Description**	**Default**
RunID	The associated Run ID given to any analysis. This will be used as the prefix for the output files generated.	N/A

PriorAlign	Options: NO – Do not use a previously generated alignment file. <File Path> - Path to a previously generated alignment file.	NO
PriorBindingSite	Options: NO – Do not use a previously generated binding sites file. <File Path> - Path to a previously generated binding sites file.	NO
InputFile	The path to the fasta file containing target sequence / sequences.	N/A
InputFileType	Input fasta file type: SS - Single target sequence MS - Multiple target sequences (unaligned) MAS - Multiple targets sequences (aligned)	SS
IdentityThreshold	The binding site identity threshold. (Explained in more detail below)	99
ConservedAnchor	The number of nucleotides from the 3' primer terminus which require an 100% identity score.	3
PrimerLength	The desired primer length e.g., 30 or range e.g., 28-32	30
ProbeRequired	The options are as follows: NO - No probe required EXO - Exo probe required NFO - Nfo probe required	NO
ProbeLength	The desired probe length / range e.g., 50 or 45-55	50
AmpliconSizeLimit	The upper limit for the amplicon length	500

NucleotideRepeatLimit	The number of tolerated single nucleotide repeats	5
MinGC	Minimum GC Content %	30
MaxGC	Maximum GC Content %	70
DimerisationThresh	The number of sites in an oligo which could cause dimerization, relative to the sequence length, expressed as a percentage	40
BackgroundCheck	Options: NO - No cross-reactivity check required. <File Path> - Path to fasta file containing all the background sequences which will be checked against.	NO
CrossReactivityThresh	The threshold between a given binding site in the target and a similar binding site in potential background sequences provided. (Explained in more detail below).	65
MaxSets	The maximum number of primer-probe sets to identify and export.	100
Threads	The number of threads available.	1
BackgroundSearchSensitivity	An option to alter the Blastn settings which will impact the sensitivity and speed of the cross-reactivity search. Speed: Fast > Basic > Advanced. Sensitivity: Advanced > Basic > Fast (Explained in more detail below).	Basic

Parameter: IdentityThreshold

The purpose of the identity threshold is to identify oligo binding sites which are conserved across all target sequences if a multi-sequence fasta file is provided by the user. As outlined previously, if a multi-sequence fasta is used as the input file the percentage identity of each

position will be assessed and recorded as part of the alignment summary. When extracting and filtering potential oligo binding sites, the average identity of the binding site will be calculated based on the positions covered. For example, **Supplementary Table 3** covers a subsection of the alignment summary for the potential oligo binding site covering positions 1 to 13. When averaged this oligo binding site will have an identity score of 0.962, which is below the default Identity Threshold 0.99, and as such be excluded from downstream analysis.

Supplementary Table 3. Alignment Summary subsection.

Identity	**Alignment Position**	**Nucleotide**
1	1	A
1	2	C
1	3	G
1	4	A
0.75	5	A
1	6	A
1	7	A
1	8	T

1	9	A
0.75	10	T
1	11	A
1	12	G
1	13	G

Parameter: CrossReactivityThresh

If a cross reactivity check is required by the user, the background fasta file specified will be used to create a nucleotide BLAST database. Each oligo binding site will subsequently be screened against this database and a cross-reactivity score calculated for each BLAST hit as follows:

$$\text{Cross Reactivity Score (CRS)} = ((LA * (PI/100))/LQ) * 100$$

LA = Length of Oligo Binding Site Alignment

PI = Percentage Identity

LQ = Length of Oligo Binding Site

For a given oligo binding site, hits are subsequently ranked according to the CRS, with the maximum CRS used to determine if the oligo binding site is above the CrossReactivityThresh parameter. If the CRS is above the threshold the binding site is excluded from downstream analysis.

Parameter: DimerisationThresh

RPA reactions are susceptible to dimerization, due to the low reaction temperature and like other nucleic acid amplification technologies, dimerization is believed to impede assay sensitivity. PrimedRPA assesses the formation of dimers at two stages: 1) the ability of oligos to form self-dimers when assessing candidate oligo binding sites, 2) the ability of oligos within a potential RPA primer set to form dimers. The dimerization score is calculated as follows.

$$\text{Dimerization Score} = (CN * 100) / (CN + MN)$$

CN = Complementary nucleotides with dimerization complex

MN = Mismatched nucleotides within the dimerization complex.

For example, the oligo 5'- CGGTAAGAAATATTTCCAAAACTTAATACCGC - 3' can form the self-dimerization complex outlined in **Supplementary Figure 2**. This complex contains 20 and 12 complementary and mismatched nucleotides respectively. As such the dimerization score for this given oligonucleotide is calculated as 62.5% which is above the default dimerization threshold of 40% and would be excluded from downstream analysis.

```
 CGGTAAGAAATATTTCCAAAACTTAATACCGC
-|||||--||-|-||---||-|-||--|||||---
CGCCATAATTCAAAACCTTTATAAAGAATGGC
```

Supplementary Figure 2. Self-dimerization potential for an example oligo.

Parameter: BackgroundSearchSensitivity

When screening oligo-binding sites, a trade-off is made in the BLAST search between speed and sensitivity. **Supplementary Table 4** outlines the BLAST search parameters changed for each setting.

Supplementary Table 4: BackgroundSeachSensitivity Parameter Options and associated BLAST search parameters.

Option	**Word Size**	**Gap Open**	**Gap Extend**	**Reward**	**Penalty**
Fast	7	5	2	1	-3
Basic	4	5	2	1	-2
Advanced	4	5	2	1	-1

References

1. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
2. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–5 (2004).
3. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Chapter 3. Adapting RPA for Colorimetric End-Point Detection.

Chapter 3: Adapting RPA for Colorimetric End-Point Detection

Premise

RPA has been successfully deployed in the detection of several pathogens, including *Plasmodium falciparum* and *knowlesi* species [1–3]. However, to compete with the cost-effectiveness of existing antigen-based malaria rapid diagnostic tests (RDTs), which average £0.30 per test, further assay development is necessary to adapt the technology (**Table 1)**. A single TwistAmp Liquid Basic reaction costs £1.60 when used according to the manufacturer's protocol. The RPA reaction volume can be reduced 90%, from 50ul to 5ul, whilst maintaining performance, shrinking the cost per reaction (£0.16) [4]. However, additional costs for RPA reaction end-point detection have to be considered. RPA can be adapted for lateral flow (LF) based detection, using antigen labelled oligos, which has enabled the detection of both *P. knowlesi* and *P. falciparum* [1,5] malaria parasites. However, the addition of a LF cassette or dipstick costs ~£2, in addition to the inclusion of Endonuclease IV, which is no longer integrated into the TwistAmp kits available from TwistDX (since 2021). The inclusion of LF detection makes the cost of an RPA-based assay at least 7 times higher than existing RDTs. In addition, the use of the LF cassette adds another step to the diagnostic assay, introducing the risk of human error which could impede diagnostic efficacy. As such, I set about finding a cost-effective alternative for reaction-end point detection which would align the unit economics of an RPA-based diagnostic with existing malaria RDTs.

Following an assessment of the literature for other mechanisms of NAAT end-point detection, I decided to explore the feasibility of colorimetric-based approaches. By including a colorimetric component into the RPA reaction, a one-step assay could be created. The use of colorimetric components results in a dramatic cost reduction, is suitable for in-field non-

specialist use similar to LF-based assays, and has utility in a research setting by enabling high-throughput assessment when combined with a UV-Vis spectroscopy.

RDT	Per Assay Cost (£)
ParaHit Malaria Pf	0.16
First Response Malaria Pf HRP2	0.20
ParaHIT Malaria Pf cassette	0.16
AdvDx Malaria Pf	0.20
FirstResponse MalAgPf/Pv CardTest	0.28
FirstResponse Mal Agp LDH/HRP2 Combo	0.28
One Step MERISCREEN Mal Pf/Pv Ag	0.23
Bioline MalAg Pf	0.60
STANDARD Q Malaria Pf/Pan Ag Test	0.40
Bioline MalAg Pf/Pv	0.48

Table 1. Cost analysis of commercially available Rapid Diagnostic Tests (RDTs) according to UNICEF (https://supply.unicef.org/), assuming GBP to USD exchange rate of £1:$1.20.

Alongside my attempt to develop a colorimetric-based RPA assay, I sought to explore RPA's potential to move beyond conventional diagnostics and into the detection of clinically relevant biomarkers. As such, I decided to target key antimalarial resistance mutations in the *P. falciparum Kelch13* gene associated with Artemisinin resistance, specifically the non-synonymous mutations C580Y, I543Y and R539T [6]. In this pursuit I created plasmids containing the *PfKelch13* region of interest and generated artificially the variants of interest via site-directed mutagenesis. Subsequently, I designed primers to target the mutations of interest, through the deliberate introduction of mismatches into the primer-template complex to confer reaction specificity to the resistance genotype (**Supplementary Table 3)**. Through this approach I designed primers that could distinguish between resistance and wild-type genotypes. The distinction was valid between 1-100 pg of DNA for I543Y and R539T and

100fg - 100pg for C580Y, representing a target copy number (TCN) of 28,000 - 28,300,000 (**Figure 1)**.

Distinction between the wild type and resistance genotypes was not possible when >1ng of template DNA (283,000,000 TCN) was used. However, considering hyper-parasitemia is defined as 250,000 parasites per ul of blood in areas with a high malaria transmission [7], a TCN of 283,000,000 will not be reached in-field unless a >1000-fold concentration occurred in DNA material from the patient. As such, this should not impede assay performance in-field. Following the preliminary success of introducing primer-template mismatches to facilitate RPA-based genotyping, a full exploration is presented in **Chapter four**.

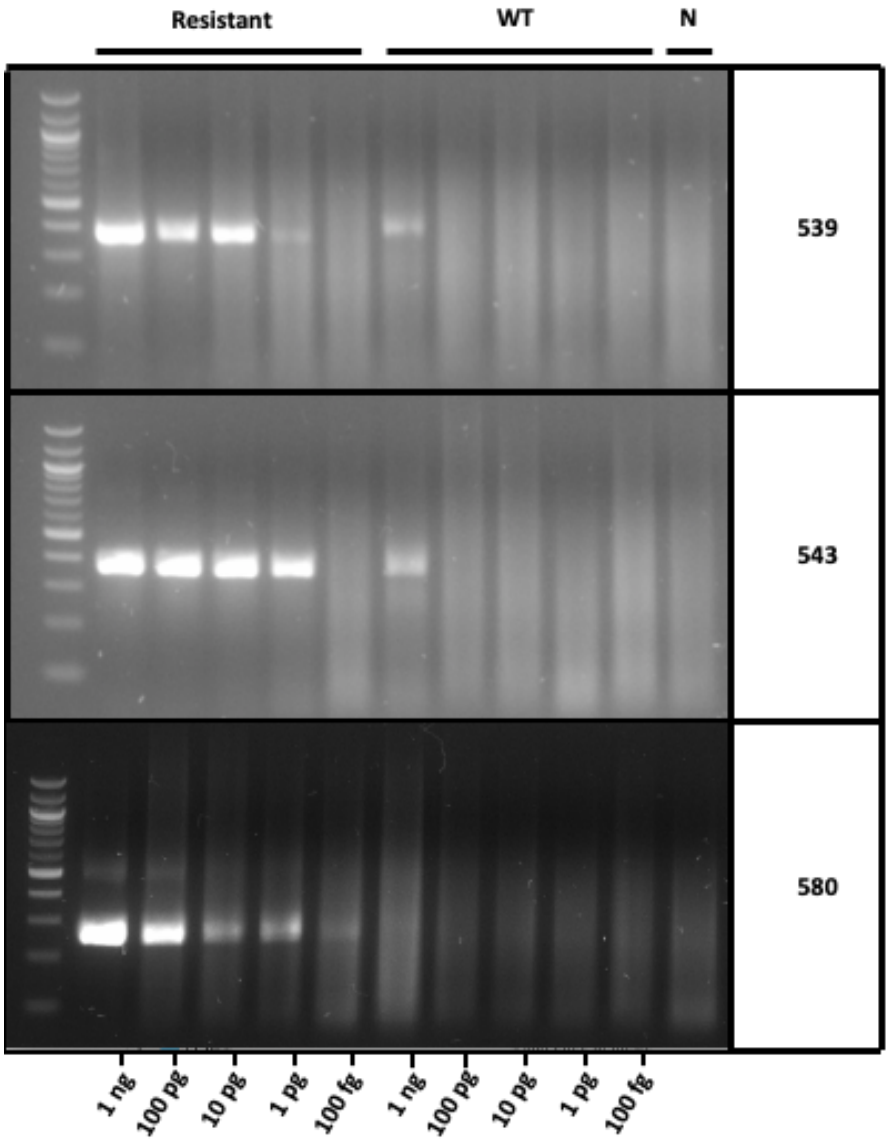


Figure 1. Assessing the feasibility of RPA-based SNP genotyping for three *PfKelch13* loci of interest (C580Y, I543Y and R539T), utilising Resistant and Wild Type *PfKelch13* synthetic constructs across a titration gradient.

Methods

Predicting *In-silico* the pH of a Solution at Equilibrium

The pHcalc python package (https://github.com/rnelsonchem/pHcalc) was used to predict the pH of defined solutions at equilibrium. Where available I utilised experimentally derived

acid-dissociation constants for compounds of interest, typically reported in product notes or associated literature. If unavailable, acid-dissociation constants were derived via the Marvin software, produced by ChemAxon. [8]

Recombinase Polymerase Amplification

Recombinase polymerase amplification was performed using the TwistDx Liquid Basic kit according to manufacturer's guidelines. Briefly each reaction consisted of 25ul reaction buffer, 3ul dNTPs, 6 ul water, 5ul of Primers, 2.5ul of 20x Enzyme Core Mix and 5ul of 10x Emix. Reactions were run for 30 minutes at 39°C using a G-Storm Thermocycler. Unless specified, 1ul of 1ng template was used in each reaction.

Polymerase Chain Reaction

Taq-based (NEB) - PCR was used during diagnostic assays such as colony screening. Briefly, reactions consisted of 6.5ul Taq2x Master Mix, 0.25ul of 100uM of each primer, 5ul of nuclease free water and 1ul of template, buffered in nuclease free water. Q5-based (NEB) PCR was used where the amplicon was required for downstream application such as plasmid sequencing. Briefly, reactions consisted of 5ul of 5x Q5 Reaction Buffer, 0.5ul 10mM dNTPs, 1.25ul of 100uM of each primer, 0.25 ul of Q5 High-Fidelity DNA Polymerase, 15.75 ul of nuclease free water and 1ul of template. Once set up, all PCR reactions were carried out in a G-Storm Thermocycler, following the recommended thermocycling conditions for each polymerase and establishing the primer annealing temperature according to the NEB Tm calculator (https://tmcalculator.neb.com/).

PfKelch13 Vector Design & Creation

I amplified a subsection of the *P. falciparum Kelch13* gene from reference *P. falciparum 3D7* lab strain DNA (**Supplementary Information)**. Following amplicon clean-up with QIAquick PCR Purification Kit (Qiagen), the insert was cloned into Pjet vector using the

CloneJET PCR Cloning Kit (ThermoFisher) according to the manufacturer's protocols. Following successful plasmid creation, I performed site-directed mutagenesis utilising the Q5® Site-Directed Mutagenesis Kit (NEB) according to manufacturer's protocols to create variant plasmids containing the SNP for C580Y, R539T and I543T. Associated primers are presented (**Supplementary Table 3)**. Successful variants were confirmed via Sanger-sequencing.

Recombinant Protein Vector Design

The Pet28a expression vector was selected due to harbouring kanamycin resistance for transformation screening and the option to introduce a N or C terminus Histidine tag for protein purification. Protein sequences for UvsX, UvsY, Gp32 and Bsu Polymerase were obtained from Uniprot and subsequently codon optimised for *E. coli*-based expression via GenSmart™ tool (GenScript). For Bsu polymerase the domain conferring 5' to 3' exonuclease activity was removed. NotI-HF and BamHI-HF target DNA sequences were manually added to the terminals of each insert alongside a smaller spacer region. The SnapGene software was then used to validate successful insert assembly *in-silico,* and subsequently each insert was ordered as a dsDNA gblock (ITD).

Recombinant Protein Vector Creation

Inserts of interest and the pet28a expression vector were digested with NotI-HF and BamHI-HF restriction enzymes (New England Biolabs) according to manufacturer's protocol in 10x Cut-Smart Buffer (NEB). DNA was subsequently cleaned using the QIAquick PCR Purification Kit (Qiagen). The inserts were subsequently ligated in the Pet28 backbone via T4 DNA ligase (NEB) under the default protocol, utilising a 3:1 insert to vector ratio as calculated by NEBioCalculator (https://nebiocalculator.neb.com /#!/ligation). Following ligation, XL10-Gold Ultracompetent (Agilent) *Escherichia coli* cells were transformed via

heat-shock. For all transformations, 2ul of ligation mix was added to 15ul of competent cells alliquotes. Post heat shock, the transformed competent cell mixture was spread onto kanamycin-treated (100 µg/mL) Luria-Bertani (LB) agar plate and incubated at 37ºC overnight. Following incubation, lone colonies were identified and subsequently screened via Taq PCR, utilising T7 forward and reverse primers covering the insert site. Successful ligations were determined by the presence of a lone PCR band per colony at the expected insert size. The respective colonies for successful ligations were incubated overnight at 37ºC in 5ml of 100 µg/mL kanamycin treated LB medium in a shaking incubator. Plasmids were subsequently isolated using QIAprep Spin Miniprep Kit (Qiagen) and quantified using the Nanodrop function on a DS-11 FX Spectrophotometer (DeNovix). Following plasmid isolation, sequence conservation was confirmed via sanger sequencing the insert region which was amplified via Q5 PCR using the T7 forward and reverse primers.

Recombinant Protein Expression & Purification

The BL21 (DE3) *E. coli* cell line was selected for recombinant protein expression. 30 ul allicotes of competent cells were transformed with 1 ul of plasmid product, following standard heat-shock protocol. The transformed cell mixtures were spread onto kanamycin-treated (100 µg/mL) Luria-Bertani (LB) agar plates and incubated at 37ºC overnight. Single colonies were subsequently selected and incubated in 10 ml culture tubes with 1ml of kanamycin treated (100 µg/mL) ZYP-50/52 autoinduction media, prepared according to authors guidelines including metallo-mix, at 20ºC for 30 hours on a shaker plate. Following successful incubation, the 500 ul of the culture was extracted and pelted by centrifugation, via a bench-top Eppendorf Centrifuge 5425 running at full speed. Supernatant was discarded and the pellet was resuspended in a lysis buffer (0.04 M Sodium Phosphate, 0.6 M NaCl, pH 7.4) and subsequently sonicated. Following sonication, the mixture was centrifuged again, and protein supernatant extracted. The supernatant was subsequently added to 1x Lemelli

buffer containing 350 mM DTT. The resulting mix was boiled for 30 minutes at 100ºC using a heating block. The presence of the recombinant protein in the boiled soluble fraction was determined via SDS-Page, using a Mini-PROTEAN electrophoresis system (Bio Rad - 1658005EDU) with 4–15% Mini-PROTEAN™ TGX Stain-Free™ Protein Gels (Bio-Rad 4568086) alongside a PageRuler Plus protein ladder (ThermoFisher). The presence of recombinant protein was determined as a strong protein band at the expected position in accordance with the PageRuler Plus protein ladder included.

Following confirmation of strong recombinant protein expression band and its presence in the soluble fraction, the remaining 500ul of the culture was used to inoculate 200ml of kanamycin treated ZYP-50/52 autoinduction media in a baffled 1L flask which was subsequently incubated again for 30 hrs at 20ºC in a shaking incubator. Following successful incubation, cultures were pelleted in a BeckMan Coulter centrifuge spinning for 30 minutes at 11,440 rcf. The cell pellets were subsequently extracted and resuspended in a lysis buffer, 5ml for each gram of pellet. The LM20 microfluidizer french press was then used for cell lysis. The soluble fraction of the lysed mixture was extracted through centrifugation, 30 minutes at 35,200 rcf, and the supernatant placed in a 50ml Falcon Tube.

DNA was subsequently removed from the supernatant via precipitation with polyethylenimine (PEI) [9,10]. 30 ml of supernatant was mixed with 6 ml of 5 NaCl, 1.17 ml of 5% PEI solution (pH 7.4) and 1.83ml of water creating a final 39ml solution in a 50ml Falcon tube containing 0.15% PEI and 1.23M NaCl, conditions under which DNA precipitates. The precipitated DNA was pelleted by centrifuging the tubes using an Eppendorf 5810R centrifuge. The DNA-free supernatant was subsequently transferred to a fresh 50ml Falcon tube. Next the proteins were precipitated through the addition of 16.48g of Ammonium Sulphate, and topped up with lysis buffer, creating an 80% ammonium sulphate 40

ml solution at which the recombinant protein of interest will precipitate. The protein precipitant was pelletized using an Eppendorf 5810R centrifuge, supernatant discarded, and the pellet resuspended in a fresh binding buffer (1% Triton X-100%, 1M NaCl, 20mM Sodium Phosphate, pH 7.4). The successful removal of dsDNA from the protein supernatant mix was checked via the Nanodrop function on a DS-11 FX Spectrophotometer (DeNovix).

To purify the recombinant protein of interest, NEBExpress® Ni resin beads were used according to the manufacturer's protocol. 4 ml of the resuspended protein in the binding buffer was applied and mixed with 1 ml of the nickel purification beads, in an 15ml falcon tube. The tubes were centrifuged for 60 seconds in an Eppendorf 5810R centrifuge. The beads were subsequently washed twice with 2 ml of wash buffer (1% Triton X-100%, 1M NaCl, 20mM Sodium Phosphate, 40mM Imidazole, pH 7.4), through application, mixing for 30 seconds and centrifugation. The recombinant protein of interest was then collected via 3 elution fractions were subsequently collected through the application of the elution buffer (1% Triton X-100%, 1M NaCl, 20mM Sodium Phosphate, 500 mM Imidazole, pH 7.4) to the beads, mixing and centrifugation. The purity of the recombinant protein in each elution fraction was subsequently checked via SDS-PAGE, as defined previously, and quantified via Quick-Start Bradford assay (Bio-Rad), according to manufacturer's protocols.

Buffer exchange was subsequently performed using a PD-10 Desalting column 8.3ml of sephadex G-25 Medium, (GE Healthcare) according to the manufacturer's gravity-based protocol, with an equilibrium buffer composed of 200uM Tris, 500 mM Potassium Acetate, 2mM DTT, pH 7.9. Bradford assay was used to determine which fraction the recombinant protein of interest eluted into. Fractions containing the recombinant protein of interest were subsequently simultaneously applied to Amicon® Ultra-15 Centrifugal Filter Units (Merck -

UFC900308) utilised with the centrifugal concentration protocol as per product supporting documentation.

SYBR Green I Assessment for End-Point Detection

SYBR Green Introduction

SYBR green I (SG) can be combined with RPA for naked-eye colorimetric distinguishment of the reaction endpoint and has been successfully applied in the detection of *P. knowlesi* and *Mycobacterium tuberculosis* [11,12]. SG is a fluorescent nucleic-acid binding dye and since its introduction in the early 1990s, has been regularly used in tandem with nucleic acid amplification techniques (NAATs), such as real-time PCR [13]. Upon binding to dsDNA, the fluorescence emission of SG is enhanced, compared to its free-state. This gain in fluorescence has been historically exploited to track the concentration of dsDNA during amplification. However, if used at a high concentration it can facilitate an orange to green dichromatic colorimetric transition, assuming dsDNA is in excess and all SG transitions from an unbound to bound state. To note, if a mix of bound and unbound SG is present, due to all binding sites within dsDNA being saturated, the solution will appear yellow.

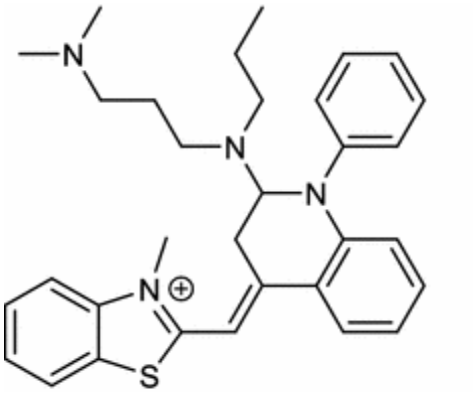


Figure 2. Chemical structure of SYBR green I (SG) with maximum excitation wavelength of 497 nm (blue) and emission wavelength is 520 nm (green).

Whilst SG is compatible with NAATs when utilised at low concentrations, the higher concentrations necessary for naked-eye visualisation have been reported to inhibit amplification [14]. As such when combined with RPA, SG forms a 2-step assay with SG being added by the practitioner on reaction completion. Under successful reaction conditions, the solution is expected to turn yellow to green upon SG addition due to the high concentration of dsDNA generated during amplification. However, if amplification was unsuccessful the solution should remain orange.

SYBR Green Results

Firstly, I sought to validate SG use with RPA in the detection of C580Y and R539T *PfKelch13* WT genotypes. Successful amplification was obtained for both targets, across a template titration range of 4ng, 40pg and 400fg (**Figure 3)**.

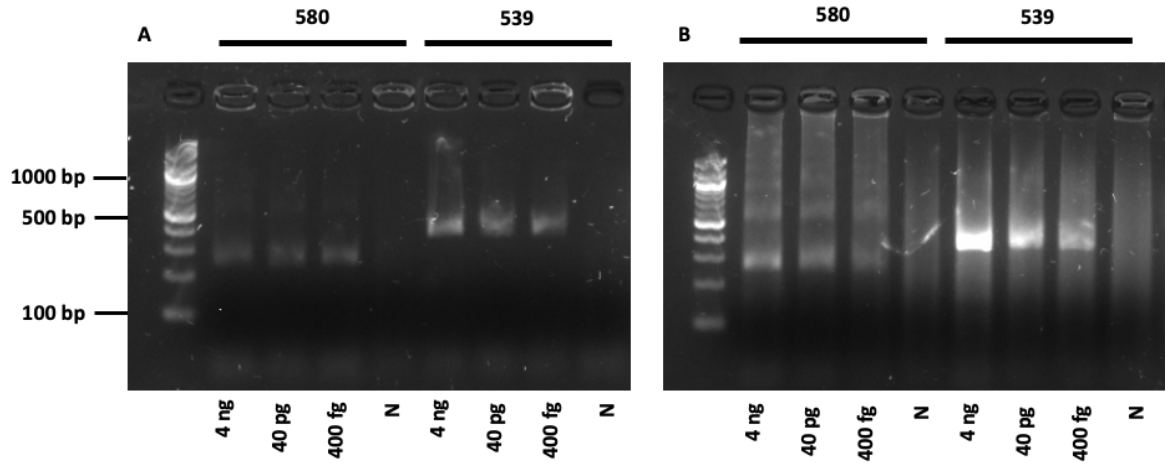


Figure 3. Amplification of C580Y and R539T *PfKelch13* wild type genotypes when running the RPA reaction for (A) 15 minutes and (B) 30 minutes, for use in SYBR green I (SG) validation assay.

Both reaction times of 15 and 30 minutes were suitable to observe successful amplification. It is clear from the intensity of the bands that the amplicon yield is higher for 30-minute reactions compared to 15 minutes as expected, particularly for the R539T WT genotype. A high level of smearing was observed when the reaction was run for 30 minutes and was present in both the positive and negative controls. Such smearing is associated with non-specific RPA amplification driven by primer secondary structure or dimerisation which is common for RPA reactions due to the low reaction temperature. For the C580Y target, a ladder pattern appears when running the reaction for 30 minutes compared to 15 minutes. This pattern is derived from the formation of amplicon secondary structure, facilitating amplicon self-priming and subsequent extension. Upon reaction completion 15 ul of each reaction was mixed with SG. Inline, with previous reports SYBR green nucleic acid gel stain was used whose stock concentration is 10,000x [11,12]. (**Figure 4; Figure 5**).

Subtle differences in end point coloration (orange vs. green) were observed for both targets under different reaction times and SG concentrations (**Figure 4; Figure 5)**. When targeting

the *PfKelch13* C580Y WT genotype, the 30-minute reaction combined with 24x SG produced a subtle difference in end-point coloration with the negative control having an orange tinge compared to the 3 target reactions which possessed a green tinge. On close inspection the 15-minute reaction with 24x SG produced a borderline end-point coloration difference, with the reactions corresponding to 4ng and 40pg of template appearing more golden compared to the negative control and 400fg reactions which appeared orange. The use of 24x SG resulted in a low colour intensity compared to the other SG concentrations, making it difficult to distinguish the borderline colour difference between the positive and negative controls. However, when increasing the SG contraction above 24x the colour intensity does increase, but no clear distinction could be made between the positive and negative controls for the C580Y WT genotype reactions. This uniform yellowish coloration obtained when using 44x, 63x and 79x SG is driven by the excess ubound SG. In comparison, for the 30-minute R539T WT genotype reactions, differences in colour were achieved between the negative control (orange) and positive reactions (green / gold) when utilising 44x, 63x and 79x SG. Unlike C580Y WT genotyping reactions, the colour distinction could be achieved at higher SG concentrations, due to the higher amplicon yield, as highlighted in **(Figure 3)**, minimising the concentration of unbound SG. Knowing that a end-point colour differential could be distinguished between a positive reaction and negative control in certain cases, I sought to explore if SG could be included into the reaction master mix to create a one-step assay. The use of SG at a concentration of 24x and 79x inhibited amplification of both *PfKelch13 WT* genotypes and no difference in colour was detected (**Figure 6; Figure 7)**.

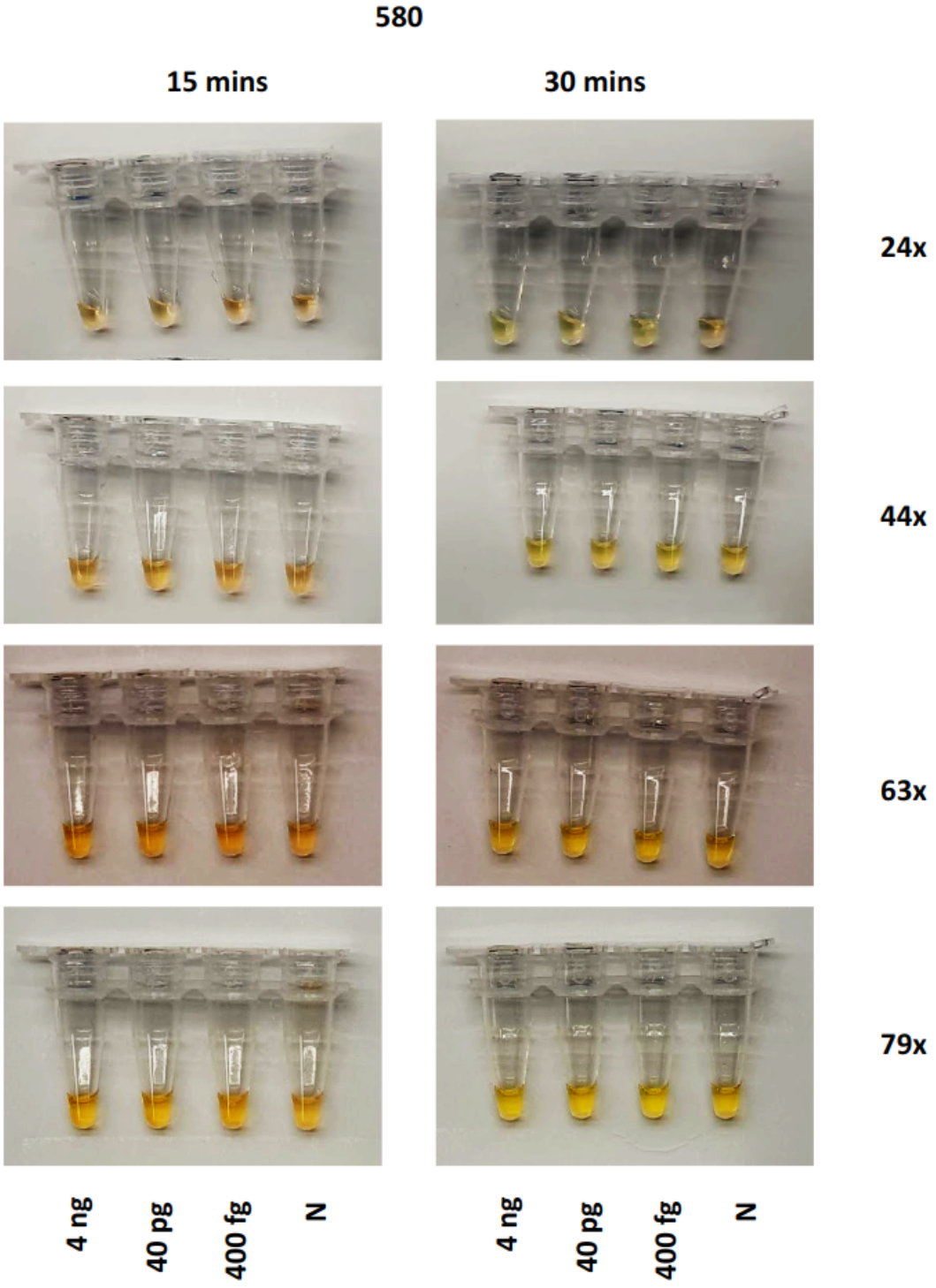


Figure 4. Evaluating the SYBR green I (SG) concentration necessary for reaction end-point visualisation when targeting the *PfKelch13* C580Y WT genotype. RPA reactions were run for 15 or 30 minutes. (N) negative control.

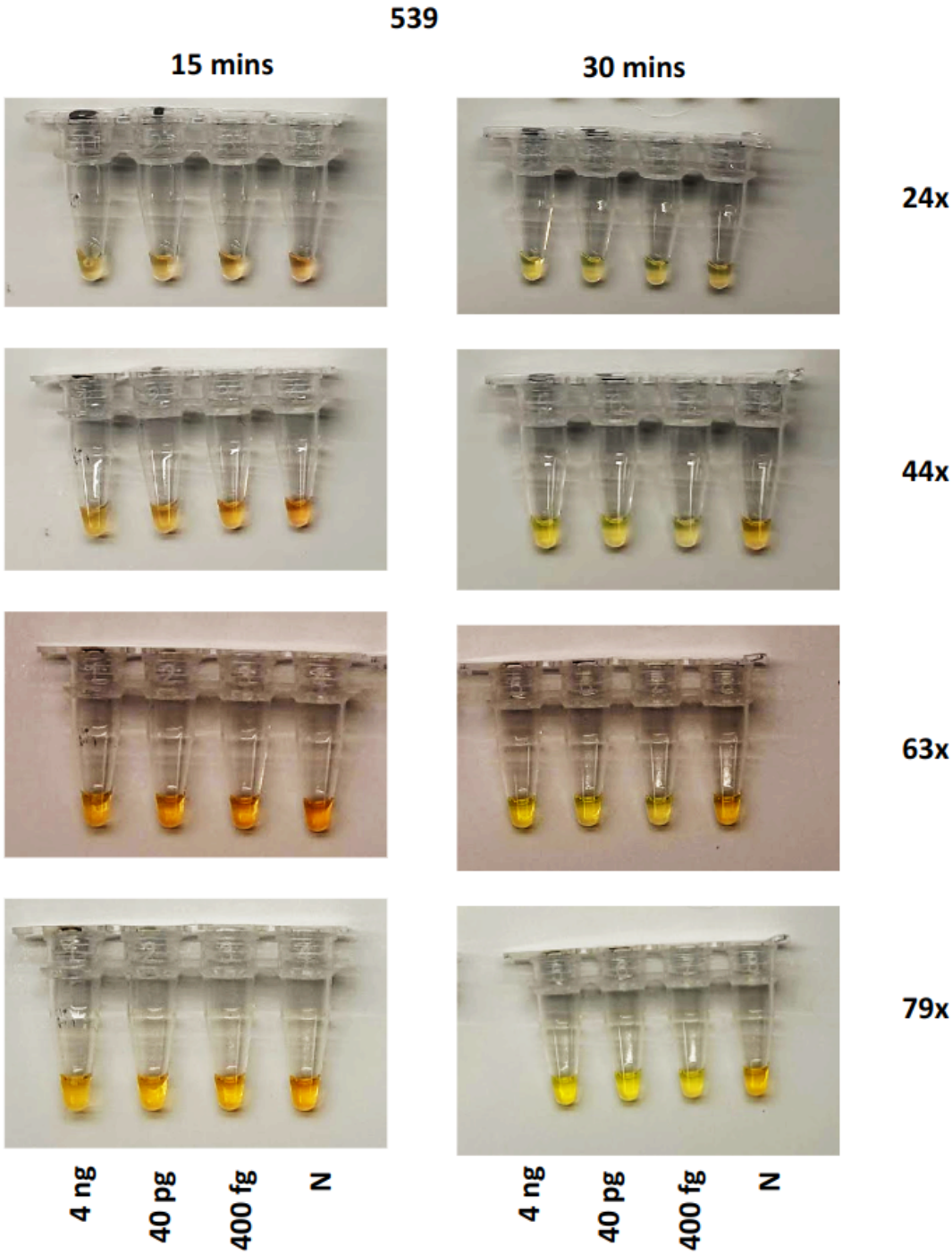


Figure 5. Evaluating the SYBR green I (SG) concentration necessary for reaction end-point visualisation when targeting the *PfKelch13* R539T WT genotype. RPA reactions were run for 15 or 30 minutes. (N) negative control

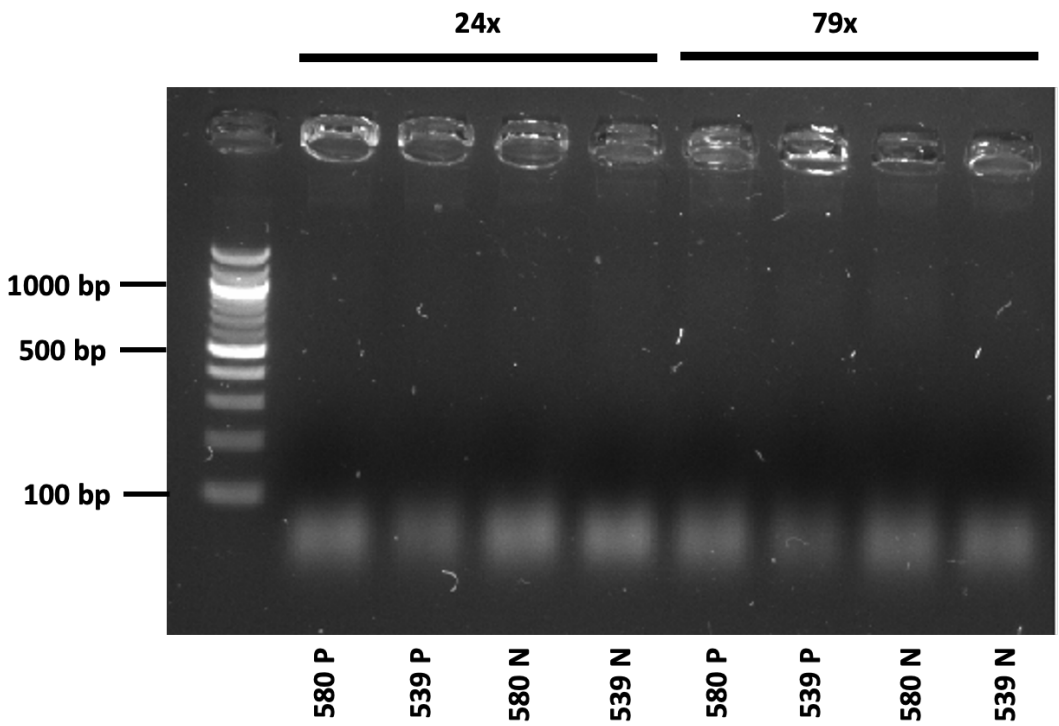


Figure 6. Assess the viability of incorporating SYBR green I (SG) to form a one-step RPA assay. Targeting the three *PfKelch13* WT genotypes of interest R539T, C580Y and I543Y.

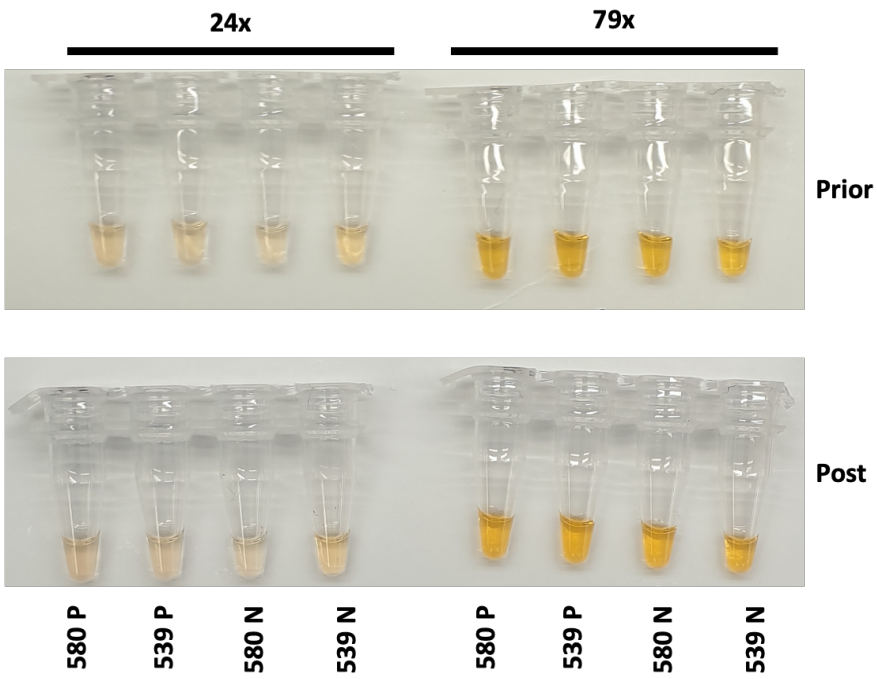


Figure 7. Assessing the colorimetric transition when incorporating SYBR green I (SG) to create a one-step RPA assay, when targeting the 3 *PfKelch13* WT genotypes of interest R539T, C580Y and I543Y prior reaction initiation and post reaction completion.

SYBR Green Discussion

Under certain experimental conditions, I was able to recreate the orange versus green SG driven reaction end-point distinction for both the detection of C580Y and R539T WT genotypes. Previous reports indicated that the optimal SG concentration for the detection of *P.knowlesi a*nd *Mycobacterium tuberculosis* was 7.4x and 14.4x, respectively. In our investigation we found it to be between 24x - 79x considering the target of interest and when running the reaction for 30 minutes. Theoretically, the use of SG at a final 24x concentration represents a 23,220x fold reduction in end-point detection costs compared to LF methods with an estimated cost of £0.92 per 10,000 assays, aligning a potential RPA-SG based diagnostic with existing RDTs.

However, our investigation highlights there are several limitations to the use of SG and a generalised one solution fits all approach, regarding RPA reaction time and concentration of SG, is not achievable. Target specific optimisation is required, to account for differences in assay performance and amplicon yield. Theoretically, the ideal SG concentration would be one that is high enough to produce a distinct polychromatic colour change that was easily distinguishable, upon successful amplification, and did not appear borderline like the 30-minute reaction of C580Y with 24x SG. However, the level of amplification achieved needs to yield the required amount of dsDNA such that all SG is bound. In turn, achieving the required yield has to be balanced with non-specific / background amplification which could reduce the green versus orange distinction between positive and negative control reactions. It is known that RPA reaction efficiency is non-uniform and tied to several factors including but not limited to, primer specificity, amplicon size and the primers / amplicon's ability to form secondary structures [15]. As expected over a fixed timeframe, differences in RPA reaction efficiency will result in a different dsDNA yield and as such the optimal concentration of SG will differ. When comparing our two targets, it appears the reaction for the R539T WT genotype is more efficient than that for C580Y WT genotype, resulting in a higher amplicon yield **(Figure 3)**. This would explain the higher concentration of SG being optimal for R539T due to achieving a higher overall dsDNA yield. In comparison the C580Y WT reaction had a lower dsDNA yield, resulting in a higher concentration of unbound SG, which in turn causes uniform coloration across both positive and negative controls, when using SG concentrations of 44, 63 or 79x. In addition, coloration optimization may not be possible for certain targets as there are variables we cannot control. For example, when using a lower target template concentration of 400fg and running the reaction for 15 minutes, across all SYBR green concentrations for the C580Y WT target, the positive reaction was indistinguishable from the negative reaction. Whilst this difference was mitigated by running the reaction for longer up to 30 minutes, one could assume

that such mitigation may not work for lower template concentrations. If this was combined with a high-level of non-specific amplification from primer-artefacts, then a SG-based distinction between positive and negative reactions would be impossible.

From a practical perspective, as target specific optimisation would be required for every SG-RPA assay, making it impossible to create a generalised reagent mix or protocol for any target like the existing TwistDX assays, the commercial viability is limited. In addition, as SG cannot be incorporated into a one-step assay and its polychromatic colour change is difficult to detect by the naked eye due to the solution remaining translucent with a low colour saturation when used at low concentrations, which are sometimes necessary to make distinction between positive and negative controls as demonstrated, I decided to explore other colorimetric approaches.

Malachite Green End Point Detection

Malachite green (MG) is another compound which has historically been combined with NAATs to facilitate reaction end-point colorimetric detection, as demonstrated by Lucchi et al, with LAMP-based detection of *Plasmodium spp* [16]. In this previous study, MG was included in the reaction master-mix making it a one-step assay, unlike SG, and upon successful amplification a colour transition from colourless to blue was reported. Like SG, successful incorporation of MG for RPA reaction end point detection would significantly reduce assay costs, £31.10 per 10,000 assays. MG was first synthesised in 1877 by Hermann Fischer and five chemical species of MG are known to exist (**Figure 8)** [17]. MG chemical species 1 is dominant when the pH is below 2.5, chemical species 5 is dominant between pH 2.5-7, and chemical species 3 is dominant when the pH is above 7 (**Figure 8)**[18].

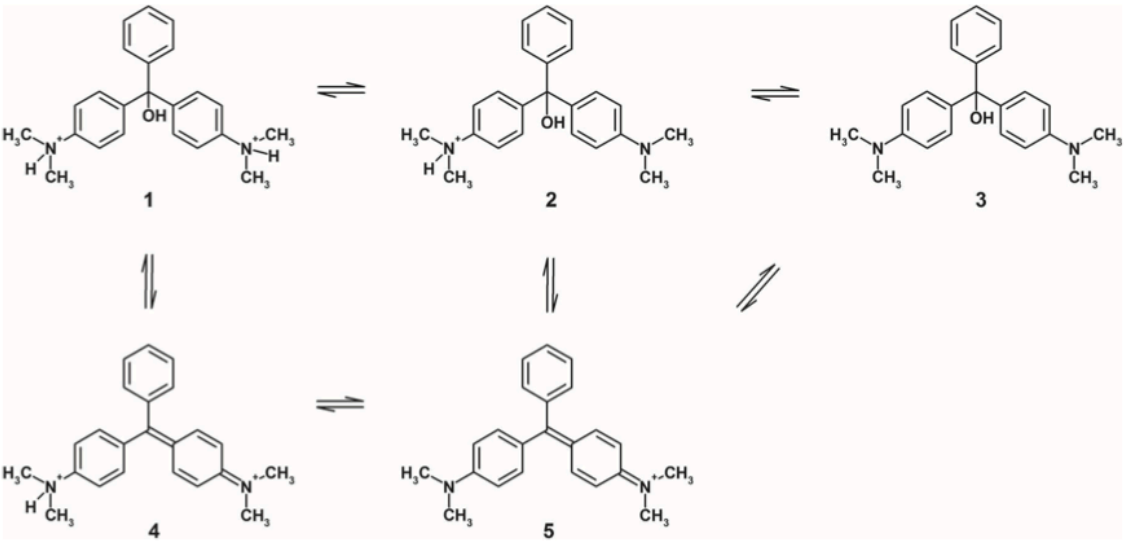


Figure 8. The 5 different chemical species of Malachite Green (MG). Chemical species 1-3 are colourless, species 4 is yellow and species 5 is blue [19]

First, I sought to determine if the RPA reaction could tolerate MG when targeting the *PfKelch13* R539T WT genotype; as to the best of my knowledge no previous attempt had been made to combine MG with RPA. MG was well tolerated by the TwistAmp Liquid Basic RPA reaction even when the core enzyme mix was diluted 5-fold (**Figure 9**). However, no change in colour was observed between positive and negative controls with all reactions remaining colourless (**Figure 9)**.

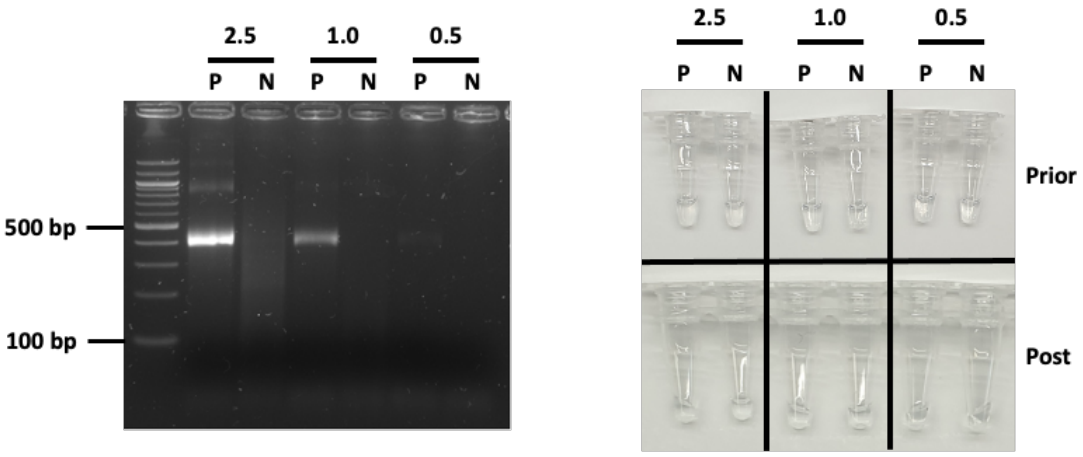


Figure 9. Assessing the tolerance of the *PfKelch13* R539T WT targeting RPA reaction to 0.004% Malachite Green (MG), when using the TwistDX Liquid Basic Kit recommend (2.5 µl) and reduced volumes (1 µl, 0.5 µl) of enzyme core mix and colour prior and post amplification.

Lucchi *et al* did not indicate the mechanism which drives the colour transition of MG upon successful amplification [16]. However, when diving into their published methodology we can rule out a change in pH as their reaction system was buffered at pH 8.8 with 40 mM Tris-HCL, which is 1500x fold higher than reported minimum buffer loop-mediated isothermal amplification systems required for pH change [20]. Nzelu et al [21] state MG is an DNA intercalating dye, and the ability of MG to form a complex with dsDNA has been confirmed within the literature [22]. However, the UV-Vis absorbance spectra of MG does not change significantly when intercalated with dsDNA, as the maximum absorption peak transitions from 616nm to 626nm, which would not result in the expected colourless to blue colour change [22]. Similar to SG, an increase in fluorescence emission of MG has been reported when bound to specifically designed nucleic acid aptamers, however the fluorescence emission peak is at ~650 nm which would result in a red coloration which again does not represent the colour change reported [23,24]. As such, I ruled out the interaction of MG with the

dsDNA produced during amplification as the cause of the colour transition. Other works which incorporated MG for endpoint NAAT-based amplification detection suggested that the colour change was from the detection of pyrophosphate, which would be released during amplification by the incorporation of dNTPs [25]. MG has long been used to detect pyrophosphate, but this has always been in combination with ammonium molybdate, which is absent in all reported MG-associated NAATs. In addition, for MG-ammonium molybdate assays the colour change transitions from yellow/green to green/blue which is not the reported by Lucchi et al [26]. Due to the lack of clarity regarding how MG-drives a colour transition upon amplification, I sought to elucidate the mechanism behind the transition and subsequently adapt the RPA reaction to facilitate this.

Interestingly, upon addition of MG to the RPA reaction, in the absence of DNA, the solution turned colourless instantly which was unexpected at pH 7.9, the pH of the RPA reaction, MG chemical species 5 **(Figure 8)** should be dominant, resulting in a blue tinge [18]. Therefore, I sought to identify if a component / components of the RPA reaction were responsible for inducing an initial colour transformation, prior to amplification, through the addition of MG to each component of the RPA TwistAmp Liquid Basic reaction (**Figure 10)**. To ensure the pH would remain fixed and not affect the colour change, each component was assessed in the presence of the 2x Rehydration Buffer. It was clear that only the E-mix resulted in the blue to colourless transition (**Figure 10**). The components of the Emix which are ATP, phosphocreatine and DTT were subsequently individually assessed (**Figure 10).** Both ATP and phosphocreatine appeared blue, however DTT induced the colour transition from blue to colourless. DTT is commonly used as an enzyme stabilising agent, preventing disulfide bond formation. Whilst DTTs presence is not mentioned directly by Lucchi et al [16] in their methods section, it is present in the Bst polymerase NEB product used. Therefore, the MG colour transition reported must be 2-step process, with DTT first interacting with MG upon reaction

composition, causing the blue to colourless transition; however this is subsequently reversed upon successful amplification.

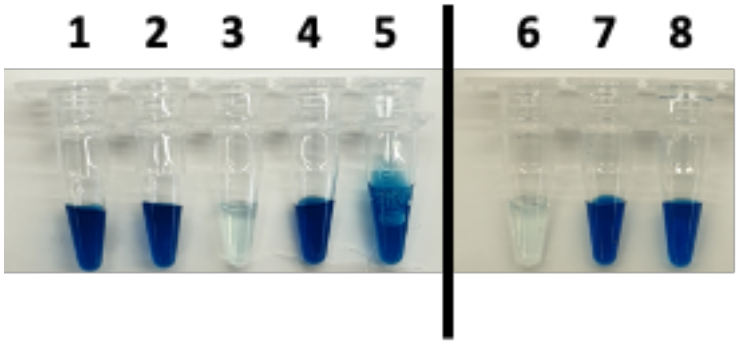


Figure 10. Deciphering which component or components of the TwistAmp Liquid Basic reaction are responsible for the initial Malachite Green (MG) colour change from blue to colourless. 1) 1x Rehydration Buffer (RB), 2) RB & 14mM MgOAc, 3) RB & 1x Emix, 4) RB & 0.6mM dNTPs, 5) RB & 1x Core Mix, 6) RB & 2mM DTT, 7) RB & 50mM Phosphocreatine, 8) RB & 3mM ATP.

After identifying that DTT was the cause of the initial colour change, I sought to create an in-house E-mix so that I could reduce the concentration of DTT, in an attempt to allow a colour transition to occur on successful amplification. When re-running the RPA assay using an in-house Emix (6 mM ATP, 50 mM dNTPs and 0.11 DTT mM) with 0.004% MG, no colour distinction was observed between the positive and negative controls (**Figure 11)**. However, upon titrating the reaction with MG, I was able to induce a subtle monochromatic colour change and see a difference between successful amplification (intense teal) and negative control (faint teal), when the MG concentration was increased to 0.012% (**Figure 11).**

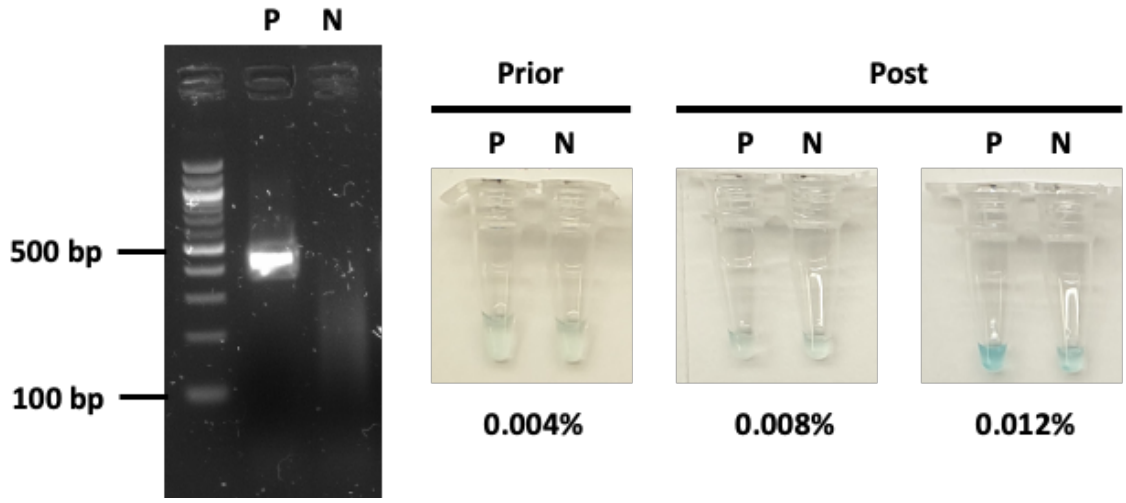


Figure 11. Optimising the Emix reagent in the TwistAmp Liquid Basic Kit and Malachite Green (MG) concentration to obtain a dichromatic colour transition for the RPA reaction targeting the *PfKelch13* R539T WT genotype. P and N correspond to the use of WT plasmid and water as samples, respectively.

Next, I sought to increase the starting concentration of MG in a one-step RPA assay to facilitate a significant colour transition (**Figure 12)**. When the MG concentration was increased above 0.012%, the RPA reaction was inhibited. In comparison when using a MG concentration of 0.008%, amplification was successful and a subtle monochromatic colour difference between the positive and negative reactions was observed at both 15 and 25 minutes, whereby the positive reaction teal colouration was borderline more saturated than the negative reaction. This subtle monochromatic colour transition is not ideal for nake-eye detection and further optimisation is required to make the transition polychromatic to ensure that the negative control remains colourless and only the positive transitions to teal upon successful amplification. As such I sought to elucidate the mechanism behind why successful amplification reverses the MG transition causing a colour change from colourless to blue as reported by Lucchi et al [16].

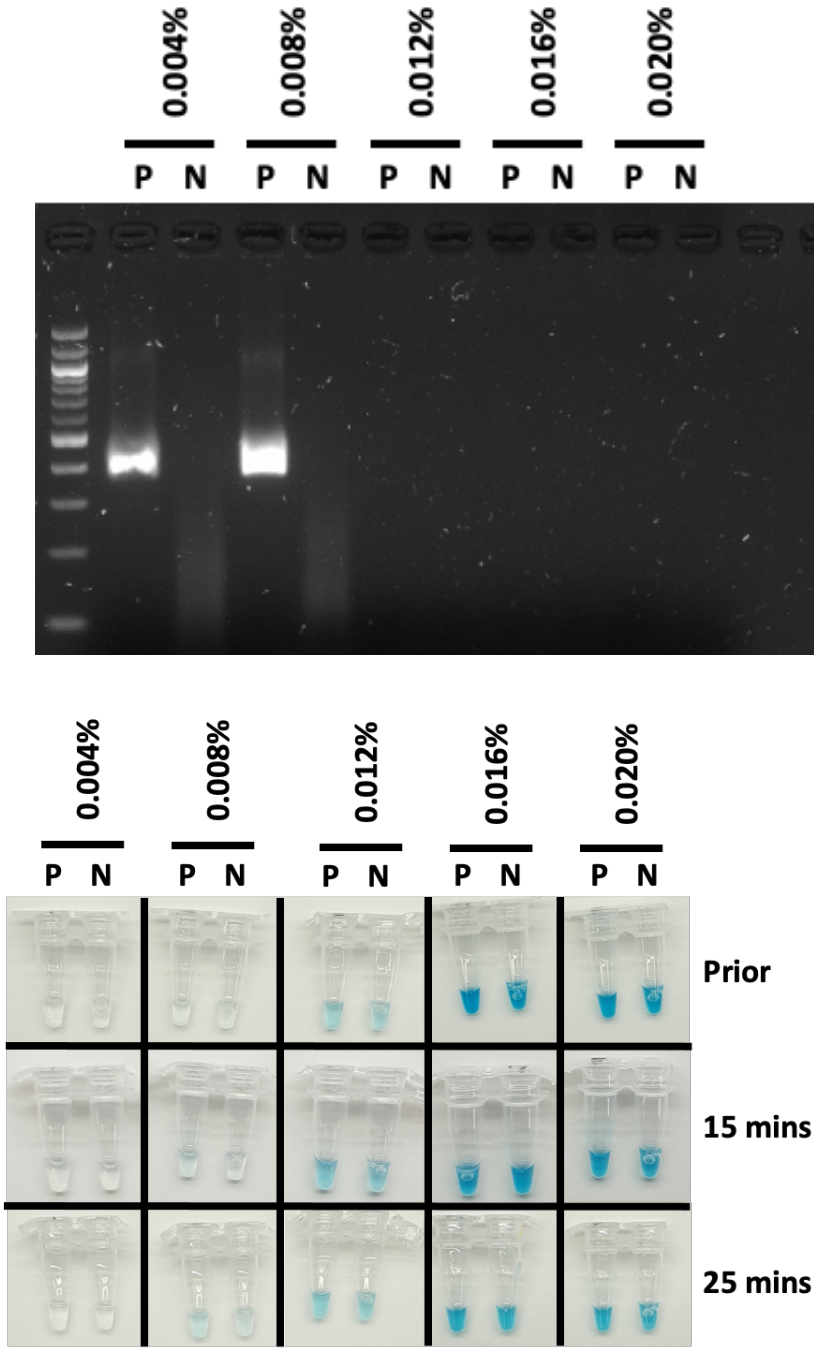


Figure 12. Tolerance and colour transition obtained for the *PfKelch13* R539T WT genotype targeting RPA reaction when increasing Malachite Green (MG) concentration. The modified TwistDX Liquid Basic Kit Emix was utilised, containing 0.11mM Dithiothreitol (DTT). The reaction mix was imaged at 15 and 25 minutes to determine if a colour transition had been obtained.

I hypothesised that DTT was responsible for the reduction of MG chemical species 5 (blue) to form species 3 (colourless) and that the transition from chemical species 3 back to species 5 was through the catalysed hydrolysis of pyrophosphate (**Figure 13)**. During successful amplification, dNTPs are incorporated into the growing DNA chain releasing pyrophosphate as a byproduct. Pyrophosphate typically undergoes hydrolysis at pH 8 to form phosphate [27] and I hypothesised that chemical species 3 of MG catalyses the hydrolysis and in doing so reforms species 5 (**Figure 13)**. Therefore, if amplification is unsuccessful, no pyrophosphate is released and subsequently the colourless chemical species 3 of MG remains dominant. In addition, if DTT is in excess compared to MG (see **Figure 9),** any species 5 of MG formed would be instantly reduced and shift the equilibrium towards the formation of the colourless species 3 of MG so that no blue colour is observed at all.

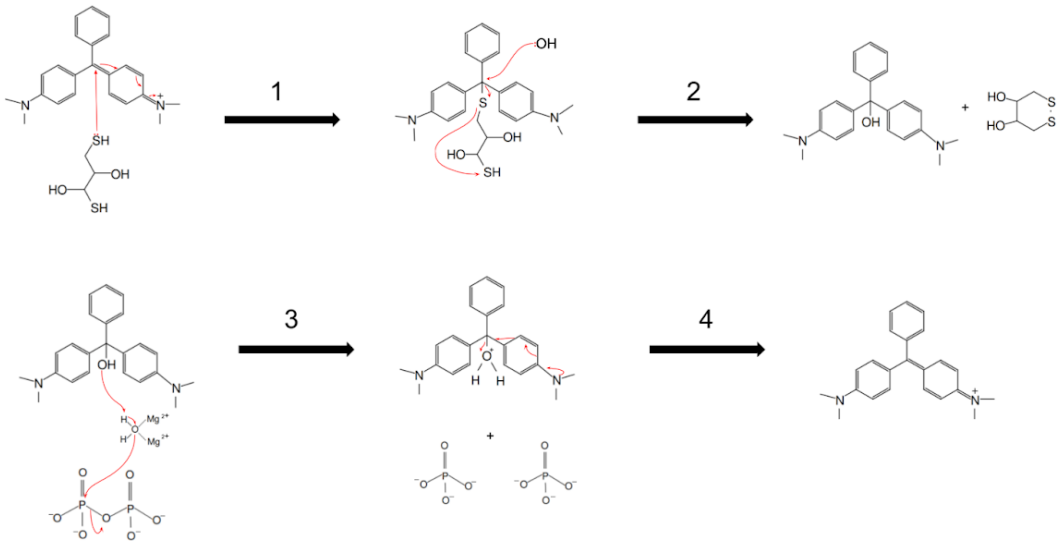


Figure 13. Hypothesised mechanism of Malachite Green (MG) use as a nucleic acid amplification technology (NAAT) end-point indicator. Steps 1 and 2 occur on reaction composition, causing initial blue to colourless dichromatic colorimetric change. Steps 3 and 4 occur upon successful amplification, whereby pyrophosphate becomes available and increases in concentration, causing colourless to blue dichromatic colorimetric change. Red arrows indicate the movement of electrons during the reaction.

To test this hypothesis, I synthetically recreated the end point of the MG reaction, with overlapping components from the RPA TwistDx Liquid kit and published, MG-LAMP assay [16]. This includes the 2x Reaction Buffer, Magnesium Acetate, DTT and MG. I then added pyrophosphate (0.6 mM) and dsDNA (40ng/ul) to artificially represent the endpoint of an RPA reaction. No change in colour was observed even after incubating the reaction mix for 60 minutes at 39°C across all combinations, including in the presence of pyrophosphate or dsDNA (**Figure 14**). This indicates that pyrophosphate hydrolysis may not drive the colourless to blue colour change and some alternative or contributing mechanism must be at play to facilitate the transition.

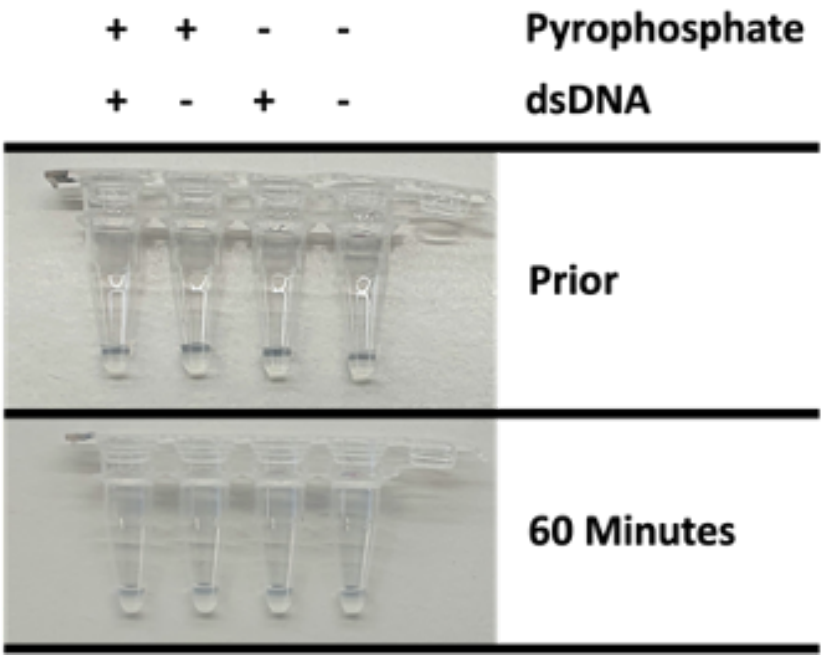


Figure 14. Synthetic recreation of Malachite Green (MG)-based RPA assay end-point detection through the addition of dsDNA and pyrophosphate, to determine cause of colourless to blue dichromatic colour transition.

As I was unable to recreate synthetically the second MG transition (colourless to blue) representing successful amplification, I decided not to pursue the utilisation of MG further; as without this understanding I could not optimise the RPA reaction. In addition, the method's reliance on DTT, which is known to have a short-shelf life when stored at room temperature [28], would be a disadvantage for in-field use by throwing out the stoichiometric ratio of DTT to MG, which facilities the first colour transition, leading to an increased rate of diagnostic failure.

pH-based Colorimetric Detection

Changes in pH can be used for colorimetric-based NAAT end point detection, as successfully demonstrated in LAMP systems such as the NEB WarmStart® Colorimetric kit which incorporates Cresol Red, a pH indicator, into an unbuffered LAMP solution. Upon successful

amplification the solution colour changes from red to yellow, due to a decrease in pH, driven by the use of dNTPs [20]. To adapt RPA for pH-based colorimetric detection a similar approach is required and so I sought to identify compatible indicators [29,30] . The candidate pH-indicators selected were Phenol Red (PR), Cresol Red (CR) and Thymol Blue (TB) which are estimated to cost between £0.23 and £0.85 per 10,000 assays. (**Table 2).** When combined with TwistDx liquid basic RPA reactions, all three indicators were tolerated (**Figure 15)**, but no colour change was observed as the RPA system remained fully buffered.

Table 2. Candidate pH colorimetric indicators.

pH Indicator	pH Range	Colour Transition
Phenol Red	7.9 [31]	Red : Yellow
Cresol Red	8.2 [32]	Purple : Yellow
Thymol Blue.	8.7 [33]	Blue : Yellow

Both Cresol Red and Thymol Blue have an acid dissociation constant (pKa) above the stated starting pH of RPA reaction buffer 7.9 [29]. I observed that the reaction containing Thymol Blue is yellow as expected (**Figure 15)**, however the Phenol Red reaction appears reddish, suggesting that the pH of the RPA reaction upon composition is not 7.9 but instead falls above 8.2. This difference in pH could be driven from commercial optimisation of the RPA reaction post-publication. For each indicator to be successfully incorporated into the RPA reaction, to enable a polychromatic colour change, the starting pH of the RPA reaction must be above the pKa of the indicator. Therefore, to enable the use of Thymol Blue the starting pH of the reaction has to be raised above 8.7. In addition, by raising the starting pH of the RPA system, we would be able to create a more sensitive colorimetric assay. This is because pH is on a logarithmic scale and under the same minimum buffered system a 10x fold increase in amplification would be required to cause a pH drop from 8 to 7 compared to 9 to 8, assuming amplification is directly correlated to pH change. Knowing that RPA reaction efficiency varies, ensuring a

minimal amount of specific amplification is enough for colour transition would theoretically cause a decrease in false negatives and enhance assay sensitivity. As such the next step was to determine the pH tolerance of the RPA reaction to determine if the pH could be raised above the pKa of Thymol Blue. A custom rehydration buffer (Tris 25 mM, PEG 10%, Potassium Acetate 100mM) was created across a pH range of 6.4 to 9.6 and assessed (**Figure 16)**. All rehydration buffers were tolerated, however a significant reduction in amplification efficiency was observed for the rehydration buffer pH 9.6 compared, based on amplicon band intensity. However, this indicates that all three pH indicators could be carried forward.

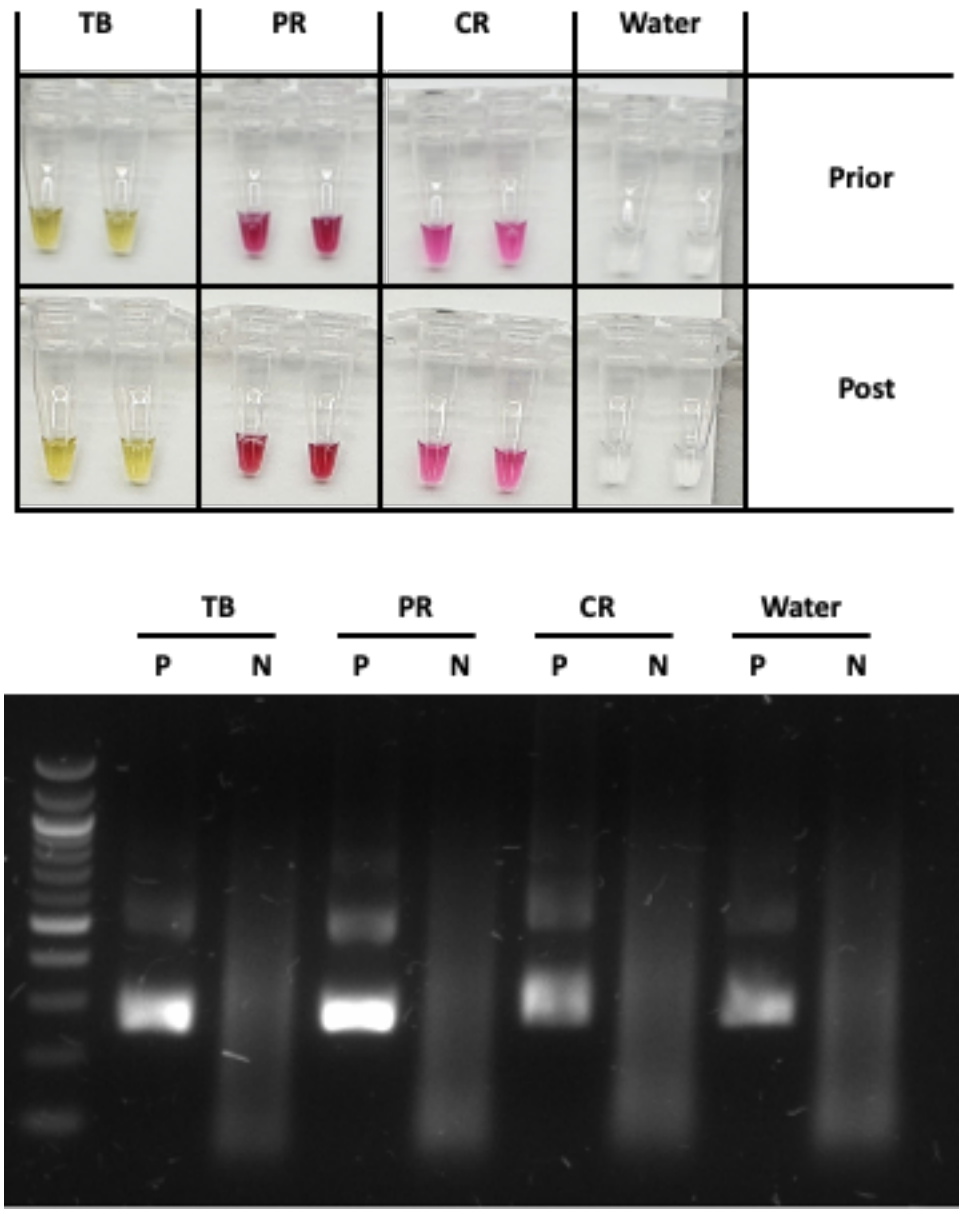


Figure 15. Assessing the tolerance of the C580Y WT targeting RPA reaction to the pH indicators Thymol Blue (TB), Phenol Red (PR) and Cresol Red (CR), alongside any colour change prior to reaction initiation and post reaction completion.

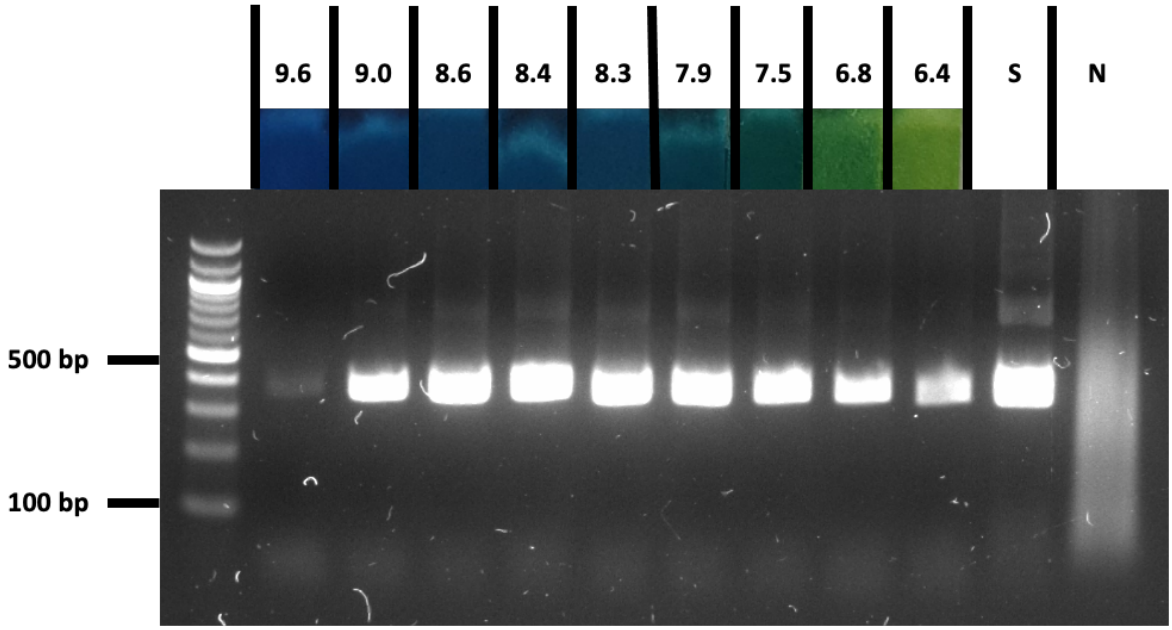


Figure 16. Assessing tolerance of the *PfKelch13* R539T WT genotype targeting RPA to changes in pH introduced through modifying the TwistAmp Liquid Basic Reaction buffer (200mM Potassium Acetate, 100mM Tris, 12% PEG) with variable pH adjusted using Acetic Acid. (S) Standard 2x Reaction Buffer. (N) Negative control. Visualisation of reaction buffer pH differences using Hydrion pH 5-9 litmus paper.

One explanation for the reaction efficiency decreasing above pH 9 is if one or more of the key RPA enzymes is precipitating. This would occur if the isoelectric point of one or more of the essential RPA enzymes is between the pH of the 20x core mix storage buffer and pH 9+. The isoelectric points of the key RPA proteins, as calculated using the Expasy tool [34,35] are shown (**Table 3)**. The TwistDx technical team were unable to release the specific isoelectric points of the proteins, their exact sequence, or their performance under different pH conditions and as such this hypothesis could not be validated. Moving forward, the next step was to optimise the mechanisms which drive the pH change to enable a clear colour transition.

Protein	Isoelectric Point
UvsX	5.29
UvsY	7.76
Gp32	5.05
BSU	5.10

Table 3. Isoelectric points of core RPA proteins and associated Uniprot IDs. UvsX (A0A023ZVM8), UvsY (P04537), Gp32 (B3IYU0), Bsu Polymerase (O34996).

RPA reaction dynamics influencing pH

To explore the mechanisms which influence the pH of the RPA reaction I took an *in-silico* approach, to calculate the pH at reaction equilibrium upon initiation and completion. Please see the **Supplementary Information** section for details of how this was achieved. The incorporation of dNTPs during the amplification and subsequent release of pyrophosphate byproduct has long been known to cause a decrease in pH [36]. Modelling this dynamic in isolation, a starting solution consisting of 200 uM dATP, dTTP, dGTP, dCTP with 3.04 mM of sodium is predicted to have a pH of 8.0. Assuming all dNTPs are incorporated into the growing DNA chain the final solution has a pH of 7.3, representing 0.7 decrease driven by the release of the pyrophosphate byproduct.

Unlike other NAATs, RPA relies on ATP which is essential to UvsX-driven priming of the dsDNA target [37]. Like the incorporation of dNTPs the hydrolysis of ATP is known to cause a decrease in pH. Modelling this in isolation, a starting solution consisting of 3 mM ATP and 11.65 mM sodium has a pH of 8.0. Assuming all ATP is hydrolysed to form 3 mM of ADP and 3 mM of phosphate the predicted pH of the final solution decreases by 0.9 to 7.1. The

concentration of dNTPs and ATP highlighted above were consistent with the original RPA paper [29], and as such, if we assume under successfully amplification conditions all ATP is hydrolysed and all dNTPs are utilised, we can state that ATP hydrolysis has a slightly greater impact on pH than the incorporation of dNTPs.

However, a third RPA reaction dynamic must be considered, which is the RPA energy replacement system that uses phosphocreatine and creatine kinase to regenerate ATP from ADP. A solution of 50 mM phosphocreatine, 50 mM ADP and 240 mM sodium has a pH of 8.0. Assuming all phosphocreatine is utilised in the regeneration of ATP, the reaction end-point solution would consist of 50 mM of ATP, 50 mM of creatine, 240 mM of sodium and has a predicted pH of 12.1. Unlike the other two RPA reaction dynamics, the energy replacement system causes an increase in pH as the reaction progresses.

Predicting the pH of the RPA Reaction

After successfully modelling the RPA reaction dynamics which cause a change in pH, the next step was to model the pH of the whole RPA reaction at equilibrium on initiation and completion assuming successful amplification. The concentration of each component was incorporated into the model according to the first publication by Piepenburg et al [29] (**Supplementary Table 2)**. When accounting for all RPA reaction components the starting pH of the RPA reaction was predicted to be 8.36, which would explain the reddish colour observed when incorporating Phenol Red (**Figure 15).** Upon completion of the reaction, the pH was predicted to be 8.32, representing a minor decrease of 0.04 and assumed the total exhaustion of ATP, the energy replacement system and incorporation of all available dNTPs.

Both Tris and Acetate in the RPA reaction act as buffering agents. Whilst Tris has long been used to buffer NAATs, the high acetate concentration is driven from the use of acetate salts,

to incorporate essential potassium and magnesium cations, as alternatives such as chlorine and sulphate salts have been shown to negative impact the functionality of RecA orthologs including UvsX [37–39]. By removing Tris from the RPA reaction, the pH on reaction completion was predicted to be 8.15, representing a 5x decrease in pH compared to the Tris-buffered system. However, the 0.21 decrease in pH is not optimal for indicator-based colorimetric detection, which ideally needs to be over 1 pH unit or more to produce a clear polychromatic colour transition. To enhance the pH change, one can increase the concentration of dNTPs and/or reduce the energy replacement system. When removing Tris and increasing the concentration of dNTPs 5-fold to 1mM the predicted pH on reaction completion was 8.01, representing a change of 0.35. However, when reducing the concentration of phosphocreatine 5-fold to 10mM the predicted pH on reaction completion was 7.67, representing a decrease of 0.69 and crosses the pKa of Phenol Red, which would theoretically result in a colour transition. Moving forward, it appears reducing the concentration of phosphocreatine will be the optimal method. In addition, by reducing the concentration of acetic acid which is used to adjust the pH of the Tris-based buffer, the starting pH can be increased to 8.50 and resulting pH on reaction completion is 7.69, representing the desired pH change which is compatible with both Phenol Red and Cresol Red pH-based indicators.

After validating *in-silico* that a significant pH change could be achieved on RPA reaction success, when modifying the reaction composition the next step was to experimentally validate the model. Ideally, this would have been through the modification of existing commercially available TwistAMP Liquid Basic kit, however the TwistDx team were unable to reveal the composition, reagent concentrations and pH of each component of the RPA Liquid Kit. As such the only route forward was to express the proteins in-house, so that I had

total control over the RPA reaction composition to both successfully validate the model and create a colorimetric-based RPA assay.

UvsX, UvsY and Gp32 Protein Expression

The first recombinant expression of T4-phage recombinant proteins was published in the late 1980s. [40,41]. Building on this previous work I constructed Pet28a-based expression vectors for UvsX, UvsY, Gp32 and Bsu polymerase. Each protein sequence was optimised for *E. coli* based expression prior to NEB-based HiFi construct assembly. For the proteins UvsX and Gp32, two expression vectors were created to enable both C and N terminal his-tagging of the recombinant proteins for nickel-based protein purification in line with the original RPA publication [29] **(Figure 17).**

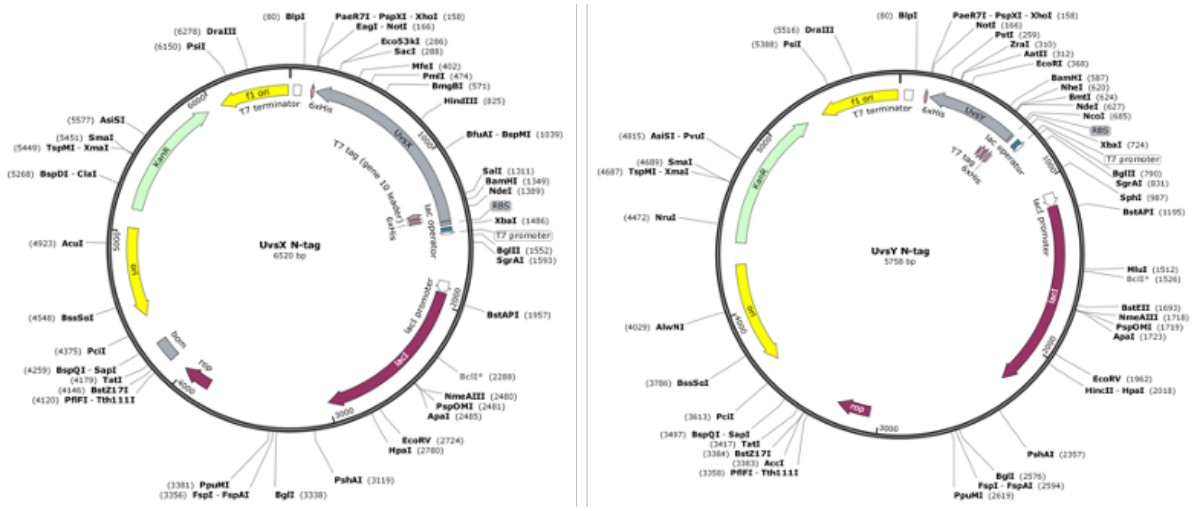


Figure 17. Plasmid maps of UvsX and UvsY expression vectors where both recombinant proteins have an N-terminal His-tag.

Following construct assembly, quality control and propagation using the *E. coli* XL10 cell line, the expression and purification of each protein was optimised in the BL21 (DE3) *E. coli* cell line (**Figure 18; Figure 19**). Due to wet-lab interruptions driven by the SARS-Cov-2 pandemic, I was unable to optimise the protein purification strategy past the final buffer

exchange and protein concentration steps (see Methods). As such this is where work on pH-based colorimetric end-point expression finished.

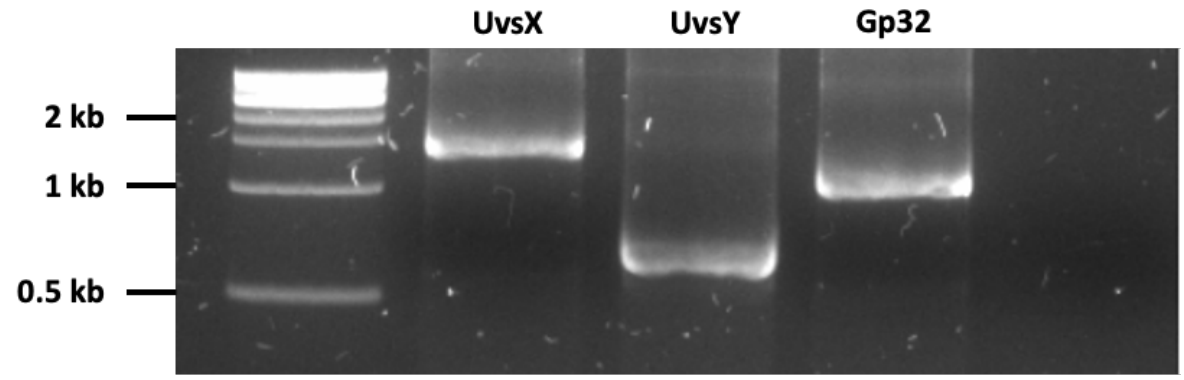


Figure 18. PCR-based screening to assess successful creation of expression vectors, UvsX, UvsY and Gp32 based on expected band size.

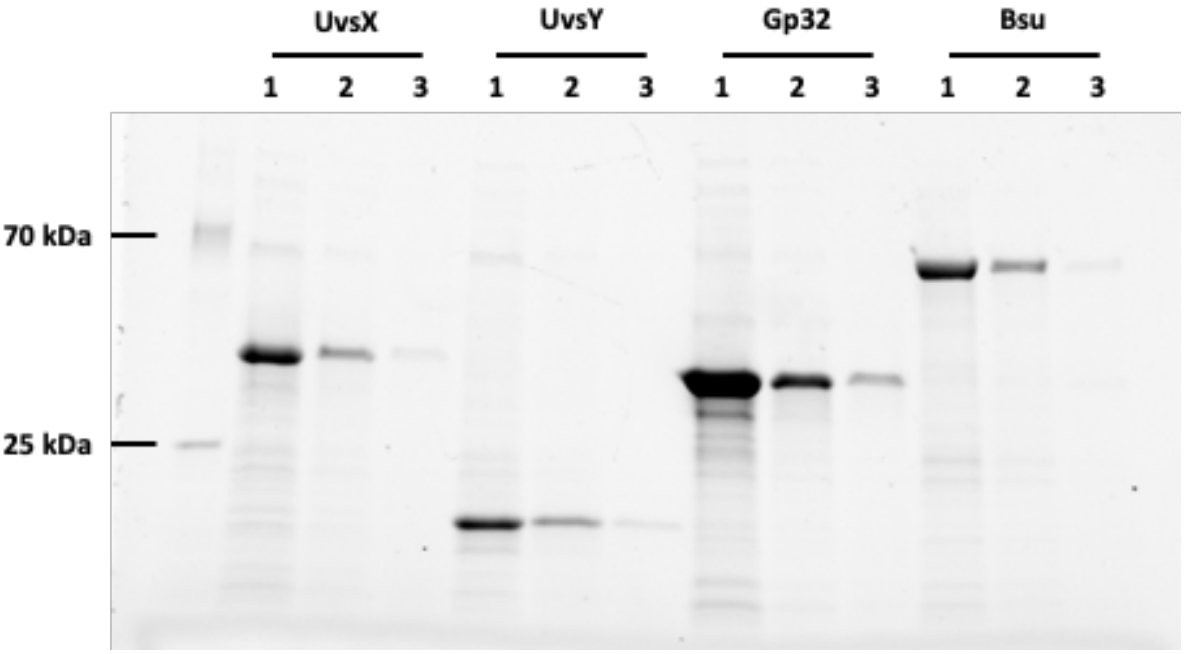


Figure 19. Assessing expression of key RPA proteins, following NEBExpress® Ni Resin-based purification, via SDS-Page. Elution fractions 1-3 are shown for each protein, alongside the PageRuler Protein Ladder.

Discussion

In this chapter I explored RPA's potential for SNP-based genotyping in the detection of markers associated with antimalarial resistance. Further work is required to investigate the true impact of introducing mismatches towards the 3'-terminus of the primer-template complex, whether this approach can be generalised for reliable RPA-based genotyping, and inferring its limits according to cofactors such as the target concentration. Some of this work is explored in **Chapter 4**.

In the pursuit of a cost-effective single-step colorimetric RPA assay I made substantial progress. I attempted to recreate the published SYBR-green RPA assay [12], and in doing so uncovered these methodologies limitations, including but not limited to its incompatibility with a one-step assay approach. In addition, I was able to successfully incorporate MG to form a one-step RPA assay and partially uncover the novel DTT-associated mechanism behind the colour change which was previously unreported [16]. Further work is needed to elucidate the full mechanism of MG-based detection, however if achieved the RPA reaction can be optimised accordingly to create a polychromatic one-step colorimetric assay. In addition, whilst I was able to demonstrate the viability of a pH-based colorimetric approach *in-silico* I was unable to complete this line of investigation due to wet-lab interruptions, but the preliminary experiments completed demonstrate that this approach has promise.

The colorimetric methods investigated all rely upon and correlate to general amplification. This dependency opens the risk of false positives driven by non-specific amplification. Whilst the low running temperature makes RPA suitable for in-field use, it favours the formation of primer-dimer artefacts or primer secondary structure, both of which can lead to non-specific amplification, leading to an increase in the risk of false positives. The introduction of specific SARMS-nucleotides into the primers can mitigate primer-driven non-

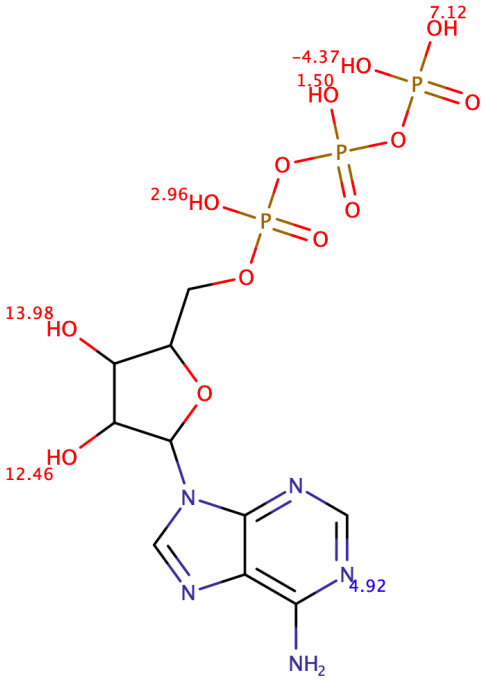
specific amplification [42], however, the risk of non-specific amplification by background DNA self-priming cannot be addressed. For example, when theoretically testing on blood samples from a suspected malaria patient, both human and potential *Plasmodium* DNA will be present. As such, further work is needed to infer the level of nonspecific background amplification when dealing with field samples, and how to optimise one-step colorimetric assays accordingly to mitigate the risk of false positives. It is conceivable that the RPA reaction could be optimised so that the threshold for colorimetric change would only be reached during exponential primer-driven amplification of the target region and not background DNA amplification.

Upon writing this thesis, several attempts by other researchers to adapt RPA for colorimetric detection have been made. This included the use of hydroxynaphthol blue, whose monochromatic transitions from dark blue to light blue is triggered by the sequestering of magnesium cations by pyrophosphate byproduct released during successful amplification [43]. Whilst this transition is monochromatic, it does represent an important step forward in the development of colorimetric RPA assays. In addition, further work is needed to ensure that no false positives are generated for UvsX catalysis of ATP to form AMP and pyrophosphate, which could impede the reaction. [43–45] In addition, Tomer et al [46], published an attempt to adapt RPA for pH-based detection, in which they successfully showed removal of the reaction buffering components resulted in a change in pH, as predicted. However, they were unable to successfully enhance the pH change by modifying the energy replacement system. This I believe was potentially due to a flaw in the experimental design as they did not adjust the reaction to ensure the starting pH of the reaction was consistent [46]. With this in mind I remain optimistic for a cost-effective pH-based RPA assay, which improves upon existing malaria RDT assays.

Supplementary Information

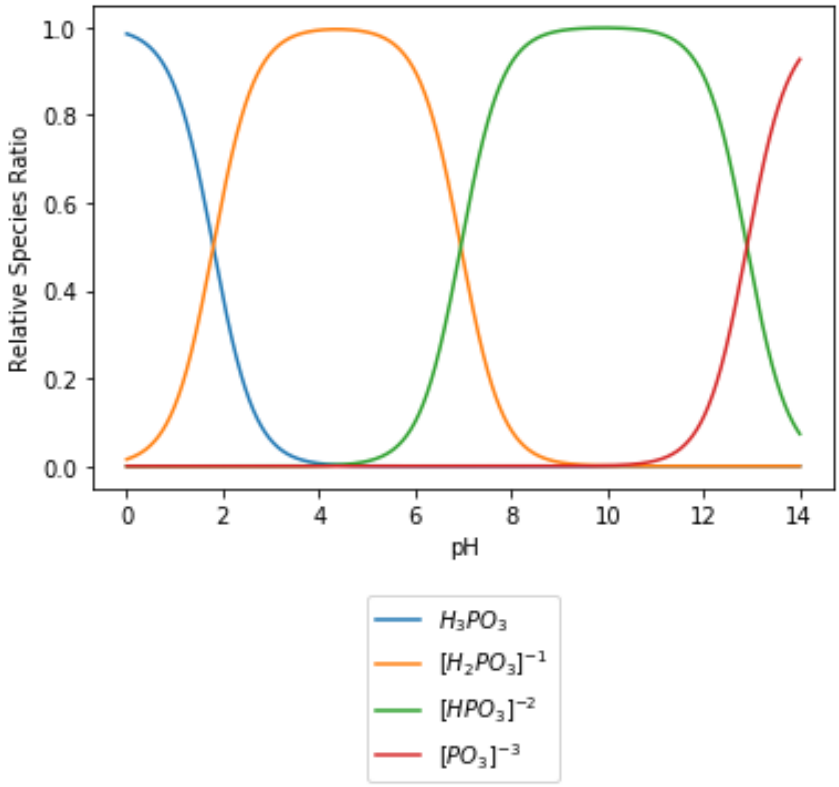
Predicting pH *in-silico*

It is possible to predict the pH of a solution when it is at equilibrium by solving a phenomenon known as the equilibrium problem. In any solution where the solutes (i.e., salts, ATP, dNTPs) are dissolved in the solvent (water) we can estimate the pH. First the equilibrium concentrations of each solute and their respective species i.e., different charge states are derived, and we subsequently calculate the concentration of hydronium ions [H_3O^+] or hydroxyl ions [OH^-] required to balance the net charge of the solution, through which we obtain our pH estimate [47]. In summary we are addressing the mass and charge balance of the solutes. To utilise the pHcalc python package, I had to derive the species for each RPA reaction component and the associated pKa values (**Supplementary Table 1)** [8]. **Supplementary Figure 1** highlights the Marvin predicted pKa values for ATP.



Supplementary Figure 1. ChemAxon Marvin predicted pKa values for each functional group of ATP. Groups shown in red are protonated.

Compounds can have multiple pKa values due to the presence of multiple functional groups which can act as an acid or base. **Supplementary Figure 2** demonstrates how the species distribution of phosphate (Pi), which has 3 pKa values, changes across a pH gradient (1-14).



Supplementary Figure 2. Changes in phosphate species abundance between pH 1-14.

Deriving *In-silico* Scalars

For reagents such as 100mM ATP (pH 7.3-7.5) and 10 mM dNTPs (pH 8.3) produced by ThermoFisher and Roche respectively, the product specifications state they have already undergone buffering with sodium hydroxide (NaOH). For the *in-silico* pH prediction to be accurate I needed to account for this buffering and as such, knowing the product concentration and final pH according to the product description we can reverse engineer the equilibrium equation to obtain the concentration of NaOH used for buffering which was 360mM and 155.7mM for the ATP and dNTP stocks respectively. Moving forward, for ATP

and dNTP-based reagents I incorporated NaOH scalars of 3.6x and 15.57x so I can accurately account for any underlying buffering.

Supplementary Table 1. pKa values for each RPA reaction component present. The charge column corresponds to the charge of the fully protonated species of each compound according to Marvin Chem Axon software prediction. To represent the dsDNA chain, I have included dGMP, dTMP, dCMP and dAMP.

Reagent	Charge	pKa
Tris	1	8.1
Phosphocreatine	1	3.35, 12.66, -0.52, 5.26
Creatine	1	3.50, 12.43
ATP	1	7.12, -4.37, 1.5, 2.96, 13.98, 12.46, 4.92
ADP	2	7.42, 1.77, 2.22, 13.98, 12.46, -1.05, 4.93
dGTP	1	7.42, 2.54, 1.04, 3.30, 13.98, 0.37, 10.16
dGMP	1	1.96, 13.98, 0.51, 10.16
dCTP	1	4.44, 13.99, 3.24, 0.89, 2.49, 7.42
dCMP	1	4.48, 13.99, 1.3
dATP	2	4.95, -1.03, 13.98, 3.29, 0.90, 1.55, 7.42
dAMP	2	3.9, -1.03, 13.98, 1.94
dTTP	0	9.96, 14.01, 3.3, 0.9, 2.54, 7.42
dTMP	0	9.96, 14.01, 1.95
Pi	0	12.9, 6.95, 1.8
PPi	0	1.7, 6.98, 3.06, 8.17
PEG	0	14.82, 15.42

Reagent	Charge	pKa
DTT	0	10.22, 15.66, 14.24, 9.62
Acetate	0	4.54
CresolRed	0	8.3, 9.52, 10.12
PhenolRed	0	9.76, 9.16, 7.9
ThymolBlue	0	10.35, 9.75, 8.9

Supplementary Table 2. The reaction compositions for predicting pH. 1) Original RPA publication, including the concentration of acetic acid required to make Tris Buffer pH 7.9 [29]. 2) Removal of Tris Buffer, 3) Removal of Tris buffer and 5x increase in dNTPs, 4) Removal of Tris Buffer and 5x decrease in Phosphocreatine, 5) Reduction in Acetic Acid.

	Molarity (M)				
	1	**2**	**3**	**4**	**5**
Tris	0.05	0	0	0	0
Phosphocreatine	0.05	0.05	0.05	0.01	0.01
ATP	0.003	0.003	0.003	0.003	0.003
dNTP	0.0002	0.0002	0.001	0.0002	0.0002
Potassium Acetate	0.1	0.1	0.1	0.1	0.1
Magnesium Acetate	0.014	0.014	0.014	0.014	0.014
PEG	0.0000062	0.0000062	0.0000062	0.0000062	0.0000062
DTT	0.002	0.002	0.002	0.002	0.002
Potassium Hydroxide	0	0	0	0	0
Acetic Acid	0.0306	0.0129	0.01285	0.01287	0.01275

Supplementary Table 3. Primer used in experiments outlined in this paper.

ID	Sequence
PfKelch13 C580Y WT	CCCCTAGATCATCAGCTATGTG
PfKelch13 C580Y Res	CCCCTAGATCATCAGCTATGGA
PfKelch13 R539T WT	GTGGTGTTACGTCAAATGGTAG
PfKelch13 R539T Res	GTGGTGTTACGTCAAATGGGAC
PfKelch13 I543T WT	CAAATGGTAGAATTTATTGTAT
PfKelch13 I543T Res	CAAATGGTAGAATTTATTGGAC
PfKelch13 Reverse	TTATTAAATGGTTGATATTGTTCAACGGAATC
R539T SDM FP	ATGGTACAATTTATTGTATTGG
R539T SDM RP	TTGACGTAACACCACAA
I543T SDM FP	GAATTTATTGTACTGGGGGATATGAT
I543T SDM RP	TACCATTTGACGTAACACCAC
C580Y SDM FP	CATCAGCTATGTATGTTGCTTTTGAT
C580Y SDM RP	ATCTAGGGGTATTCAAAGGTGCC

gBlocks utilised in recombinant Protein Expression

> gBlock for UvsX and UvsY

AGGACTCTAAGTGTAGGATCCATGAGTGATTTAAAATCAAGGCTAATAAAGGCG

TCGACCAGCAAACTGACGGCGGAACTGACCGCTAGCAAGTTCTTCAACGAGAAG

GACGTTGTTCGCACCAAAATCCCGATGATGAACATCGCGCTCTCTGGTGAGATCA

CGGGCGGCATGCAGAGTGGCCTGCTGATTCTGGCGGGTCCAAGCAAGAGCTTTA

AATCCAATTTTGGTCTTACCATGGTGAGCAGCTATATGCGTCAATATCCGGATGC

CGTTTGCCTGTTCTATGATTCGGAATTTGGTATTACCCCGGCATACCTGCGTAGCA

TGGGTGTCGATCCGGAACGCGTGATCCACACCCCTGTTCAGTCTCTGGAACAGCT

GCGTATTGATATGGTTAATCAGTTGGACGCCATCGAACGTGGTGAAAAAGTGGT

GGTCTTTATCGACTCGCTGGGTAATCTGGCGAGCAAGAAAGAGACGGAAGATGC

TCTGAACGAAAAGGTGGTTAGCGATATGACCCGTGCGAAAACGATGAAAAGCTT

GTTCCGTATTGTTACCCCGTACTTCTCCACTAAGAACATCCCGTGCATCGCGATTA

ACCATACTTACGAAACCCAGGAGATGTTTAGCAAAACCGTGATGGGTGGCGGCA

CCGGCCCGATGTATAGCGCGGACACCGTGTTCATTATTGGTAAACGTCAGATTAA

AGACGGCTCAGATTTGCAGGGTTACCAATTTGTGCTGAATGTTGAGAAGTCCAGA

ACAGTCAAGGAGAAGTCAAAGTTTTTCATCGACGTGAAATTCGATGGCGGCATC

GACCCGTACAGCGGTTTGTTAGACATGGCCCTGGAGCTGGGTTTTGTTGTAAAAC

CGAAAAACGGCTGGTACGCACGTGAGTTCCTGGACGAGGAAACGGGCGAAATG

ATTCGCGAAGAGAAGTCCTGGCGTGCTAAAGACACCAATTGTACCACCTTTTGGG

GTCCGCTGTTTAAGCACCAACCGTTCCGCGATGCGATCAAGCGCGCATATCAACT

CGGAGCGATCGACTCTAACGAGATTGTGGAAGCAGAGGTTGACGAGCTCATCAA

CAGCAAGGTCGAGAAGTTCAAATCCCCAGAAAGCAAGTCTAAATCTGCGGCTGA

TTTGGAAACGGATTTGGAACAATTATCGGACATGGAAGAGTTTAACGAGTAGGC

GGCCGCAGGTGATTCATCTGAAGCGTTACTGTAGGTGGATCCATGAGGTTGGAA

GATTTACAAGAGGAACTAAAGAAGGACGTTTTTATCGATTCTACGAAACTGCAA

TACGAGGCGGCGAACAACGTGATGCTGTATTCCAAATGGTTGAATAAACATAGC

AGTATTAAGAAGGAGATGCTGCGTATTGAAGCGCAGAAAAAGGTTGCGCTGAAG

GCTCGCCTGGACTATTACTCCGGCCGTGGTGACGGCGATGAATTCAGCATGGATC

GTTATGAGAAATCGGAAATGAAAACCGTTTTAAGCGCAGATAAAGACGTCTTGA

AGGTGGATACCAGCCTGCAATACTGGGGTATTCTGTTGGACTTCTGCAGCGGCGC

TCTCGACGCGATCAAGAGCCGTGGTTTCGCCATCAAGCACATCCAGGATATGCGC

GCATTTGAGGCCGGTAAATAGGCGGCCGCATTGCGAGATGGATC

> gBlock for Gp32

AGCTTATCGAGGTCAGGATCCATGTTTAAAAGGAAGTCAACAGCTGAACTAGCC

GCGCAAATGGCTAAATTGAACGGTAACAAAGGCTTTAGCTCCGAGGATAAAGGT

GAGTGGAAACTTAAACTTGATAACGCTGGTAACGGCCAAGCAGTGATTCGTTTCC

TGCCGAGCAAGAACGACGAGCAGGCGCCATTTGCCATTCTGGTTAATCATGGTTT

CAAGAAGAATGGCAAGTGGTATATCGAAACGTGCAGCTCGACCCACGGCGATTA

CGACAGCTGCCCGGTCTGTCAGTATATCAGCAAGAACGATCTGTACAATACCGAT

AATAAAGAATACAGCCTGGTCAAGCGCAAAACCAGCTATTGGGCAAATATCCTG

GTTGTTAAAGATCCGGCAGCTCCGGAAAACGAAGGCAAGGTGTTTAAGTACCGT

TTCGGCAAAAAAATTTGGGATAAGATCAACGCAATGATCGCAGTTGACGTGGAA

ATGGGCGAGACGCCGGTTGACGTGACCTGCCCGTGGGAAGGTGCGAACTTTGTA

CTGAAGGTGAAGCAGGTTAGCGGTTTCTCTAATTATGATGAGAGCAAGTTCCTGA

ACCAGTCTGCCATCCCGAATATTGACGACGAGTCCTTTCAGAAAGAGCTGTTTGA

GCAAATGGTTGATCTGAGCGAAATGACCAGCAAGGACAAGTTCAAAAGCTTCGA

GGAATTGAACACGAAATTCGGCCAAGTTATGGGTACTGCGGTCATGGGTGGTGC

GGCGGCGACCGCTGCGAAAAAAGCCGACAAGGTGGCGGATGACTTAGATGCGTT

TAACGTGGACGACTTCAACACCAAGACCGAAGACGACTTCATGTCCTCAAGTTCC

GGTTCTAGCTCCTCGGCGGATGATACCGACTTGGACGACCTCCTGAACGATTTGT

AGGCGGCCGCAGCCGGATTTAGCGA

> gBlock for Bsu polymerase fragment.

AGCTTATCGAGGTCAGGATCCATGACAGAAAGGAAGAAATTAGTTCTAGTAGAC

GGCAACTCCCTGGCATACCGCGCGTTCTTTGCACTGCCACTTTTGTCAAATGATA

AGGGCGTTCACACCAATGCAGTTTACGGCTTCGCCATGATTTTGATGAAAATGCT

GGAGGACGAGAAACCGACCCATATGCTGGTGGCCTTCGACGCCGGTAAAACCAC

CTTCCGCCACGGCACCTTTAAAGAGTACAAAGGCGGTCGCCAGAAGACGCCGCC

TGAACTGAGTGAGCAGATGCCGTTCATTCGCGAATTGCTTGATGCCTACCAAATC

TCTCGTTACGAACTGGAGCAATATGAGGCGGACGACATCATCGGCACCTTGGCG

AAAAGCGCGGAAAAGGACGGCTTCGAGGTTAAAGTTTTCTCCGGTGATAAAGAT

CTGACGCAGCTGGCGACGGATAAGACGACCGTTGCTATCACCAGAAAGGGTATC

ACGGATGTGGAGTTCTATACCCCGGAGCACGTGAAAGAAAAATATGGACTGACC

CCGGAGCAAATTATCGACATGAAGGGCCTGATGGGTGACTCTTCCGACAACATT

CCGGGTGTGCCGGGTGTGGGTGAGAAGACTGCGATTAAATTGCTGAAACAGTTT

GATTCCGTGGAGAAGCTGCTCGAATCCATTGATGAGGTGAGCGGCAAGAAATTG

AAGGAGAAGTTAGAAGAATTTAAAGACCAGGCGTTGATGAGCAAGGAATTGGC

GACCATTATGACCGATGCTCCGATCGAAGTTAGCGTGAGCGGTCTGGAATACCA

GGGTTTCAATCGTGAACAAGTTATTGCGATTTTTAAAGACTTAGGCTTCAACACC

CTGCTGGAACGTCTCGGCGAGGATAGCGCGGAGGCCGAACAAGATCAAAGCTTA

GAGGACATCAATGTTAAAACGGTAACGGACGTTACCTCCGATATTCTGGTGTCGC

CGAGCGCGTTCGTGGTTGAACAAATTGGCGACAACTATCATGAAGAGCCGATTC

TGGGTTTTTCTATTGTTAATGAAACCGGCGCGTATTTTATCCCGAAAGATATCGC

AGTCGAGTCGGAGGTTTTCAAAGAGTGGGTTGAGAACGACGAGCAGAAAAAGTG

GGTGTTCGACAGCAAACGTGCGGTAGTTGCACTGCGCTGGCAGGGCATCGAATT

GAAGGGCGCGGAATTCGATACCCTGTTAGCTGCCTACATTATCAATCCGGGTAAC

AGCTATGATGATGTGGCGAGCGTCGCCAAAGACTACGGTCTGCATATCGTGTCCT

CTGACGAGAGCGTTTACGGTAAGGGCGCTAAACGTGCGGTGCCAAGCGAGGACG

TCTTGTCGGAACATCTGGGTCGTAAAGCGCTGGCGATTCAGAGCCTGCGTGAAA

AGTTGGTTCAAGAGCTGGAGAACAACGATCAGCTGGAGTTGTTCGAGGAGCTGG

AGATGCCGCTGGCACTCATACTGGGTGAAATGGAAAGCACGGGCGTAAAGGTGG

ACGTTGATCGTCTGAAGCGCATGGGTGAAGAGCTCGGCGCAAAGTTGAAGGAGT

ACGAAGAGAAGATCCACGAAATCGCTGGCGAGCCGTTTAATATCAACTCTCCGA

AACAACTGGGTGTCATCCTGTTTGAAAAGATTGGTCTGCCGGTTGTCAAGAAAAC

CAAGACTGGCTATTCCACTTCAGCGGACGTGCTGGAAAAATTGGCCGACAAACA

CGATATTGTGGATTATATCCTGCAATATCGTCAAATCGGCAAACTTCAAAGCACC

TATATCGAGGGTCTGCTGAAGGTGACTCGCCCAGATTCCCATAAAGTGCACACGC

GTTTTAACCAGGCCTTGACCCAGACCGGTCGCCTGAGCTCTACCGACCCGAATCT

GCAAAACATTCCGATTCGTCTTGAAGAAGGTCGTAAAATCCGCCAGGCATTTGTT

CCGAGCGAAAAGGACTGGCTGATCTTCGCGGCTGACTACAGCCAGATCGAATTG

CGTGTTCTGGCGCATATTAGCAAGGACGAGAACCTGATTGAAGCCTTTACCAACG

ACATGGATATCCACACCAAAACGGCTATGGATGTATTCCACGTCGCAAAAGACG

AGGTGACCTCCGCCATGCGTCGTCAAGCGAAGGCGGTGAACTTCGGTATCGTTTA

TGGGATTAGCGATTACGGTTTGTCTCAGAACCTGGGTATCACCCGTAAAGAGGCG

GGTGCGTTCATCGATCGTTACCTGGAGAGCTTTCAGGGTGTGAAAGCGTACATGG

AAGACTCTGTCCAGGAGGCTAAGCAGAAAGGCTATGTTACAACTCTCATGCATC

GTCGCCGCTACATTCCGGAACTGACGTCCCGTAACTTCAACATCAGATCGTTTGC

GGAACGTACCGCTATGAATACCCCGATCCAAGGTAGCGCTGCGGACATCATTAA

AAAGGCTATGATCGATATGGCGGCGAAGCTGAAGGAGAAGCAGCTGAAAGCCC

GTCTGCTGCTGCAGGTCCACGATGAACTGATTTTCGAAGCACCGAAAGAGGAAA

TTGAGATCCTGGAGAAGCTCGTTCCGGAAGTCATGGAACACGCGCTGGCTTTAG

ATGTCCCGTTGAAGGTTGATTTTGCAAGCGGTCCGTCGTGGTATGACGCTAAGTA

GGCGGCCGCAGCCGGATTTAGCGA

Pfkelch13 Region of Interest

CTAGAAGAAATAATTGTGGTGTTACGTCAAATGGTAGAATTTATTGTATTGGGGG

ATATGATGGCTCTTCTATTATACCGAATGTAGAAGCATATGATCATCGTATGAAA

GCATGGGTAGAGGTGGCACCTTTGAATACCCCTAGATCATCAGCTATGTGTGTTG

CTTTTGATAATAAAATTTATGTCATTGGTGGAACTAATGGTGAGAGATTAAATTC

TATTGAAGTATATGAAGAAAAAATGAATAAATGGGAACAATTTCCATATGCCTT

ATTAGAAGCTAGAAGTTCAGGAGCAGCTTTTAATTACCTTAATCAAATATATGTT

GTTGGAGGTATTGATAATGAACATAACATATTAGATTCCGTTGAACAATATCAAC

CATTTAATAAAAGATGGCAATTTCTAAATGGTGTACCAGAGAAAAAAATGAATT

TTGGAGCTGCCACATTGTCAGATTCTTATATAA

References

1. Kersting, S., Rausch, V., Bier, F. F. & von Nickisch-Rosenegk, M. Rapid detection of Plasmodium falciparum with isothermal recombinase polymerase amplification and lateral flow analysis. *Malar. J.* **13**, 99 (2014).
2. Sun, Y. *et al.* One-tube SARS-CoV-2 detection platform based on RT-RPA and CRISPR/Cas12a. *J. Transl. Med.* **19**, 74 (2021).
3. Strayer-Scherer, A., Jones, J. B. & Paret, M. L. Recombinase Polymerase Amplification Assay for Field Detection of Tomato Bacterial Spot Pathogens. *Phytopathology* **109**, 690–700 (2019).
4. Lillis, L. *et al.* Factors influencing Recombinase polymerase amplification (RPA) assay outcomes at point of care. *Mol. Cell. Probes* **30**, 74–78 (2016).
5. Lai, M.-Y., Ooi, C.-H. & Lau, Y.-L. Recombinase Polymerase Amplification Combined with a Lateral Flow Strip for the Detection of Plasmodium knowlesi. *Am. J. Trop. Med. Hyg.* **98**, 700–703 (2018).
6. Oboh, M. A. *et al.* Status of Artemisinin Resistance in Malaria Parasite Plasmodium falciparum from Molecular Analyses of the Kelch13 Gene in Southwestern Nigeria. *Biomed Res. Int.* **2018**, 2305062 (2018).
7. Wilairatana, P., Tangpukdee, N. & Krudsood, S. Definition of hyperparasitemia in severe falciparum malaria should be updated. *Asian Pac. J. Trop. Biomed.* **3**, 586 (2013).
8. Prediction of dissociation constant using microconstants. https://chemaxon.com/poster/prediction-of-dissociation-constant-using-microconstants.
9. Burgess, R. R. [1] Use of polyethyleneimine in purification of DNA-binding proteins. in *Methods in Enzymology* vol. 208 3–10 (Academic Press, 1991).

10. Cordes, R. M., Sims, W. B. & Glatz, C. E. Precipitation of nucleic acids with poly(ethyleneimine). *Biotechnol. Prog.* **6**, 283–285 (1990).
11. Singpanomchai, N. *et al.* Naked eye detection of the Mycobacterium tuberculosis complex by recombinase polymerase amplification-SYBR green I assays. *J. Clin. Lab. Anal.* **33**, e22655 (2019).
12. Lai, M. Y. & Lau, Y. L. Detection of Plasmodium knowlesi using recombinase polymerase amplification (RPA) combined with SYBR Green I. *Acta Trop.* **208**, 105511 (2020).
13. Ponchel, F. *et al.* Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. *BMC Biotechnol.* **3**, 18 (2003).
14. Nath, K., Sarosy, J. W., Hahn, J. & Di Como, C. J. Effects of ethidium bromide and SYBR® Green I on different polymerase chain reaction systems. *J. Biochem. Biophys. Methods* **42**, 15–29 (2000).
15. Higgins, M. *et al.* PrimedRPA: primer design for recombinase polymerase amplification assays. *Bioinformatics* **35**, 682–684 (2019).
16. Lucchi, N. W., Ljolje, D., Silva-Flannery, L. & Udhayakumar, V. Use of Malachite Green-Loop Mediated Isothermal Amplification for Detection of Plasmodium spp. Parasites. *PLoS One* **11**, e0151437 (2016).
17. Chieng, H. I., Lim, L. B. L. & Priyantha, N. Enhancing adsorption capacity of toxic malachite green dye through chemically modified breadnut peel: equilibrium, thermodynamics, kinetics and regeneration studies. *Environ. Technol.* **36**, 86–97 (2015).
18. Cigén, R., Ekström, C.-G., Holm, A., Nielsen, P. H. & Munch-Petersen, J. Equilibrium and kinetic studies on Halide derivatives of malachite green. 1. Ortho-fluoro malachite green. *Acta Chem. Scand.* **17**, 1189–1195 (1963).

19. Cooksey, C. J. Quirks of dye nomenclature. 6. Malachite green. *Biotech. Histochem.* **91**, 438–444 (2016).
20. Tanner, N. A., Zhang, Y. & Evans, T. C., Jr. Visual detection of isothermal nucleic acid amplification using pH-sensitive dyes. *Biotechniques* **58**, 59–68 (2015).
21. Nzelu, C. O. *et al.* Development of a loop-mediated isothermal amplification method for rapid mass-screening of sand flies for Leishmania infection. *Acta Trop.* **132**, 1–6 (2014).
22. Hu, X., Jiao, K., Sun, W. & You, J.-Y. Electrochemical and spectroscopic studies on the interaction of malachite green with DNA and its application. *Electroanalysis* **18**, 613–620 (2006).
23. Kolpashchikov, D. M. Binary malachite green aptamer for fluorescent detection of nucleic acids. *J. Am. Chem. Soc.* **127**, 12442–12443 (2005).
24. Babendure, J. R., Adams, S. R. & Tsien, R. Y. Aptamers switch on fluorescence of triphenylmethane dyes. *J. Am. Chem. Soc.* **125**, 14716–14717 (2003).
25. Jothikumar, P., Narayanan, J. & Hill, V. R. Visual endpoint detection of Escherichia coli O157:H7 using isothermal Genome Exponential Amplification Reaction (GEAR) assay and malachite green. *J. Microbiol. Methods* **98**, 122–127 (2014).
26. Vardakou, M., Salmon, M., Faraldos, J. A. & O'Maille, P. E. Comparative analysis and validation of the malachite green assay for the high throughput biochemical characterization of terpene synthases. *MethodsX* **1**, 187–196 (2014).
27. Simão, A. M. S., Bolean, M., Hoylaerts, M. F., Millán, J. L. & Ciancaglini, P. Effects of pH on the production of phosphate and pyrophosphate by matrix vesicles' biomimetics. *Calcif. Tissue Int.* **93**, 222–232 (2013).
28. SIGMA. DL-Dithiothreitol Product Information.
29. Piepenburg, O., Williams, C. H., Stemple, D. L. & Armes, N. A. DNA detection using recombination proteins. *PLoS Biol.* **4**, e204 (2006).

30. Li, J., Macdonald, J. & von Stetten, F. Review: a comprehensive summary of a decade development of the recombinase polymerase amplification. *Analyst* **144**, 31–67 (2018).
31. SIGMA. Phenol Red Product Infomation. https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/347/859/p4633pis.pdf.
32. Yuan, S. & DeGrandpre, M. D. Evaluation of indicator-based pH measurements for freshwater over a wide range of buffer intensities. *Environ. Sci. Technol.* **42**, 6092–6099 (2008).
33. Yimkosol, W. & Dangkulwanich, M. Finding the pKa Values of a Double-Range Indicator Thymol Blue in a Remote Learning Activity. *J. Chem. Educ.* **98**, 3930–3934 (2021).
34. Bjellqvist, B., Basse, B., Olsen, E. & Celis, J. E. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* **15**, 529–539 (1994).
35. Bjellqvist, B. *et al.* The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14**, 1023–1031 (1993).
36. Salm, E. *et al.* Electrical detection of nucleic acid amplification using an on-chip quasi-reference electrode and a PVC REFET. *Anal. Chem.* **86**, 6968–6975 (2014).
37. Formosa, T. & Alberts, B. M. Purification and characterization of the T4 bacteriophage uvsX protein. *J. Biol. Chem.* **261**, 6107–6118 (1986).
38. Bell, C. E. Structure and mechanism of Escherichia coli RecA ATPase. *Mol. Microbiol.* **58**, 358–366 (2005).
39. Menetski, J. P. & Kowalczykowski, S. C. Interaction of recA protein with single-stranded DNA. Quantitative aspects of binding affinity modulation by nucleotide cofactors. *J. Mol. Biol.* **181**, 281–295 (1985).

40. Hinton, D. M. & Nossal, N. G. Cloning of the bacteriophage T4 uvsX gene and purification and characterization of the T4 uvsX recombination protein. *J. Biol. Chem.* **261**, 5663–5673 (1986).
41. Kodadek, T., Gan, D. C. & Stemke-Hale, K. The Phage T4 uvsY Recombination Protein Stabilizes Presynaptic Filaments*. *J. Biol. Chem.* **264**, 16451–16457 (1989).
42. Sharma, N., Hoshika, S., Hutter, D., Bradley, K. M. & Benner, S. A. Recombinase-based isothermal amplification of nucleic acids with self-avoiding molecular recognition systems (SAMRS). *Chembiochem* **15**, 2268–2274 (2014).
43. Priti, Jangra, S., Baranwal, V. K., Dietzgen, R. G. & Ghosh, A. A rapid field-based assay using recombinase polymerase amplification for identification of Thrips palmi, a vector of tospoviruses. *J. Pest Sci.* **94**, 219–229 (2021).
44. Goto, M., Honda, E., Ogura, A., Nomoto, A. & Hanaki, K.-I. Colorimetric detection of loop-mediated isothermal amplification reaction by using hydroxy naphthol blue. *Biotechniques* **46**, 167–172 (2009).
45. Farb, J. N. & Morrical, S. W. Functional complementation of UvsX and UvsY mutations in the mediation of T4 homologous recombination. *Nucleic Acids Res.* **37**, 2336–2345 (2009).
46. Tomar, S., Lavickova, B. & Guiducci, C. Recombinase polymerase amplification in minimally buffered conditions. *Biosens. Bioelectron.* **198**, 113802 (2022).
47. Baeza-Baeza, J. J. & García-Álvarez-Coque, M. C. Systematic Approach To Calculate the Concentration of Chemical Species in Multi-Equilibrium Problems. *J. Chem. Educ.* **88**, 169–173 (2011).

Chapter 4. Characterising the Impact of Primer-Template Mismatches on Recombinase Polymerase Amplification.

RESEARCH PAPER COVER SHEET

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

SECTION A – Student Details

Student ID Number	1702842	**Title**	Mr
First Name(s)	Matthew		
Surname/Family Name	Higgins		
Thesis Title	Developing an RPA-based Molecular Barcoding Tool for Plasmodium Malaria		
Primary Supervisor	Prof. Taane Clark		

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?	Journal of Molecular Diagnostics		
When was the work published?	2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
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?	Journal of Molecular Diagnostics
Please list the paper's authors in the intended authorship order:	Matthew Higgins, Oliver W. Stringer, Daniel Ward, Jenny Andrews, Matthew S. Forrest, Susana Campino, Taane G. Clark
Stage of publication	**In press**

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 performed all data processing and subsequent analysis to assess the impact of primer-template mismatches on the RPA reaction. I wrote the first draft of the manuscript which was then circulated to supervisors and co-authors.

SECTION E

Student Signature	[redacted]
Date	26/06/2022

Supervisor Signature	[redacted]
Date	26/6/2022

The Journal of Molecular Diagnostics, Vol. 24, No. 11, November 2022

the Journal of Molecular Diagnostics
jmdjournal.org

Characterizing the Impact of Primer-Template Mismatches on Recombinase Polymerase Amplification

Matthew Higgins,* Oliver W. Stringer,† Daniel Ward,* Jennifer M. Andrews,* Matthew S. Forrest,‡ Susana Campino,* and Taane G. Clark*§

From the Faculty of Infectious and Tropical Diseases and the Faculty of Epidemiology and Population Health,§ London School of Hygiene and Tropical Medicine, London; the Section of Paediatric Infectious Disease,† Department of Infectious Disease, Imperial College London, London; and Wolfson College,‡ University of Cambridge, Cambridge, United Kingdom*

Accepted for publication August 16, 2022.

Address correspondence to Taane G. Clark, D.Phil., Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel St., London, United Kingdom.
E-mail: taane.clark@lshtm.ac.uk.

Recombinase polymerase amplification (RPA) is an isothermal amplification assay that has been ubiquitously utilized in the detection of infectious agents. Like any nucleic acid amplification technology, primer-template complementarity is critical to RPA reaction success. Mismatches arising in the primer-template complex are known to impact reaction kinetics, invalidate downstream analysis, such as nucleic acid quantification, and result in false negatives if used in a diagnostic capacity. Although the impact of specific primer-template mismatches has been well characterized for techniques such as PCR, characterization remains limited for RPA. Through our study, we systematically characterize the impact of mismatches on the RPA reaction, when located in the 3′-anchor region of the primer-template complex. Our investigation identified that the nucleotides involved, as well as position of each mismatch, influence the size of the impact, with terminal cytosine-thymine and guanine-adenine mismatches being the most detrimental. The presence of some mismatch combinations, such as a penultimate cytosine-cytosine and a terminal cytosine-adenine mismatch pairing, led to complete RPA reaction inhibition. Through the successful characterization of 315 mismatch combinations, researchers can optimize their RPA assay accordingly and seek to implement RPA technology for rapid, in-field genotyping. *(J Mol Diagn 2022, 24: 1207–1216; https://doi.org/10.1016/j.jmoldx.2022.08.005)*

Recombinase polymerase amplification (RPA) is an isothermal nucleic acid amplification technique (NAAT) that has been ubiquitously implemented in the detection of human and plant pathogens.[1,2] The RPA system relies on three T4 phage proteins, UvsX, UvsY, and Gp32. UvsX and UvsY facilitate priming of the DNA target through the assembly of a nucleoprotein filament, and Gp32 stabilizes the displaced single-stranded DNA during D-loop formation.[3] Together, this equates to the denaturation and primer annealing steps of a typical PCR cycle. RPA's performance at 37°C to 42°C makes it ideal for use in low-resource field settings, as demonstrated during the 2015 Ebola outbreak.[4] Unlike other common NAATs, such as PCR and loop-mediated isothermal amplification,[5–7] under certain conditions, RPA can be performed in the absence of a heat block,[5] highlighting its potential as the basis for future diagnostics.

RPA reaction success depends on robust primer design, like any NAAT.[8] This process balances several factors, which include the following: ensuring primer specificity by maximizing Watson-Crick nucleotide base pairing,[9] minimizing the potential of off-site binding to nontarget DNA, and minimizing primer-derived secondary structures, which can impede the reaction with varying degrees of severity.[10] Technique-specific primer design software has

Supported by Biotechnology and Biological Sciences Research Council (BBSRC) London Interdisciplinary Doctoral Programme (LIDo) PHD studentship (M.H.); Medical Research Council UK grants MR/M01360X/1, MR/R025576/1, and MR/R020973/1 (S.C.); Medical Research Council UK grants MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1 (T.G.C.); and BBSRC grant BB/R013063/1 (T.G.C.).
M.S.F., S.C., and T.G.C. contributed equally to this work.
Disclosures: TwistDX provided reagents necessary to conduct the study.

https://doi.org/10.1016/j.jmoldx.2022.08.005

Table 1 Loci of Interest

Gene	Gene orientation	Wild type (+/−)	Mutation (+/−)	SNPedia identifier	Description
cyp2c9	+	(A/T)	(C/G)	*rs1057910*	*cyp2c9* Encodes a member of the cytochrome P450 superfamily of enzymes, which is a key component in the xenobiotic metabolism of warfarin. This mutation has been shown to decrease enzyme activity, reducing rates of warfarin clearance and, as such, increasing sensitivity.[25]
cyp2c9	+	(G/C)	(A/T)	*rs4244285*	This mutation produces a nonfunctional truncated enzyme. Subsequently, the rate of warfarin clearance is reduced, and individuals have increased sensitivity.[25]
vkorc1	+	(C/G)	(G/C)	*rs8050894*	Warfarin inhibits the enzyme activity of vitamin K epoxide reductase complex C1 encoded by the *vkorc1* gene. The mutation highlighted confers resistance and, as such, a higher dose of warfarin is required for effective treatment.[27]
hbb	−	(T/A)	(A/T)	*rs334*	Carriers of the homozygous *TT* genotype develop sickle cell disease, increasing the underlying risk of blood clots. Low-dose warfarin treatment has been shown to reduce the risk of adverse effects linked to clotting.[28]

been developed, such as PrimedRPA, Primer3, and PrimerExplorer, to assist this task and overcome limitations associated with manual primer design.[8,11] However, the presence of unknown polymorphic sites in the primer binding region can compromise an NAAT reaction, causing nucleotide mismatches and reducing the stability of the primer-template complex.[12,13] This reduction is particularly an issue for organisms with high genetic diversity, limited genomic characterization, or error-prone replication systems.[14] Previous work has shown how mismatches can acutely disrupt NAAT amplification.[15,16] Initial reports highlighted that RPA has a high tolerance to polymorphisms in the primer/probe binding sites,[17] but subsequent work has identified that primer-template mismatches can impact RPA reaction efficiency and success, although they were unable to predict the impact based on the presence of a single mismatch.[18] Single or multiple mismatches located toward the 3′ terminus of the primer result in the most severe disruption for PCR, significantly reducing amplification efficiency and, in certain cases, preventing amplification altogether.[15] However, this phenomenon can be utilized to facilitate single-nucleotide polymorphism (SNP) genotyping, as demonstrated by amplification-refractory mutation system–PCR, kompetitive allele-specific PCR, and other techniques.[19,20] The genotyping of informative SNPs can personalize treatment choices and inform related fields, such as pharmacogenomics, where these biomarkers can be integrated into treatment decision pathways.[21] For example, several SNPs have been identified in metabolic genes that confer increased sensitivity or tolerance to the widely used warfarin anticoagulant drug,[22] and whose detection can inform the correct dosing in individuals with a high risk of thromboembolism, lowering the risk of adverse drug events due to underlying xenobiotic metabolism heterogeneity.[23–25]

This study aims to build on previous work and systematically characterize the impact of mismatches on the RPA reaction across four human genetic loci with clinical relevance to warfarin treatment. Two loci are situated within the *cyp2c9* gene, linked to warfarin clearance, whereas the other two loci are found within genes associated with altered warfarin dosage levels (*vkorc1* and *hbb*). The study centers on mismatches located in the primer anchor region, defined as the pre-ante-penultimate to 3′-terminal position, and attempts to build a mismatch classification system that can predict the impact of a given mismatch on RPA reaction success and kinetics. For this investigation, formamidopyrimidine DNA glycosylase (fpg) probes were utilized as, unlike the commonly used exo probes, they cannot act as extendable primers after cleavage and as such do not influence reaction kinetics. In addition, the fpg enzyme utilized for fpg probe cleavage has no 3′-5′ exonuclease activity, which could reduce the length of the primers removing mismatch loci under investigation. Understanding the impact on reaction kinetics is vital in determining whether the presence of a given mismatch will compromise techniques, which rely on the kinetic profile, such as RPA-based nucleic acid quantification.[26] Furthermore, this characterization could aid in the adaptation of RPA for in-field rapid SNP genotyping.

Materials and Methods

Chemicals and Oligonucleotide Design

All RPA reactions were performed using the TwistAmp fpg kit (TwistDx Ltd., Cambridge, UK). Oligonucleotide primers were sourced from Eurogentec (Seraing, Belgium) and TwistAmp fpg probes from LGC Biosearch Technologies (Petaluma, CA). Four nonsynonymous SNPs linked to warfarin metabolic changes were identified (Table 1). The double-stranded DNA templates housing each locus were procured from Twist Bioscience (San Francisco, CA) and subsequently diluted to the desired copy number in Tris-EDTA with 1 ng/μL poly(2′-deoxyinosinic-2′-

Table 2 dsDNA Templates

SNP	Sequence
>*rs105791*	5′-TTTAAGTTTGCATATACTTCCAGCACTATAATTTAAATTTATAATGATGTTTGGATACCTTCATGATTCATATACCCCTGA ATTGCTACAACAAATGTGCCATTTTTCTCCTTTTCCATCAGTTTTTACTTGTGTCTTATCAGCTAAAGTCCAGGAAGAGATT GAACGTGTGATTGGCAGAAACCGGAGCCCCTGCATGCAAGACAGGAGCCACATGCCCTACACAGATGCTGTGGTGCACGAGG TCCAGAGATACNTTGACCTTCTCCCCACCAGCCTGCCCCATGCAGTGACCTGTGACATTAAATTCAGAAACTATCTCATTCC CAAGGTAAGTTTGTTTCTCCTACACTGCAACTCCATGTTTTCGAAGTCCCCAAATTCATAGTATCATTTTTAAACCTCTACC ATCACCGGGTGAGAGAAGTGCATAACTCATATGTATGGCAGTTTAACTGGACTTTCTCTTGTTTCCAGTTTGGGGCTATAAA GGTTTGTAACAGGTCCTAGTGTCTGGCAGTGTGTGTTCTCCAGATTTATTATCTTTCTTCAAGATTGGTTTGGCTACTCTTA GGTGCTTATATTTCCAAATAATT-3′
>*rs334*	5′-GCACTTTCTTGCCATGAGCCTTCACCTTAGGGTTGCCCATAACAGCATCAGGAGTGGACAGATCCCCAAAGGACTCAAAGA ACCTCTGGGTCCAAGGGTAGACCACCAGCAGCCTAAGGGTGGGAAAATAGACCAATAGGCAGAGAGAGTCAGTGCCTATCAG AAACCCAAGAGTCTTCTCTGTCTCCACATGCCCAGTTTCTATTGGTCTCCTTAAACCTGTCTTGTAACCTTGATACCAACCT GCCCAGGGCCTCACCACCAACTTCATCCACGTTCACCTTGCCCCACAGGGCAGTAACGGCAGACTTCTCC**N**CAGGAGTCAGA TGCACCATGGTGTCTGTTTGAGGTTGCTAGTGAACACAGTTGTGTCAGAAGCAAATGTAAGCAATAGATGGCTCTGCCCTGA CTTTTATGCCCAGCCCTGGCTCCTGCCCTCCCTGCTCCTGGGAGTAGATTGGCCAACCCTAGGGTGTGGCTCCACAGGGTGA GGTCTAAGTGATGACAGCCGTACCTGTCCTTGGCTCTTCTGGCACTGGCTTAGGAGTTGGACTTCAAACCCTCAGCCCTCCC TCTAAGATATATCTCTTGGCCCCATACCATCAGTACAAATTGCTACTAAAAACATCCTCCTTTGCAAGTGTATTTACGTAAT ATTTGG-3′
>*rs424428*	5′-ACCATCTTATATTTCAAGATTGTAGAGAAGAATTGTTGTAAAAAGTAAGAGAATTAATATAAAGATGCTTTTATACTATCA AAAGCAGGTATAAGTCTAGGAAATGATTATCATCTTTGATTCTCTTGTCAGAATTTTCTTTCTCAAATCTTGTATAATCAGA GAATTACTACACATGTACAATAAAAATTTCCCCATCAAGATATACAATATATTTTATTTATATTTATAGTTTTAAATTACAA CCAGAGCTTGGCATATTGTATCTATACCTTTATTAAATGCTTTTAATTTAATAAATTATTGTTTTCTCTTAGATATGCAATA ATTTTCCCACTATCATTGATTATTTCCC**N**GGAACCCATAACAAATTACTTAAAAACCTTGCTTTTATGGAAAGTGATATTTT GGAGAAAGTAAAAGAACACCAAGAATCGATGGACATCAACAACCCTCGGGACTTTATTGATTGCTTCCTGATCAAAATGGAG AAGGTAAAATGTTAACAAAAGCTTAGTTATGTGACTGCTTGCGTATTTGTGATTCATTGACTAGTTTTGTGTTTACTACGGA TGTTTAACAGGTCAAGGAGTAATGCTTGAGAAGCATATTTAAGTTTTTATTGTATGCATGAATATCCAGTAAGCATCATAGA AAATGTAAAATTAAAT-3′
>*rs8050894*	5′-ACATGGCGAGACACCATCTCTACCAAAAAAAAACAAAAACAAAAATTAGCTGGGCATAGTGGTGCACGCCTGTGATTCCAG CTGCTTGGGAGGCTAAGGTGGGAGGATCCCTTGGGCAGGGAGGCAGAGGTTGCCATGAACTGAGATCACGCCAGTGCACACT AAGGGCATCCTAGACCTCACTTTGGGCAACAGAGCCAGACCCTGTCTCAAAACAACAACAAACAAAAAACCTGGGGACCTAG GATGTCTTTAAGGGCCCTTCAGCCTCTAACAGTACTTAAACCAATTAAAAGACTCCTGTTAGTTACCTCCCCACATCCCCAC CCGCAGGACGCTC**N**GTGATGAGCAGCTAGCTGGCTGTCAGCTGTGTGGATCACCAAGATTGCATGGAGTGGGGCTGAGCTGA CCAAGGGGGATGAGGGGCGGGGCGGGGCGGGCAGGGAGGGGGCGGAGCCACTCACCTAACAATAGCTGTAGTGTGTAGAAGA TGCAACCGAATATGCTGTTGGATTGATTGAGGATGCTGTCCTGTCCCAGCACATGCTCCACCAGCCCGAAACCCCTGCCCCA CCTGGCAGAGGGGTGGGGTGGGGTGGAACCAGGTTAGGACTGTCAACCCAGTGCCTTGGACCCTGCCCGAGAAAG-3′

These sequences were procured from Twist Bioscience. The N value (in bold) indicates the single-nucleotide polymorphism (SNP) site that was modified to generate four template variants per loci.
dsDNA, double-stranded DNA.

deoxycytidylic acid) sodium salt (Table 2). Seven assays were designed, targeting the four selected loci (Table 3). For each assay, the SNP was located in a primer binding region, with 52 dynamic primer variants generated through the exchange of one or two nucleotides from the pre-antepenultimate to the 3′ terminal position (Table 3). As such, the impact of a single and/or combined mismatches on amplification could be studied, while accounting for their relative position.

RPA Amplification

All reactions followed the recommended TwistAmp fpg protocol. A total of 600 nmol/L of forward and reverse primers, 120 nmol/L of probe, DNA template, 1× rehydration buffer, and DNAse-free water to a total volume of 47.5 μL were added to each lyophilized TwistAmp fpg pellet. A clean 2-mm bearing ball was then added to each tube to allow magnetic mixing to occur. Reactions were simultaneously initiated through the addition of 2.5 μL of 280 mmol/L magnesium acetate to the lids of the reaction tubes (strip of eight), capping the tubes carefully and spinning the magnesium acetate into the rehydrated material (total reaction volume 50 μL). Reactions were vortexed briefly and spun down once again before being placed into T8-ISO fluorescence readers manufactured by Axxin (Melbourne, VIC, Australia). Reactions were run at 39°C for 20 minutes, with readings taken every 10 to 15 seconds with an Opto PWM Duty FAM setting of 17 or 20.

Mismatch Characterization

For each assay, the reliable limit of detection (rLOD) was established (1000 to 5000 copies) in the absence of

Table 3 RPA Primers and Probe Groups Used in This Investigation

Group	Target	Sense	Role	Sequence
1	RS4244285	−	FP	5′-AAATTACAACCAGAGCTTGGCATATTGTATCTATA-3′
1	RS4244285	−	RP	5′-GCAAGGTTTTTAAGTAATTTGTTATGGGT**TCCN**-3′
1	RS4244285	−	Probe	5′-TCTTAGATATGCAATAATTTTCCCACT(dR[FAM])TCA(dT[BHQ-1])TGATTATTTCC-3′
2	RS1057910	−	FP	5′-ATCAGCTAAAGTCCAGGAAGAGATTGAACGTGTGA-3′
2	RS1057910	−	RP	5′-GCATGGGGCAGGCTGGTGGGGAGAAGGT**CAAN**-3′
2	RS1057910	−	Probe	5′-TGGCAGAAACCGGAGCCCCTGCATGCAA(dR[FAM])ACAG(dT[BHQ-1])AGCCACATG-3′
3	RS8050894	−	FP	5′-CTTCAGCCTCTAACAGTACTTAAACCAATTA-3′
3	RS8050894	−	RP	5′-CACACAGCTGACAGCCAGCTAGCTGCTCAT**CACN**-3′
3	RS8050894	−	Probe	5′-[FAM]AA[dR-BHQ-1]ACTCCTGTTAGTTACCTCCCCACATCC-3′
4	RS334	−	FP	5′-CATCTATTGCTTACATTTGCTTCTGACACAAC-3′
4	RS334	−	RP	5′-CCCACAGGGCAGTAACGGCAGACTTC**TCCN**-3′
4	RS334	−	Probe	5′-CAGGAGTCAGATGCACCATGGTGTCT(dR[FAM])TT(dT[BHQ-1])GAGGTTGCTAGT-3′
5	RS4244285	+	FP	5′-ATAATTTTCCCACTATCATTGATTATTT**CCCN**-3′
5	RS4244285	+	RP	5′-CTTTTGTTAACATTTTACCTTCTCCATTTTGAT-3′
5	RS4244285	+	Probe	5′-CACTTTCCATAAAAGCAAGGTTTTTAA(dR[FAM])TAA(dT[BHQ-1])TTGTTATGGGT-3′
6	RS1057910	+	FP	5′-AGATGCTGTGGTGCACGAGGTCCAGAGA**TACN**-3′
6	RS1057910	+	RP	5′-CAGTGTAGGAGAAACAAACTTACCTTGGGAATGAGA-3′
6	RS1057910	+	Probe	5′-TTAATGTCACAGGTCACTGCATGGGGCAGGCT(dR[FAM])G(dT[BHQ-1])GGGGAGAAGGT-3′
7	RS334	+	FP	5′-CAACCTCAAACAGACACCATGGTGCATCTGACTC**CTGN**-3′
7	RS334	+	RP	5′-GCCCAGTTTCTATTGGTCTCCTTAAACCTGTCTTG-3′
7	RS334	+	Probe	5′-CTGCCGTTACTGCCCTGTGGGGCAA(dR[FAM])G(dT[BHQ-1])GAACGTGGATGAA-3′

Bold indicates dynamic primer variants; N-terminal nucleotide covers loci of interest.
FP, forward primer; RP, reverse primer; RPA, recombinase polymerase amplification.

mismatches. The assessment of all following dynamic primers proceeded with two technical replicates against each relevant template variant. Primers resulting in a lone 3′-terminal mismatch were assessed against 1×, 10×, and 1000× the rLOD, whereas primers resulting in a lone internal mismatch were assessed at 1× the rLOD. Finally, primers that introduced multiple mismatches were assessed at 100× and 500× the rLOD. In each experiment, a primer with full complementarity to the target site was included to act as an internal standard and assessed at 1× the rLOD. Multiples of the rLOD were used in mismatch reactions as previous work has shown the introduction of mismatches reduced reaction sensitivity.[15,16] Mismatches were classified categorically according to the nucleotides present in the anchor region, disregarding complementary base-pairing positions. For example, the following primer (5′–CCCT′)−template (3′–GAGT) complex anchor region would be categorized as (?C?T-?A?T). In total, our data set covered 315 unique primer-template mismatch combinations.

Reaction Kinetic Profiling and Thermodynamic Calculations

The fluorescence profile of each reaction was extracted from the T8-ISO output, and a custom python script was used to normalize all reactions against their respective baseline (*https://github.com/MatthewHiggins2017/RPALogisticModelling*, last accessed August 25, 2022). Reaction success was established through a standard minimum fluorescence threshold criterion. Each successful reaction was modeled using a generalized logistic function (Richards' curve), via the Scipy python package (*https://scipy.org*, last accessed August 25, 2022). For each reaction, the time to positivity (TP), maximum gradient (MG), and time to maximum gradient (TMG) were derived (Figure 1). The 3′ Gibbs free energy was determined for each mismatch combination, according to the nearest-neighbor thermodynamic model using the full anchor region sequence. This model used prederived values, which are validated under crowding conditions present in the RPA

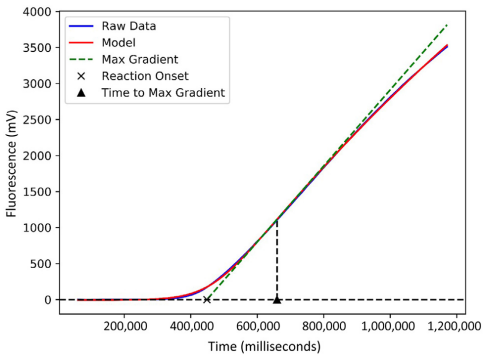


Figure 1 Modeling of the fluorescent curve after baseline normalization via a generalized logistic function and subsequent derivation of reaction kinetic metrics. The fluorescence profile corresponds to primer-probe group 6 ???C-???T terminal mismatch. Max, maximum.

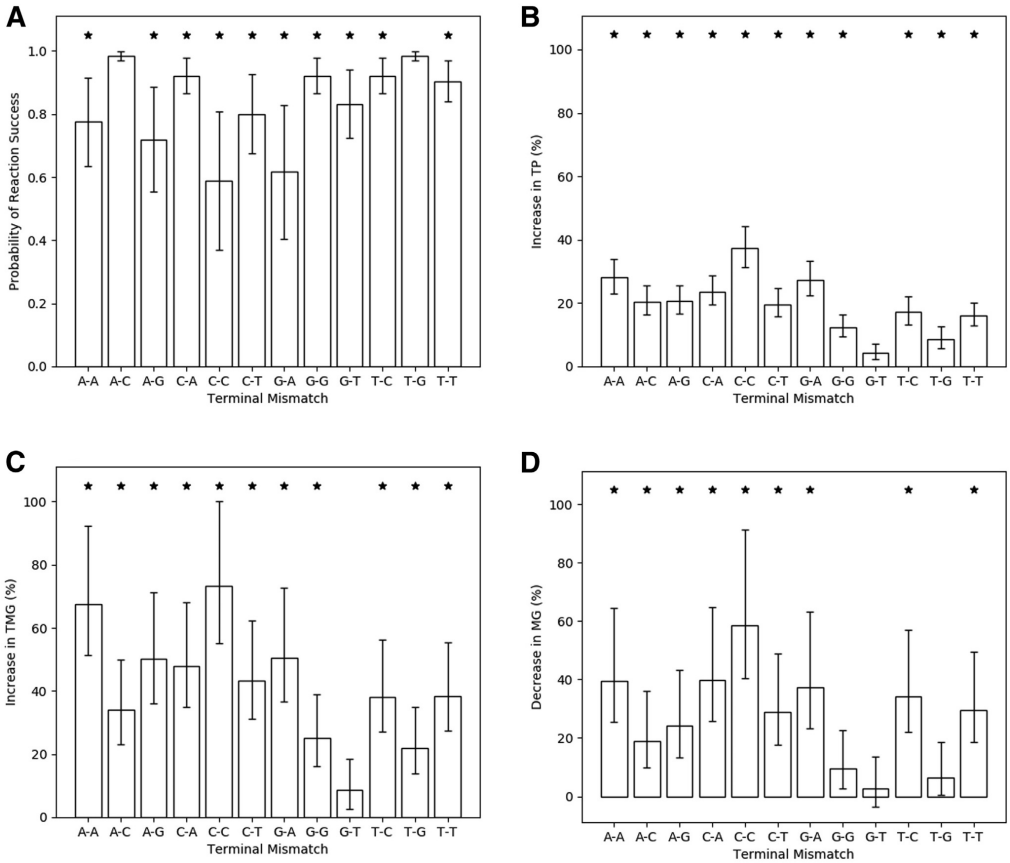


Figure 2 Impact of terminal mismatches on recombinase polymerase amplification reaction success (**A**); increase in time to positivity (TP; **B**); increase in time to maximum gradient (TMG; **C**); and decrease in maximum gradient (MG; **D**) when compared with primer-template complexes with complete complementarity. $^{*}P < 0.05$.

reaction.[29,30] Calculations were performed using a custom python script.

Statistical Analysis

To explore the impact of mismatch(es) on the odds of RPA reaction success, Firth logistic regression was implemented using the logistf package in R statistical software version 4.1.3 (*https://www.r-project.org*). This method was selected on the basis of its ability to handle complete separation events, such as a given mismatch combination inhibiting the RPA reaction across all experiments. In addition, Firth logistic regression can account for representational imbalances in the data set, which exist for primers that introduced two mismatches, due to limitations imposed by the nucleotides adjacent to the SNP of interest, which remained unadjusted in our double-stranded DNA templates. Area under the receiver-operating characteristic curve (AUC) analysis was implemented to assess model performance.

The impact of mismatches on successful reaction RPA kinetics (TP, MG, and TMG) was investigated using a robust linear mixed model, using the robustlmm R package.[31] The mixed model allowed us to account for the hierarchical data structure introduced as experiments were discretely clustered according to the primer-probe groups. Inclusion of a random effect variable for experiment accounted for human introduced variation, which could arise, including subtle time delays between the addition of magnesium acetate to start the reaction to placing the eight-tube strip in the fluorescence reader. Inclusion of a random effect variable for the target accounted for intrinsic

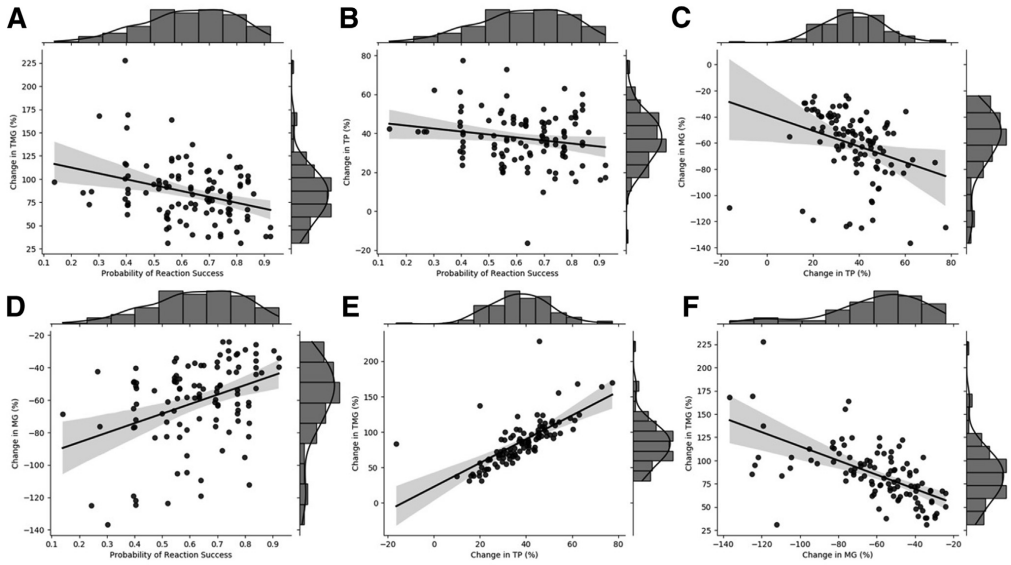


Figure 3 Spearman correlation plots for recombinase polymerase amplification reaction metrics of interest. For each comparison, linear regression was used to establish a line of best fit (solid black) and the associated 95% CI (shaded gray) **A:** Change in time to maximum gradient (TMG) versus probability of reaction success. **B:** Change in time to positivity (TP) versus probability of reaction success. **C:** Change in maximum gradient (MG) versus change in TP. **D:** Change in MG versus probability of reaction success. **E:** Change in TMG versus change in TP. **F:** Change in TMG versus change in MG.

performance differences between the seven assays, which could be linked to amplicon length, secondary structure formation, and a range of other factors beyond the scope of this investigation. The fixed variable, Opto PWM Duty FAM, was included in the model to account for calibration differences between the T8-ISO fluorescence readers utilized. A robust approach was chosen because of the heavy tailed residual distribution, ensuring all assumptions associated with statistical inference were met. To assess mixed model performance, conditional R^2 values were derived according to Nakagawa and Schielzeth.[32] In all models, the baseline level of the mismatch categorical variable (????-????) represented complete complementary binding of the anchor region, allowing us to compare the impact of primer-template mismatches on the RPA reaction against primer binding with complete complementarity.

Results

First Models for RPA Reaction Success and Kinetics

Across 501 experiments covering 315 unique mismatch combinations, a total of 3543 reactions of 4008 were classified as successful (88.4%). The model established to investigate mismatch impact on reaction success achieved an AUC of 0.88 (Supplemental Table S1 provides estimated coefficients). More than 150 mismatch combinations (compared with ????-????) were identified to have a significant impact on reaction success (159 with $P < 0.05$), representing 50.4% of all mismatches investigated. The double-stranded DNA template copy number was found not to have a significant impact on the probability of reaction success ($P = 0.549$). The models derived for the TP, MG, and TMG achieved conditional (and adjusted for model size) R^2 values of 0.867 (0.854), 0.862 (0.850), and 0.800 (0.781), respectively. The template copy number had a significant impact across all three reaction kinetic metrics (TP, MG, and TMG), where a unit increase in the copy

Table 4 Spearman Correlation Values Obtained between Different RPA Reaction Metrics

Metric 1	Metric 2	ρ	*P* value
Probability of reaction success	Change in TP (%)	−0.138	1.60×10^{-1}
Probability of reaction success	Change in MG (%)	0.357	1.72×10^{-4}
Change in TP (%)	Change in MG (%)	−0.489	1.04×10^{-7}
Change in TP (%)	Change in TMG (%)	0.823	2.61×10^{-27}
Probability of reaction success	Change in TMG (%)	−0.260	7.23×10^{-3}
Change in MG (%)	Change in TMG (%)	−0.649	5.41×10^{-14}

MG, maximum gradient; RPA, recombinase polymerase amplification; TMG, time to maximum gradient; TP, time to positivity.

Table 5 Reclassification of Primer-Template Complex to Investigate Positional Impact on RPA Reaction Success and Kinetics

New classification	Original classification	New classification sample size
P	????-????	1002
T	???A-???G	510
1n	??A?-??G?	186
2n	?A??-?G??	192
3n	A???-G???	192
T1n	??AA-??GG	624
T2n	?A?A-?G?G	630
T3n	A??A-G??G	672

RPA, recombinase polymerase amplification.

number resulted in a decrease in TP and TMG, while increasing the MG ($P < 0.001$) (Supplemental Table S1). Overall, 252, 188, and 250 mismatches (compared with ????-????) had a significant ($P < 0.05$) impact on TP, MG, and TMG, respectively. In total, 106 mismatches significantly impacted all the reaction kinetics metrics, as well as the probability of reaction success.

The Impact of Primer-Template Mismatch Constituents on RPA Reaction Kinetics

To investigate whether the constituents of a given primer-template mismatch alter the impact on RPA kinetic profile, the impact of single terminal mismatches was studied. This impact was assessed when expressing model coefficients relative to primer-template complexes with complete complementary binding (Figure 2). Of the 12 possible terminal mismatches, 8 resulted in a significant difference ($P < 0.05$) across all reaction metrics, and a quarter of all possible terminal mismatches resulted in the probability of reaction success decreasing below 0.8. The impact of a given mismatch deviates depending on the nucleotide constituents. The cytosine-cytosine terminal mismatches (primer-template) appear to be most detrimental to the probability of reaction success (0.59), closely followed by the guanine-adenine terminal mismatch (0.62). This pattern continues for the increase in TP for cytosine-cytosine (37.3%) and guanine-adenine (27.4%) terminal mismatches. The cytosine-cytosine terminal mismatch also results in the largest decrease in MG (58.4%) and increase in TMG (73.2%). Only two terminal mismatches do not significantly impact the probability of reaction success, thymine-guanine (0.98) and adenine-cytosine (0.98). The only terminal mismatch to not significantly affect any of the reaction kinetics is guanine-thymine (primer-template), but this was found to have a significant impact on the probability of reaction success (0.83). The pairwise Spearman correlation between each reaction metric was determined, considering only those mismatches that resulted in a significant impact across all metrics (Figure 3 and Table 4). A significant pairwise correlation ($P < 0.05$) was found between all metrics apart from the probability of reaction success versus a change in TP. The strongest correlation (ρ: 0.823) was observed between the change in TP and TMG metrics, which is to be expected.

The Impact of Primer-Template Mismatch Location on RPA Reaction Success and Kinetics

Next, the detrimental effect of a primer-template mismatch according to the relative positioning of mismatches in the anchor region was studied. Primer-template complexes were categorized into eight groups (T, 1n, 2n, 3n, T1n, T2n, and T3n; compared with P − ????-????) (Table 5). With this new classification system, models for each metric were fitted (Supplemental Table S2 provides estimated coefficients). The updated model for the probability of reaction success achieved an AUC of 0.72, whereas the updated mixed models for TP, MG, and TMG obtained conditional (and adjusted) R^2 values of 0.792 (0.791), 0.826 (0.826), and 0.656 (0.655), respectively. The predictive performance of the updated models is inferior to those including unique mismatch combinations, but they contain fewer parameters. From a generalized standpoint, any mismatches positioned in the anchor region may have a significant impact on the probability of reaction success (Figure 4). When considering the positional impact of lone primer-template mismatches (T, 1n, 2n, and 3n), those located at the terminal position are the most detrimental, resulting in a 17.7% increase in TP, a 19.4% decrease in MG, and a 51.7% increase in TMG, with a probability of reaction success of 0.877 (Figure 4). Lone mismatches located outside of the terminal position do not have a significant impact on TP, whereas those located in the 1n and 3n positions do significantly ($P < 0.05$) impact MG (9.7% and 15.3%) and TMG (13.2% and 11.1%), respectively. The presence of two mismatches is more detrimental across all RPA reaction metrics than the presence of a lone mismatch, regardless of position. For TP and MG, the size of the impact decreases as the distance between the secondary mismatch and terminal mismatch grows. The impact on TP decreases from 34.17% to 27.35% and on MG from 40.13% to 32.10% for T1n and T3n, respectively. However, this is not the case for reaction success and TMG, where T3n and T1n appear to be the most detrimental, respectively.

The Impact of Anchor Region Stability on RPA Reaction Success and Kinetics

To determine whether the anchor region stability could be used to predict changes in RPA reaction success and kinetics, the Gibbs free energy for each primer-template mismatch was estimated. The thermodynamic potential reflects the stability of the primer-template complex anchor region and is currently a selection feature in popular primer design software (eg, PrimerExplorer and Primer3). Primer-

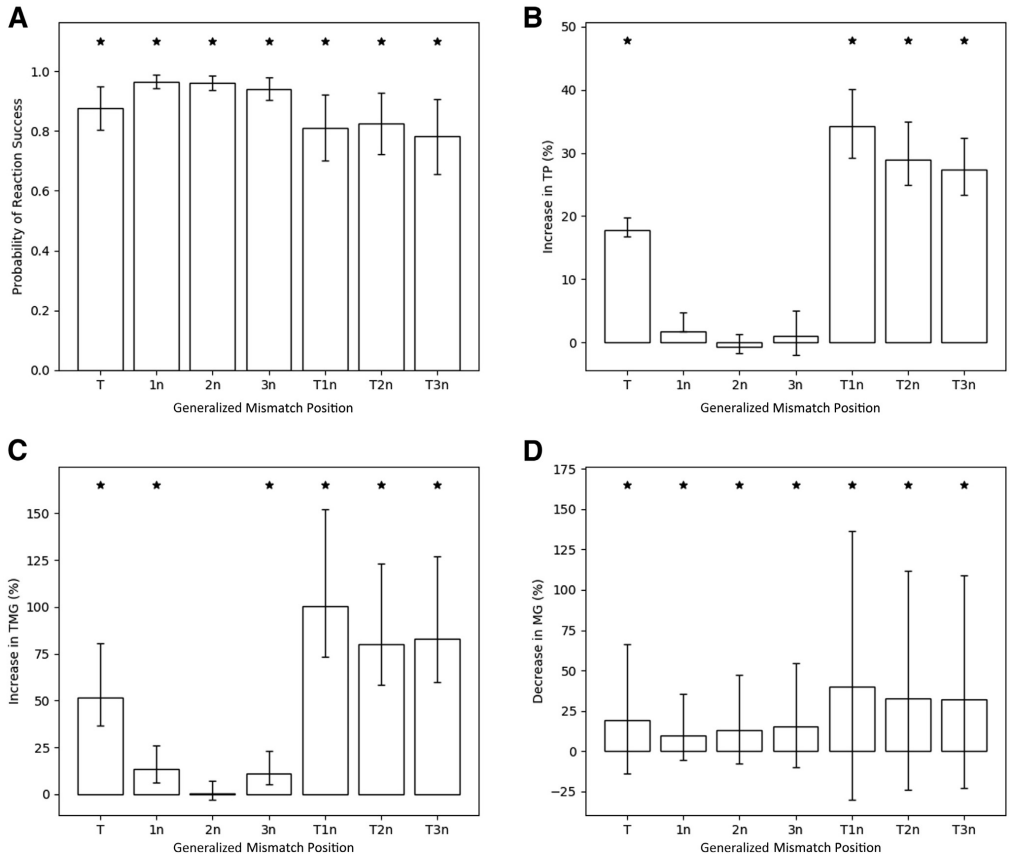


Figure 4 Positional impact of mismatches on probability of reaction success (**A**); increase in time to positivity (TP; **B**); increase in time to maximum gradient (TMG; **C**); and decrease in maximum gradient (MG; **D**) when compared with primer-template complexes with complete complementarity. $^*P < 0.05$.

template complexes with mismatches in the terminal and penultimate positions were excluded because of limitations of values available used to generate the nearest-neighbor thermodynamic model. The Gibbs free energy values were used to replace the categorical mismatch values in revised models for each metric (Supplemental Table S3 provides estimated coefficients). The AUC obtained for the probability of reaction success was 0.734, whereas the conditional (and adjusted) R^2 values for TP, MG, and TMG were 0.658 (0.658), 0.857 (0.857), and 0.550 (0.550), respectively. Across all RPA reaction metrics, the Gibbs free energy variable had a significant impact ($P < 0.05$), where a unit increase in energy resulted in an increase in TMG and TP and a decrease in MG (Supplemental Table S3). These directions of effect were expected as an increase in the Gibbs free energy represents a decrease in anchor region stability.

To enable an accurate comparison to prior mismatch classification models, the positional-based classification models were refitted, excluding the T1n data (Supplemental Table S4). For the updated positional models, the probability of reaction success AUC was 0.748, whereas the R^2 values for TP, MG, and TMG were 0.711 (0.710), 0.874 (0.874), and 0.600 (0.599), respectively. By comparing measures of model fit, utilizing adjusted-conditional R^2 metric, the performance of using Gibbs free energy to the positional-based mismatch classification is similar. However, across all metrics, positional-based mismatch classification narrowly outperforms models that include the Gibbs free energy. In addition, using positional-based classification allows combined mismatches in the penultimate and terminal position of the primer-template mismatch complex to be accounted for.

Discussion

Our investigation has shown the detrimental impact of primer-template mismatches on RPA amplification when located toward the 3′ primer terminus. To the best of our knowledge, this is the first study to systematically explore RPA kinetics via modeling the reaction fluorescence profile using a generalized logistic function. As expected, when classifying mismatches according to position, the presence of multiple mismatches resulted in a greater impact on the RPA reaction, compared with the presence of a lone mismatch. Multiple mismatches, positioned adjacently in the 3′ primer terminal and penultimate position, were found to be the most detrimental across all reaction kinetic metrics. Our analysis reveals that the impact of a given mismatch combination is not only dependent on the relative position in the anchor region but also the nucleotides involved. Most mismatch combinations significantly impacted at least one reaction measure, with just over a third of mismatches (106/315) significantly impacting all metrics considered. Specifically, a terminal cytosine-cytosine mismatch was the most detrimental to the RPA reaction efficiency, followed by a guanine-adenine. However, adenine-cytosine and thymine-guanine mismatches were highly tolerated, rarely resulting in amplification failure, mirroring the impact these mismatches have in PCR.[15]

Characterizing the stability of the primer-template complex anchor region via the gold standard nearest-neighbor approach did not outperform positional classification. This result suggests that the position and nucleotides involved in a particular mismatch are more informative than the stability of the primer-template complex. Such an insight aligns with our current understanding of polymerase fidelity and the concept of active site tightness, which highlights the nucleotides involved in a mismatch govern the impact on the polymerase due to differences in steric hindrances.[33] Further research is required to quantify the steric hindrance induced by different mismatch combinations and, subsequently, if this quantifiable parameter can be used to predict the impact of a given mismatch on RPA reaction kinetics.

Our investigation highlights the importance of considering variation in primer binding sites for RPA diagnostic applications, as a single mismatch has the potential to reduce the probability of reaction success to 0.589, compromising both sensitivity and specificity. Addressing the impact of mismatch on reaction success is a potential issue for robust SNP profiling. The introduction of specific mismatches needs to completely inhibit the reaction, enabling binary classification to indicate the presence or absence of a particular genotype. Within the scope of our investigation, the deliberate introduction of certain T1n mismatch combinations, such as ??CA-??CC and ??TC-??CC, completely inhibited the RPA reaction, whereas the corresponding single penultimate mismatches, ??C?-??C? and ??T?-??C?, only mildly retarded the RPA reaction kinetics. Alternatively, the RPA reaction could be designed to guarantee reaction success in the presence of mismatches, while maintaining heterogeneity in reaction kinetics, which could be used for SNP classification. The strong correlation between reaction kinetic metrics, such as TP and TMG, could be used to enhance the feasibility and reliability of a metric clustering approach to determine SNP presence. To achieve the desired changes in mismatch impact on reaction success, a variety of strand-displacing polymerases should be screened to identify those more sensitive and tolerant to mismatch combinations.

In summary, we have systematically investigated the impact of terminal mismatches on the RPA reaction across several clinically relevant biomarkers. Through implementing a range of statistical models, we have determined the impact of 315 mismatch combination on the RPA reaction, highlighting to RPA users the pitfalls of bad primer design and proving a foundation on which to build for RPA-based SNP genotyping. We hope that our description on RPA mismatch tolerance will form the foundation of improved RPA primer design using computational programs, such as PrimedRPA, aiding in the design of robust assays, especially in targets with high genetic diversity. The implementation of RPA-based SNP genotyping and diagnostics for infections and diseases, especially in high burden populations, has the potential to inform clinical and surveillance decision making, leading to personalized treatment of patients with improved outcomes and healthier populations.

Acknowledgment

We thank TwistDX for providing reagents necessary to conduct the study.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2022.08.005*.

References

1. El Wahed AA, Patel P, Maier M, Pietsch C, Rüster D, Böhlken-Fascher S, Kissenkötter J, Behrmann O, Frimpong M, Diagne MM, Faye M, Dia N, Shalaby MA, Amer H, Elgamal M, Zaki A, Ismail G, Kaiser M, Corman VM, Niedrig M, Landt O, Faye O, Sall AA, Hufert FT, Truyen U, Liebert UG, Weidmann M: Suitcase lab for rapid detection of SARS-CoV-2 based on recombinase polymerase amplification assay. Anal Chem 2021, 93:2627−2634
2. Babu B, Ochoa-Corona FM, Paret ML: Recombinase polymerase amplification applied to plant virus detection and potential implications. Anal Biochem 2018, 546:72−77
3. Piepenburg O, Williams CH, Stemple DL, Armes NA: DNA detection using recombination proteins. PLoS Biol 2006, 4:e204
4. Yang M, Ke Y, Wang X, Ren H, Liu W, Lu H, Zhang W, Liu S, Chang G, Tian S, Wang L, Huang L, Liu C, Yang R, Chen Z:

Development and evaluation of a rapid and sensitive EBOV-RPA test for rapid diagnosis of Ebola virus disease. Sci Rep 2016, 6:26943
5. Crannell ZA, Rohrman B, Richards-Kortum R: Equipment-free incubation of recombinase polymerase amplification reactions using body heat. PLoS One 2014, 9:e112146
6. Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, Amino N, Hase T: Loop-mediated isothermal amplification of DNA. Nucleic Acids Res 2000, 28:E63
7. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY: A polymorphic DNA marker genetically linked to Huntington's disease. Nature 1983, 306:234–238
8. Higgins M, Ravenhall M, Ward D, Phelan J, Ibrahim A, Forrest MS, Clark TG, Campino S: PrimedRPA: primer design for recombinase polymerase amplification assays. Bioinformatics 2019, 35:682–684
9. Watson JD, Crick FH: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953, 171:737–738
10. Lobato IM, O'Sullivan CK: Recombinase polymerase amplification: basics, applications and recent advances. Trends Analyt Chem 2018, 98:19–35
11. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM: Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 2007, 35:W71–W74
12. Ziegler K, Steininger P, Ziegler R, Steinmann J, Korn K, Ensser A: SARS-CoV-2 samples may escape detection because of a single point mutation in the N gene. Euro Surveill 2020, 25:2001650
13. Allawi HT, SantaLucia J Jr: Thermodynamics of internal C.T mismatches in DNA. Nucleic Acids Res 1998, 26:2694–2701
14. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R: Extremely high mutation rate of HIV-1 in vivo. PLoS Biol 2015, 13: e1002251
15. Stadhouders R, Pas SD, Anber J, Voermans J, Mes THM, Schutten M: The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5′ nuclease assay. J Mol Diagn 2010, 12:109–117
16. Kwok S, Kellogg DE, McKinney N, Spasic D, Goda L, Levenson C, Sninsky JJ: Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. Nucleic Acids Res 1990, 18:999–1005
17. Boyle DS, Lehman DA, Lillis L, Peterson D, Singhal M, Armes N, Parker M, Piepenburg O, Overbaugh J: Rapid detection of HIV-1 proviral DNA for early infant diagnosis using recombinase polymerase amplification. MBio 2013, 4:e00135-13
18. Daher RK, Stewart G, Boissinot M, Boudreau DK, Bergeron MG: Influence of sequence mismatches on the specificity of recombinase polymerase amplification technology. Mol Cell Probes 2015, 29:116–121
19. Little S: Amplification-refractory mutation system (ARMS) analysis of point mutations. Curr Protoc Hum Genet 2001. Chapter 9: Unit 9.8
20. Semagn K, Babu R, Hearne S, Olsen M: Single nucleotide polymorphism genotyping using kompetitive allele specific PCR (KASP): overview of the technology and its application in crop improvement. Mol Breed 2014, 33:1–14
21. Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, Peterson JF, Van Driest SL: Pharmacogenomics. Lancet 2019, 394: 521–532
22. Pirmohamed M: Warfarin: almost 60 years old and still causing problems. Br J Clin Pharmacol 2006, 62:509–511
23. Budnitz DS, Lovegrove MC, Shehab N, Richards CL: Emergency hospitalizations for adverse drug events in older Americans. N Engl J Med 2011, 365:2002–2012
24. Johnson JA, Cavallari LH: Warfarin pharmacogenetics. Trends Cardiovasc Med 2015, 25:33–41
25. Perera MA, Cavallari LH, Limdi NA, Gamazon ER, Konkashbaev A, Daneshjou R, et al: Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. Lancet 2013, 382:790–796
26. Crannell ZA, Rohrman B, Richards-Kortum R: Development of a quantitative recombinase polymerase amplification assay with an internal positive control. J Vis Exp 2015:52620
27. AL-Eitan LN, Almasri AY, Khasawneh RH: Effects of CYP2C9 and VKORC1 polymorphisms on warfarin sensitivity and responsiveness during the stabilization phase of therapy. Saudi Pharm J 2019, 27: 484–490
28. Ahmed S, Siddiqui AK, Iqbal U, Sison CP, Shahid RK, Sheth M, Patel DV, Russo LA: Effect of low-dose warfarin on D-dimer levels during sickle cell vaso-occlusive crisis: a brief report. Eur J Haematol 2004, 72:213–216
29. Ghosh S, Takahashi S, Endoh T, Tateishi-Karimata H, Hazra S, Sugimoto N: Validation of the nearest-neighbor model for Watson–Crick self-complementary DNA duplexes in molecular crowding condition. Nucleic Acids Res 2019, 47:3284–3294
30. SantaLucia J Jr: A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A 1998, 95:1460–1465
31. Koller M: robustlmm: An R package for robust estimation of linear mixed-effects models. J Stat Softw 2016, 75:1–24
32. Nakagawa S, Schielzeth H: A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol Evol 2013, 4:133–142
33. Kool ET: Active site tightness and substrate fit in DNA replication. Annu Rev Biochem 2002, 71:191–219

Supplementary Information

dsDNA Templates

The following sequences were procured from Twist Bioscience. The N value indicates the SNP site which was modified to create four template variants per loci.

> rs105791

TTTAAGTTTGCATATACTTCCAGCACTATAATTTAAATTTATAATGATGTTTGGATACCTTCATGATTCATATACCCCTGAATTGCTACAACAAATGTGCCATTTTTCTCCTTTTCCATCAGTTTTTACTTGTGTCTTAT
CAGCTAAAGTCCAGGAAGAGATTGAACGTGTGATTGGCAGAAACCGGAGCCCCTGCATGCAAGACAGGAGCCACATGCCCTACACAGATGCTGTGGTGCACGAGGTCCAGAGATACNTTGACCTTCTCCCCACCA
GCCTGCCCCATGCAGTGACCTGTGACATTAAATTCAGAAACTATCTCATTCCCAAGGTAAGTTTGTTTCTCCTACACTGCAACTCCATGTTTTCGAAGTCCCCAAATTCATAGTATCATTTTTAAACCTCTACCATCA
CCGGGTGAGAGAAGTGCATAACTCATATGTATGGCAGTTTAACTGGACTTTCTCTTGTTTCCAGTTTGGGGCTATAAAGGTTTGTAACAGGTCCTAGTGTCTGGCAGTGTGTGTTCTCCAGATTTATTATCTTTCTTC
AAGATTGGTTTGGCTACTCTTAGGTGCTTATATTTCCAAATAATT

> rs334

GCACTTTCTTGCCATGAGCCTTCACCTTAGGGTTGCCCATAACAGCATCAGGAGTGGACAGATCCCCAAAGGACTCAAAGAACCTCTGGGTCCAAGGGTAGACCACCAGCAGCCTAAGGGTGGGAAAATAGACCA
ATAGGCAGAGAGAGTCAGTGCCTATCAGAAACCCAAGAGTCTTCTCTGTCTCCACATGCCCAGTTTCTATTGGTCTCCTTAAACCTGTCTTGTAACCTTGATACCAACCTGCCCAGGGCCTCACCACCAACTTCATC
CACGTTCACCTTGCCCCACAGGGCAGTAACGGCAGACTTCTCCNCAGGAGTCAGATGCACCATGGTGTCTGTTTGAGGTTGCTAGTGAACACAGTTGTGTCAGAAGCAAATGTAAGCAATAGATGGCTCTGCCCTG
ACTTTTATGCCCAGCCCTGGCTCCTGCCCTCCCTGCTCCTGGGAGTAGATTGGCCAACCCTAGGGTGTGGCTCCACAGGGTGAGGTCTAAGTGATGACAGCCGTACCTGTCCTTGGCTCTTCTGGCACTGGCTTAGG
AGTTGGACTTCAAACCCTCAGCCCTCCCTCTAAGATATATCTCTTGGCCCCATACCATCAGTACAAATTGCTACTAAAAACATCCTCCTTTGCAAGTGTATTTACGTAATATTTGG

> rs424428

ACCATCTTATATTTCAAGATTGTAGAGAAGAATTGTTGTAAAAAGTAAGAGAATTAATATAAAGATGCTTTTATACTATCAAAAGCAGGTATAAGTCTAGGAAATGATTATCATCTTTGATTCTCTTGTCAGAATTT
TCTTTCTCAAATCTTGTATAATCAGAGAATTACTACACATGTACAATAAAAATTTCCCCATCAAGATATACAATATATTTTATTTATATTTATAGTTTTAAATTACAACCAGAGCTTGGCATATTGTATCTATACCTTT
ATTAAATGCTTTTAATTTAATAAATTATTGTTTTCTCTTAGATATGCAATAATTTTCCCACTATCATTGATTATTTCCCNGGAACCCATAACAAATTACTTAAAAACCTTGCTTTTATGGAAAGTGATATTTTGGAGAA
AGTAAAAGAACACCAAGAATCGATGGACATCAACAACCCTCGGGACTTTATTGATTGCTTCCTGATCAAAATGGAGAAGGTAAAATGTTAACAAAAGCTTAGTTATGTGACTGCTTGCGTATTTGTGATTCATTGA
CTAGTTTTGTGTTTACTACGGATGTTTAACAGGTCAAGGAGTAATGCTTGAGAAGCATATTTAAGTTTTTATTGTATGCATGAATATCCAGTAAGCATCATAGAAAATGTAAAATTAAAT

> rs8050894

ACATGGCGAGACACCATCTCTACCAAAAAAAAAACAAAAACAAAAATTAGCTGGGCATAGTGGTGCACGCCTGTGATTCCAGCTGCTTGGGAGGCTAAGGTGGGAGGATCCCTTGGGCAGGGAGGCAGAGGTTGC
CATGAACTGAGATCACGCCAGTGCACACTAAGGGCATCCTAGACCTCACTTTGGGCAACAGAGCCAGACCCTGTCTCAAAACAACAACAAACAAAAAACCTGGGGACCTAGGATGTCTTTAAGGGCCCTTCAGCC
TCTAACAGTACTTAAACCAATTAAAAGACTCCTGTTAGTTACCTCCCCACATCCCCACCCGCAGGACGCTCNGTGATGAGCAGCTAGCTGGCTGTCAGCTGTGTGGATCACCAAGATTGCATGGAGTGGGGCTGAG
CTGACCAAGGGGGATGAGGGGCGGGGCGGGGCGGGCAGGGAGGGGGCGGAGCCACTCACCTAACAATAGCTGTAGTGTGTAGAAGATGCAACCGAATATGCTGTTGGATTGATTGAGGATGCTGTCCTGTCCCA
GCACATGCTCCACCAGCCCGAAACCCCTGCCCCACCTGGCAGAGGGGTGGGGTGGGGTGGAACCAGGTTAGGACTGTCAACCCAGTGCCTTGGACCCTGCCCGAGAAAG

Supplementary Table 1. Coefficients obtained when generalising model according to mismatch position and nucleotides involved.

	TMG		MG		TP		RS	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Log-Odds*	*CI*
(Intercept)	330627.10 ***	195611.47 – 465642.73	15687.49 ***	15687.48 – 15687.50	399899.43 ***	314657.58 – 485141.27	4.79 ***	3.72 – 5.95
???A-???A	222840.63 ***	180816.04 – 264865.22	-6211.07 ***	-6211.07 – -6211.06	112112.30 ***	90753.67 – 133470.93	-3.55 ***	-4.71 – -2.42
???A-???C	112442.86 ***	71996.48 – 152889.24	-2976.08 ***	-2976.08 – -2976.08	82062.97 ***	61463.23 – 102662.71	-0.66	-2.84 – 4.22
???A-???G	165933.74 ***	122688.76 – 209178.71	-3790.75 ***	-3790.75 – -3790.75	82324.47 ***	60356.11 – 104292.82	-3.84 ***	-4.97 – -2.78
???C-???A	158801.43 ***	118027.97 – 199574.88	-6225.35 ***	-6225.36 – -6225.35	94595.30 ***	73839.76 – 115350.84	-2.32 *	-3.74 – -0.59
???C-???C	242133.99 ***	197812.11 – 286455.86	-9159.97 ***	-9159.97 – -9159.97	148997.26 ***	126531.81 – 171462.72	-4.43 ***	-5.51 – -3.44
???C-???T	143096.31 ***	103437.64 – 182754.97	-4505.54 ***	-4505.54 – -4505.54	78667.89 ***	58505.76 – 98830.03	-3.40 ***	-4.55 – -2.27
???G-???A	167423.56 ***	119930.79 – 214916.32	-5830.96 ***	-5830.96 – -5830.95	109622.98 ***	85546.26 – 133699.70	-4.31 ***	-5.43 – -3.27
???G-???G	82974.67 ***	42072.69 – 123876.65	-1510.25 ***	-1510.25 – -1510.25	49521.17 ***	28701.13 – 70341.20	-2.32 *	-3.74 – -0.59
???G-???T	28383.63	-13354.45 – 70121.72	-402.74 ***	-402.74 – -402.73	16765.61	-4460.23 – 37991.44	-3.19 ***	-4.40 – -1.94
???T-???C	126320.89 ***	85499.23 – 167142.54	-5346.72 ***	-5346.72 – -5346.72	68835.34 ***	48055.63 – 89615.05	-2.32 *	-3.74 – -0.59
???T-???G	72241.43 ***	31795.05 – 112687.81	-1009.39 ***	-1009.39 – -1009.38	34702.74 ***	14103.00 – 55302.48	-0.66	-2.84 – 4.22
???T-???T	126932.93 ***	88481.24 – 165384.62	-4638.55 ***	-4638.55 – -4638.55	63977.98 ***	44412.23 – 83543.73	-2.54 **	-3.85 – -1.06
??A?-??C?	8882.93	-57928.31 – 75694.16	-1341.93 ***	-1341.93 – -1341.93	-674.04	-34231.69 – 32883.62	-2.25	-4.53 – 2.65
??A?-??G?	56525.29 ***	27917.45 – 85133.12	-1383.82 ***	-1383.82 – -1383.82	28014.04 ***	13656.50 – 42371.58	-2.50 **	-3.80 – -1.04
??AA-??CA	169873.39 *	36619.24 – 303127.54	-2965.71 ***	-2965.71 – -2965.70	69210.28 *	2221.82 – 136198.74	-3.55	-6.23 – 1.43
??AA-??CC	258699.50 ***	124768.94 – 392630.07	-6439.56 ***	-6439.56 – -6439.55	146923.05 ***	79566.88 – 214279.22	-3.55	-6.23 – 1.43
??AA-??CG	194654.38 **	60723.81 – 328584.95	-4188.56 ***	-4188.56 – -4188.55	124597.53 ***	57241.35 – 191953.70	-3.55	-6.23 – 1.43
??AA-??GA	295093.80 ***	226821.90 – 363365.70	-8282.45 ***	-8282.45 – -8282.44	162917.48 ***	128822.25 – 197012.71	-5.17 ***	-6.56 – -3.84
??AA-??GC	412508.11 ***	357018.65 – 467997.58	-11421.03 ***	-11421.03 – -11421.03	251989.72 ***	224148.46 – 279830.98	-3.56 ***	-5.11 – -1.74
??AA-??GG	317702.59 ***	254706.82 – 380698.35	-10423.95 ***	-10423.96 – -10423.95	201884.60 ***	170309.04 – 233460.16	-4.62 ***	-6.02 – -3.22
??AC-??CA	120630.46	-12623.68 – 253884.61	-2837.47 ***	-2837.48 – -2837.46	78008.32 *	11019.86 – 144996.77	-3.55	-6.23 – 1.43
??AC-??CC	170336.36 *	2770.43 – 337902.30	-8462.80 ***	-8462.81 – -8462.79	105696.15 *	22104.07 – 189288.22	-5.16 ***	-7.80 – -2.53
??AC-??CT	99514.53	-33739.62 – 232768.67	-2585.94 ***	-2585.95 – -2585.93	56210.05	-10778.41 – 123198.50	-3.55	-6.23 – 1.43
??AC-??GA	237528.43 ***	184934.03 – 290122.84	-6304.66 ***	-6304.67 – -6304.66	148012.16 ***	121578.31 – 174446.02	-1.8	-4.07 – 3.10
??AC-??GC	343007.13 ***	284922.43 – 401091.83	-5627.18 ***	-5627.19 – -5627.18	241119.16 ***	211964.02 – 270274.31	-3.13 *	-4.87 – -0.78
??AC-??GT	249395.30 ***	192891.77 – 305898.83	-8447.99 ***	-8448.00 – -8447.99	132495.58 ***	104101.25 – 160889.90	-3.56 ***	-5.12 – -1.75
??AG-??CA	-92735.07	-225989.22 – 40519.08	-147.69 ***	-147.70 – -147.68	-30419.65	-97408.10 – 36568.81	-3.55	-6.23 – 1.43
??AG-??CG	401099.27 ***	267168.70 – 535029.84	-9331.04 ***	-9331.05 – -9331.04	156405.96 ***	89049.79 – 223762.14	-3.55	-6.23 – 1.43
??AG-??CT	80065.28	-53188.87 – 213319.43	-3665.87 ***	-3665.88 – -3665.87	47588.59	-19399.86 – 114577.05	-3.55	-6.23 – 1.43
??AG-??GA	250757.12 ***	184108.79 – 317405.46	-5700.18 ***	-5700.18 – -5700.17	172776.95 ***	139365.45 – 206188.44	-3.95 ***	-5.56 – -2.10
??AG-??GG	215356.05 ***	153842.54 – 276869.57	-3767.40 ***	-3767.40 – -3767.40	137503.24 ***	106588.18 – 168418.31	-3.73 ***	-5.31 – -1.90
??AG-??GT	160914.68 ***	100290.27 – 221539.08	-1043.82 ***	-1043.83 – -1043.82	95663.36 ***	65250.17 – 126076.55	-3.73 ***	-5.31 – -1.90
??AT-??CC	204708.06 **	70777.50 – 338638.63	-3290.23 ***	-3290.24 – -3290.23	165430.70 ***	98074.53 – 232786.87	-3.55	-6.23 – 1.43

??AT-??CG	386898.26 ***	291615.66 – 482180.87	-9305.96 ***	-9305.97 – -9305.96	166263.42 ***	118343.28 – 214183.56	-2.96	-5.41 – 1.97
??AT-??CT	93253.32	-1076.13 – 187582.78	-4018.65 ***	-4018.65 – -4018.64	32368.28	-15033.61 – 79770.17	-2.96	-5.41 – 1.97
??AT-??GC	312312.82 ***	248764.55 – 375861.08	-9961.42 ***	-9961.43 – -9961.42	195455.91 ***	163557.94 – 227353.87	-3.32 *	-5.09 – -0.96
??AT-??GG	252137.57 ***	172146.31 – 332128.83	-2435.16 ***	-2435.17 – -2435.16	140973.55 ***	101027.34 – 180919.76	-5.48 ***	-6.96 – -4.09
??AT-??GT	372941.36 ***	321196.27 – 424686.45	-14924.94 ***	-14924.95 – -14924.94	192870.75 ***	166907.47 – 218834.03	-3.41 **	-4.94 – -1.60
??C?-??C?	54704.31	-12106.93 – 121515.54	-2652.27 ***	-2652.28 – -2652.27	17167.06	-16390.59 – 50724.72	-2.25	-4.53 – 2.65
??C?-??T?	50841.4	-15943.67 – 117626.47	-6265.61 ***	-6265.61 – -6265.61	608.25	-32941.25 – 34157.74	-2.23	-4.51 – 2.67
??CA-??CA	160112.08 *	26857.93 – 293366.23	-5134.30 ***	-5134.30 – -5134.29	87415.11 *	20426.66 – 154403.57	-3.55	-6.23 – 1.43
??CA-??CG	203339.49 *	35773.55 – 370905.42	-9142.71 ***	-9142.72 – -9142.71	84369.33 *	777.26 – 167961.41	-5.16 ***	-7.80 – -2.53
??CA-??TA	753183.79 ***	619057.26 – 887310.31	-18686.20 ***	-18686.21 – -18686.20	182839.28 ***	115415.23 – 250263.32	-5.21 ***	-7.31 – -3.14
??CA-??TG	387244.52 ***	253118.00 – 521371.04	-22844.82 ***	-22844.83 – -22844.82	54300.87	-13123.17 – 121724.92	-3.6	-6.31 – 1.39
??CC-??CC	415498.89 ***	281568.33 – 549429.46	-8184.81 ***	-8184.81 – -8184.80	197031.24 ***	129675.07 – 264387.41	-3.55	-6.23 – 1.43
??CC-??CT	293937.86 ***	160683.71 – 427192.01	-6940.57 ***	-6940.58 – -6940.57	115903.95 ***	48915.50 – 182892.41	-3.55	-6.23 – 1.43
??CC-??TT	250873.22 ***	116746.70 – 384999.74	-10958.34 ***	-10958.35 – -10958.34	149193.41 ***	81769.37 – 216617.45	-3.6	-6.31 – 1.39
??CG-??CA	184005.17 **	50751.02 – 317259.32	-5204.48 ***	-5204.49 – -5204.48	102352.68 **	35364.22 – 169341.13	-3.55	-6.23 – 1.43
??CG-??CG	-9144.29	-143074.85 – 124786.28	-3384.32 ***	-3384.33 – -3384.32	33326.17	-34030.00 – 100682.34	-3.55	-6.23 – 1.43
??CG-??CT	119709.71	-13544.44 – 252963.86	-3683.98 ***	-3683.98 – -3683.97	72328.52 *	5340.06 – 139316.97	-3.55	-6.23 – 1.43
??CG-??TA	278270.69 ***	144126.56 – 412414.83	-15917.12 ***	-15917.13 – -15917.12	49412.72	-18017.30 – 116842.73	-3.6	-6.31 – 1.39
??CG-??TG	176913.74 ***	81355.90 – 272471.58	-12071.67 ***	-12071.68 – -12071.67	109312.77 ***	61297.27 – 157328.26	-3.01	-5.50 – 1.93
??CG-??TT	123903.28	-10223.24 – 258029.80	-8377.09 ***	-8377.10 – -8377.08	77459.38 *	10035.34 – 144883.43	-3.6	-6.31 – 1.39
??CT-??CC	298149.14 ***	164218.57 – 432079.70	-8046.95 ***	-8046.95 – -8046.94	123786.98 ***	56430.81 – 191143.15	-3.55	-6.23 – 1.43
??CT-??CG	413014.68 ***	308341.83 – 517687.52	-9179.31 ***	-9179.31 – -9179.30	178677.53 ***	126216.30 – 231138.77	-4.31 **	-6.34 – -1.85
??CT-??CT	114326.06 *	19996.60 – 208655.51	-1640.40 ***	-1640.40 – -1640.39	74130.15 **	26728.26 – 121532.04	-2.96	-5.41 – 1.97
??CT-??TC	109363.94	-24780.20 – 243508.08	-17861.12 ***	-17861.13 – -17861.12	4391.47	-63038.55 – 71821.48	-3.6	-6.31 – 1.39
??CT-??TG	63144.96	-70981.56 – 197271.48	-5570.86 ***	-5570.87 – -5570.85	45596.98	-21827.07 – 113021.02	-3.6	-6.31 – 1.39
??CT-??TT	205920.30 ***	127317.33 – 284523.28	-16445.02 ***	-16445.03 – -16445.02	69303.78 ***	29824.05 – 108783.52	-2.65	-5.05 – 2.28
??G?-??G?	12020.42	-16186.36 – 40227.20	-2021.59 ***	-2021.59 – -2021.59	-8643.51	-22812.20 – 5525.18	-1.61	-3.20 – 0.67
??G?-??T?	61762.74	-5022.33 – 128547.81	-3525.69 ***	-3525.70 – -3525.69	3482.91	-30066.59 – 37032.41	-2.23	-4.51 – 2.67
??GA-??GA	358919.89 ***	301786.91 – 416052.87	-9380.54 ***	-9380.54 – -9380.53	183630.78 ***	154951.65 – 212309.91	-3.98 ***	-5.44 – -2.39
??GA-??GC	265610.05 ***	213026.32 – 318193.78	-8762.83 ***	-8762.84 – -8762.83	169123.55 ***	142693.49 – 195553.61	-1.8	-4.07 – 3.10
??GA-??GG	377627.49 ***	312275.24 – 442979.74	-13050.93 ***	-13050.93 – -13050.92	217227.95 ***	184512.55 – 249943.35	-4.90 ***	-6.29 – -3.54
??GA-??TA	385700.86 ***	251574.34 – 519827.38	-12091.22 ***	-12091.23 – -12091.22	244751.58 ***	177327.53 – 312175.62	-5.21 ***	-7.31 – -3.14
??GA-??TG	236442.66 ***	102316.14 – 370569.19	-10156.51 ***	-10156.52 – -10156.51	153376.76 ***	85952.71 – 220800.80	-3.6	-6.31 – 1.39
??GC-??GA	230780.03 ***	171447.18 – 290112.89	-8011.66 ***	-8011.67 – -8011.66	140397.86 ***	110685.61 – 170110.11	-4.32 ***	-5.74 – -2.85
??GC-??GC	355399.66 ***	295606.23 – 415193.10	-12497.13 ***	-12497.13 – -12497.13	191269.12 ***	161288.08 – 221250.16	-3.73 ***	-5.31 – -1.90
??GC-??GT	281479.58 ***	215979.32 – 346979.84	-8472.92 ***	-8472.92 – -8472.92	153091.51 ***	120318.69 – 185864.33	-4.91 ***	-6.30 – -3.54
??GC-??TT	167553.93 *	33427.41 – 301680.45	-5197.63 ***	-5197.64 – -5197.62	109856.04 **	42431.99 – 177280.08	-3.6	-6.31 – 1.39

??GG-??GA	237321.10 ***	175808.17 – 298834.04	-4402.38 ***	-4402.39 – -4402.38	151792.75 ***	120878.36 – 182707.14	-2.13	-4.43 – 2.78
??GG-??GG	127122.59 ***	65609.07 – 188636.11	-5414.97 ***	-5414.97 – -5414.97	78495.01 ***	47579.94 – 109410.07	-3.73 ***	-5.31 – -1.90
??GG-??GT	232524.98 ***	166694.04 – 298355.91	-8310.30 ***	-8310.31 – -8310.30	125013.20 ***	92064.44 – 157961.96	-4.53 ***	-5.99 – -3.04
??GG-??TA	253435.54 **	85698.85 – 421172.23	-15632.70 ***	-15632.71 – -15632.69	81840.23	-1811.35 – 165491.82	-5.21 ***	-7.88 – -2.55
??GG-??TG	376019.74 ***	280461.89 – 471577.58	-12072.77 ***	-12072.77 – -12072.76	190926.09 ***	142910.60 – 238941.59	-3.01	-5.50 – 1.93
??GG-??TT	127451.24	-6675.28 – 261577.76	-4135.88 ***	-4135.89 – -4135.88	87996.92 *	20572.88 – 155420.96	-3.6	-6.31 – 1.39
??GT-??GC	306157.09 ***	242608.83 – 369705.35	-9610.55 ***	-9610.55 – -9610.54	179676.34 ***	147778.38 – 211574.30	-3.32 *	-5.09 – -0.96
??GT-??GG	240191.85 ***	176506.30 – 303877.40	-6051.89 ***	-6051.89 – -6051.89	115692.48 ***	83710.81 – 147674.14	-4.17 ***	-5.67 – -2.56
??GT-??GT	341993.87 ***	284788.24 – 399199.50	-16411.99 ***	-16412.00 – -16411.99	181330.99 ***	152674.95 – 209987.02	-4.43 ***	-5.80 – -3.05
??GT-??TC	275094.72 ***	140950.58 – 409238.85	-17615.61 ***	-17615.61 – -17615.60	61180.71	-6249.30 – 128610.73	-3.6	-6.31 – 1.39
??GT-??TG	181657.09 **	47530.57 – 315783.61	-4831.71 ***	-4831.71 – -4831.70	111274.52 **	43850.48 – 178698.56	-3.6	-6.31 – 1.39
??GT-??TT	148414.99 ***	69812.01 – 227017.96	-7522.60 ***	-7522.60 – -7522.59	87398.75 ***	47919.01 – 126878.48	-2.65	-5.05 – 2.28
??T?-??C?	-12993.24	-79804.48 – 53817.99	-2649.05 ***	-2649.05 – -2649.05	2230.03	-31327.62 – 35787.69	-2.25	-4.53 – 2.65
??T?-??G?	9808.6	-18199.60 – 37816.79	-383.26 ***	-383.26 – -383.26	8626.25	-5446.32 – 22698.83	-0.49	-2.65 – 4.39
??T?-??T?	38632.88	-35922.66 – 113188.43	-3504.86 ***	-3504.86 – -3504.85	-11408.08	-48737.61 – 25921.44	-4.11 ***	-5.71 – -2.28
??TA-??CA	46654.97	-86599.18 – 179909.12	-3578.03 ***	-3578.03 – -3578.02	39049.54	-27938.91 – 106038.00	-3.55	-6.23 – 1.43
??TA-??CC	365022.17 ***	231091.61 – 498952.74	-6869.35 ***	-6869.35 – -6869.34	186434.25 ***	119078.07 – 253790.42	-3.55	-6.23 – 1.43
??TA-??CG	310508.94 ***	176578.37 – 444439.50	-8473.96 ***	-8473.97 – -8473.96	133777.17 ***	66420.99 – 201133.34	-3.55	-6.23 – 1.43
??TA-??GA	277016.55 ***	220896.55 – 333136.54	-10410.71 ***	-10410.72 – -10410.71	152593.16 ***	124414.02 – 180772.30	-3.56 ***	-5.11 – -1.74
??TA-??GC	212979.14 ***	160395.41 – 265562.86	-6098.04 ***	-6098.04 – -6098.04	146586.76 ***	120156.70 – 173016.82	-1.8	-4.07 – 3.10
??TA-??GG	258374.63 ***	200181.62 – 316567.64	-11958.07 ***	-11958.07 – -11958.06	155191.97 ***	125976.35 – 184407.60	-3.98 ***	-5.44 – -2.39
??TA-??TA	259909.67 ***	125783.15 – 394036.19	-7678.39 ***	-7678.40 – -7678.38	174779.65 ***	107355.61 – 242203.70	-5.21 ***	-7.31 – -3.14
??TA-??TG	347316.09 ***	213189.56 – 481442.61	-11758.06 ***	-11758.06 – -11758.05	174751.09 ***	107327.04 – 242175.13	-3.6	-6.31 – 1.39
??TC-??CA	166967.43	-58.37 – 333993.22	-6782.27 ***	-6782.28 – -6782.26	12714.71	-70581.35 – 96010.78	-5.16 ***	-7.80 – -2.53
??TC-??CT	184367.28 **	51113.13 – 317621.43	-5598.11 ***	-5598.12 – -5598.10	91434.29 **	24445.84 – 158422.75	-3.55	-6.23 – 1.43
??TC-??GA	277677.72 ***	223910.94 – 331444.50	-6730.77 ***	-6730.77 – -6730.77	161960.68 ***	134960.14 – 188961.22	-2.97 *	-4.69 – -0.63
??TC-??GC	396409.82 ***	332418.62 – 460401.01	-13004.04 ***	-13004.04 – -13004.04	236663.99 ***	204655.97 – 268672.01	-4.53 ***	-5.99 – -3.04
??TC-??GT	268553.21 ***	212049.68 – 325056.75	-8835.36 ***	-8835.37 – -8835.36	164262.09 ***	135867.77 – 192656.41	-3.56 ***	-5.12 – -1.75
??TC-??TT	230625.22 ***	96498.69 – 364751.74	-7643.48 ***	-7643.49 – -7643.47	147157.75 ***	79733.71 – 214581.79	-3.6	-6.31 – 1.39
??TG-??CA	-33025.27	-166279.42 – 100228.88	-1546.95 ***	-1546.96 – -1546.94	13492.05	-53496.40 – 80480.51	-3.55	-6.23 – 1.43
??TG-??CG	274503.26 ***	140572.69 – 408433.83	-7746.48 ***	-7746.49 – -7746.47	156304.39 ***	88948.22 – 223660.56	-3.55	-6.23 – 1.43
??TG-??CT	-104485.22	-237739.37 – 28768.93	-4922.25 ***	-4922.26 – -4922.24	-53731.85	-120720.31 – 13256.61	-3.55	-6.23 – 1.43
??TG-??GA	141367.67 ***	79854.74 – 202880.61	-5407.31 ***	-5407.32 – -5407.31	94834.79 ***	63920.40 – 125749.18	-2.13	-4.43 – 2.78
??TG-??GG	128353.25 ***	61693.11 – 195013.40	-1233.10 ***	-1233.11 – -1233.10	66247.38 ***	32831.67 – 99663.09	-4.53 ***	-5.99 – -3.04
??TG-??GT	263985.31 ***	207480.99 – 320489.63	-10423.47 ***	-10423.47 – -10423.46	66508.83 ***	38114.27 – 94903.39	-1.95	-4.23 – 2.95
??TG-??TA	-31657.77	-165801.91 – 102486.37	-1406.21 ***	-1406.22 – -1406.20	8179.19	-59250.82 – 75609.21	-3.6	-6.31 – 1.39
??TG-??TG	163836.45 ***	68278.61 – 259394.30	-10177.76 ***	-10177.77 – -10177.76	96513.03 ***	48497.54 – 144528.52	-3.01	-5.50 – 1.93

??TG-??TT	217564.31 **	83437.79 – 351690.83	-25231.24 ***	-25231.25 – -25231.23	85093.77 *	17669.73 – 152517.82	-3.6	-6.31 – 1.39
??TT-??CC	391322.15 ***	257391.59 – 525252.72	-8140.07 ***	-8140.08 – -8140.07	142105.45 ***	74749.28 – 209461.62	-3.55	-6.23 – 1.43
??TT-??CT	45032.48	-49296.98 – 139361.93	-576.73 ***	-576.74 – -576.73	42907.81	-4494.08 – 90309.70	-2.96	-5.41 – 1.97
??TT-??GC	323467.57 ***	261954.64 – 384980.51	-11226.03 ***	-11226.03 – -11226.02	171417.80 ***	140503.41 – 202332.19	-2.13	-4.43 – 2.78
??TT-??GG	149081.56 ***	81589.19 – 216573.93	-1233.84 ***	-1233.84 – -1233.84	95451.77 ***	61581.93 – 129321.62	-4.53 ***	-5.99 – -3.04
??TT-??GT	331837.32 ***	274633.13 – 389041.51	-14807.13 ***	-14807.13 – -14807.12	194416.29 ***	165760.70 – 223071.87	-4.43 ***	-5.80 – -3.05
??TT-??TG	228539.53 ***	94413.01 – 362666.06	-5513.23 ***	-5513.24 – -5513.23	151533.95 ***	84109.91 – 218957.99	-3.6	-6.31 – 1.39
??TT-??TT	270159.82 ***	191556.85 – 348762.79	-12468.25 ***	-12468.26 – -12468.25	139004.22 ***	99524.48 – 178483.95	-2.65	-5.05 – 2.28
?A??-?A??	-70040.50 *	-136865.68 – -3215.32	-2889.94 ***	-2889.94 – -2889.93	-39207.12 *	-72768.18 – -5646.06	-2.29	-4.59 – 2.61
?A??-?G??	-13726.71	-47551.24 – 20097.82	-2747.84 ***	-2747.84 – -2747.84	-4395.76	-21368.08 – 12576.56	-1.97	-3.57 – 0.32
?A?A-?A?A	113019.02	-20235.13 – 246273.16	-2761.86 ***	-2761.87 – -2761.85	69481.89 *	2493.43 – 136470.34	-3.55	-6.23 – 1.43
?A?A-?A?C	208283.22 **	74258.79 – 342307.65	-5865.16 ***	-5865.17 – -5865.16	111858.36 **	44469.67 – 179247.05	-3.55	-6.23 – 1.43
?A?A-?A?G	192698.49 **	58674.07 – 326722.92	-4975.56 ***	-4975.57 – -4975.55	100888.08 **	33499.39 – 168276.77	-3.55	-6.23 – 1.43
?A?A-?G?A	341998.07 ***	278400.90 – 405595.23	-11620.62 ***	-11620.63 – -11620.62	192021.44 ***	160105.93 – 223936.96	-3.31 *	-5.07 – -0.96
?A?A-?G?C	260716.82 ***	182254.47 – 339179.16	-11880.13 ***	-11880.13 – -11880.12	168436.89 ***	129003.76 – 207870.01	-2.59	-4.96 – 2.33
?A?A-?G?G	155230.88 **	59689.47 – 250772.29	-1692.85 ***	-1692.86 – -1692.85	111109.56 ***	63098.02 – 159121.10	-4.57 ***	-6.34 – -2.65
?A?C-?A?A	21471.32	-111782.83 – 154725.46	-3470.68 ***	-3470.68 – -3470.67	28735.89	-38252.57 – 95724.34	-3.55	-6.23 – 1.43
?A?C-?A?C	371115.36 ***	208452.60 – 533778.11	-6671.76 ***	-6671.77 – -6671.75	147835.38 ***	67040.16 – 228630.61	-6.01 ***	-8.48 – -4.01
?A?C-?A?T	130126.63	-3127.52 – 263380.78	-4338.47 ***	-4338.48 – -4338.47	48495.49	-18492.97 – 115483.95	-3.55	-6.23 – 1.43
?A?C-?G?A	167355.23 ***	98994.74 – 235715.73	-3919.66 ***	-3919.66 – -3919.65	114979.14 ***	80622.38 – 149335.91	-2.33	-4.65 – 2.59
?A?C-?G?C	178573.22 ***	117020.81 – 240125.64	-1877.91 ***	-1877.92 – -1877.91	133722.65 ***	102793.94 – 164651.35	-2.11	-4.41 – 2.79
?A?C-?G?T	190413.82 ***	133878.86 – 246948.78	-2462.97 ***	-2462.97 – -2462.96	129696.41 ***	101290.85 – 158101.96	-1.94	-4.22 – 2.96
?A?G-?A?A	-87323.62	-181653.07 – 7005.84	215.53 ***	215.53 – 215.54	-17578.35	-64980.24 – 29823.54	-2.96	-5.41 – 1.97
?A?G-?A?G	132872.72	-1151.71 – 266897.15	-6478.78 ***	-6478.79 – -6478.77	44376.53	-23012.16 – 111765.22	-3.55	-6.23 – 1.43
?A?G-?A?T	22645.53	-110608.62 – 155899.68	-582.13 ***	-582.14 – -582.13	11578.78	-55409.68 – 78567.23	-3.55	-6.23 – 1.43
?A?G-?G?A	238782.90 ***	155739.39 – 321826.40	-3371.94 ***	-3371.95 – -3371.94	166338.45 ***	124686.96 – 207989.94	-3.86 **	-5.73 – -1.46
?A?G-?G?G	109069.53 ***	47464.58 – 170674.49	8583.40 ***	8583.40 – 8583.40	113987.86 ***	83040.84 – 144934.88	-2.11	-4.41 – 2.79
?A?G-?G?T	221644.11 ***	140181.41 – 303106.80	-9573.53 ***	-9573.53 – -9573.52	130545.29 ***	89797.09 – 171293.49	-4.71 ***	-6.32 – -3.04
?A?T-?A?C	278099.42 ***	144074.99 – 412123.85	-6338.62 ***	-6338.62 – -6338.61	101856.87 **	34468.18 – 169245.56	-3.55	-6.23 – 1.43
?A?T-?A?G	393770.72 ***	298458.42 – 489083.01	-8717.10 ***	-8717.10 – -8717.09	187188.56 ***	139257.94 – 235119.19	-2.96	-5.41 – 1.97
?A?T-?A?T	-11406.39	-144660.53 – 121847.76	-3525.90 ***	-3525.91 – -3525.89	-3634.21	-70622.67 – 63354.24	-3.55	-6.23 – 1.43
?A?T-?G?C	303638.29 ***	212249.94 – 395026.64	-16515.43 ***	-16515.44 – -16515.43	182376.74 ***	136653.17 – 228100.30	-4.57 ***	-6.34 – -2.65
?A?T-?G?G	190079.29 ***	111616.95 – 268541.64	-2635.07 ***	-2635.07 – -2635.06	134273.38 ***	94840.26 – 173706.51	-2.59	-4.96 – 2.33
?A?T-?G?T	282905.25 ***	183344.07 – 382466.43	-19619.79 ***	-19619.79 – -19619.78	164091.71 ***	114516.73 – 213666.68	-5.92 ***	-7.59 – -4.44
?C??-?A??	-68968.25 *	-135793.43 – -2143.07	-2418.04 ***	-2418.04 – -2418.04	-21486.35	-55047.41 – 12074.72	-2.29	-4.59 – 2.61
?C??-?T??	49958.09 *	10774.25 – 89141.93	-2014.84 ***	-2014.85 – -2014.84	24087.50 *	4426.10 – 43748.89	-2.29 *	-3.91 – -0.00
?C?A-?A?A	198250.21 **	64996.06 – 331504.35	-3534.44 ***	-3534.45 – -3534.44	86008.78 *	19020.33 – 152997.24	-3.55	-6.23 – 1.43

?C?A-?A?C	257315.70 ***	123291.27 – 391340.13	-5190.23 ***	-5190.24 – -5190.23	208184.47 ***	140795.78 – 275573.16	-3.55	-6.23 – 1.43
?C?A-?A?G	295405.65 ***	161381.22 – 429430.07	-6431.16 ***	-6431.17 – -6431.16	111764.84 **	44376.15 – 179153.53	-3.55	-6.23 – 1.43
?C?A-?T?A	318446.46 ***	250050.62 – 386842.30	-7874.21 ***	-7874.21 – -7874.21	176775.09 ***	142407.80 – 211142.39	-2.36	-4.72 – 2.55
?C?A-?T?C	179858.07 ***	109871.24 – 249844.90	-130.72 ***	-130.72 – -130.71	123429.45 ***	88114.09 – 158744.80	-2.36	-4.72 – 2.55
?C?A-?T?G	143544.33 **	50175.79 – 236912.87	-3532.35 ***	-3532.36 – -3532.35	97142.01 ***	50352.68 – 143931.35	-4.61 ***	-6.40 – -2.66
?C?C-?A?A	27569.96	-105684.19 – 160824.11	-1796.19 ***	-1796.20 – -1796.18	37353.79	-29634.67 – 104342.24	-3.55	-6.23 – 1.43
?C?C-?A?C	335684.93 ***	216710.54 – 454659.33	-6225.95 ***	-6225.95 – -6225.94	160262.69 ***	100910.05 – 219615.32	-5.16 ***	-7.22 – -3.12
?C?C-?A?T	68043.2	-65210.95 – 201297.35	-2958.97 ***	-2958.97 – -2958.96	35436.45	-31552.00 – 102424.91	-3.55	-6.23 – 1.43
?C?C-?T?A	382419.82 ***	305645.05 – 459194.58	-9191.96 ***	-9191.97 – -9191.96	183764.05 ***	145287.72 – 222240.39	-4.24 ***	-5.94 – -2.35
?C?C-?T?C	302166.56 ***	206653.02 – 397680.10	-7249.28 ***	-7249.28 – -7249.27	207141.60 ***	159140.53 – 255142.68	-4.61 ***	-6.40 – -2.66
?C?C-?T?T	203099.89 ***	111276.94 – 294922.84	-7500.11 ***	-7500.12 – -7500.11	96944.18 ***	51003.95 – 142884.41	-4.61 ***	-6.40 – -2.66
?C?G-?A?A	28944	-65385.45 – 123273.46	-3370.84 ***	-3370.85 – -3370.84	19614.12	-27787.77 – 67016.01	-2.96	-5.41 – 1.97
?C?G-?A?G	34897.52	-99126.91 – 168921.95	-5970.32 ***	-5970.33 – -5970.31	38194.02	-29194.67 – 105582.71	-3.55	-6.23 – 1.43
?C?G-?A?T	179693.05 **	46438.90 – 312947.20	-6128.87 ***	-6128.88 – -6128.86	82839.69 *	15851.24 – 149828.15	-3.55	-6.23 – 1.43
?C?G-?T?A	192201.07 ***	100380.98 – 284021.15	-7639.44 ***	-7639.44 – -7639.43	118695.71 ***	72756.29 – 164635.14	-4.61 ***	-6.40 – -2.66
?C?G-?T?G	303258.67 ***	224760.46 – 381756.87	-11192.56 ***	-11192.57 – -11192.56	133277.38 ***	93833.31 – 172721.45	-4.24 ***	-5.94 – -2.35
?C?G-?T?T	152130.78 **	60307.83 – 243953.73	-5651.53 ***	-5651.54 – -5651.53	87802.43 ***	41862.20 – 133742.66	-4.61 ***	-6.40 – -2.66
?C?T-?A?C	209323.03 **	75298.60 – 343347.46	-5300.66 ***	-5300.67 – -5300.65	132413.52 ***	65024.83 – 199802.22	-3.55	-6.23 – 1.43
?C?T-?A?G	228592.75 ***	133280.45 – 323905.04	-8522.93 ***	-8522.94 – -8522.93	126130.67 ***	78200.05 – 174061.30	-2.96	-5.41 – 1.97
?C?T-?A?T	24321.93	-108932.22 – 157576.08	-1142.67 ***	-1142.67 – -1142.66	26010.14	-40978.31 – 92998.60	-3.55	-6.23 – 1.43
?C?T-?T?C	331485.13 ***	203563.86 – 459406.41	-10220.34 ***	-10220.34 – -10220.33	123338.57 ***	59424.96 – 187252.17	-5.21 ***	-7.31 – -3.14
?C?T-?T?G	139771.08 ***	61315.91 – 218226.25	-6164.43 ***	-6164.43 – -6164.42	85702.68 ***	46273.24 – 125132.11	-2.63	-5.02 – 2.30
?C?T-?T?T	408754.21 ***	337360.35 – 480148.08	-12268.10 ***	-12268.11 – -12268.10	188641.73 ***	152825.64 – 224457.82	-4.43 ***	-5.99 – -2.77
?G??-?A??	-53094.79	-122861.97 – 16672.40	-2905.45 ***	-2905.46 – -2905.45	-11822.93	-46805.84 – 23159.97	-3.51 **	-5.29 – -1.16
?G??-?G??	34196.54	-115.28 – 68508.37	-5141.89 ***	-5141.89 – -5141.88	18461.38 *	1250.59 – 35672.17	-2.51 **	-3.92 – -0.78
?G??-?T??	17307.37	-21325.71 – 55940.46	-1671.22 ***	-1671.23 – -1671.22	-10428.42	-29822.71 – 8965.87	-1.15	-3.34 – 3.73
?G?A-?A?A	101641.63	-31612.52 – 234895.78	-5256.39 ***	-5256.40 – -5256.39	68969.44 *	1980.99 – 135957.90	-3.55	-6.23 – 1.43
?G?A-?G?A	329048.37 ***	263118.82 – 394977.93	-9975.27 ***	-9975.27 – -9975.26	180720.93 ***	147678.13 – 213763.74	-3.94 ***	-5.54 – -2.09
?G?A-?G?C	211971.87 ***	133509.52 – 290434.21	-11809.67 ***	-11809.67 – -11809.66	141130.76 ***	101697.64 – 180563.89	-2.59	-4.96 – 2.33
?G?A-?G?G	98116.51 *	2575.10 – 193657.92	967.85 ***	967.84 – 967.85	59450.51 *	11438.97 – 107462.05	-4.57 ***	-6.34 – -2.65
?G?A-?T?A	382173.73 ***	292680.50 – 471666.96	-7600.69 ***	-7600.69 – -7600.68	204888.66 ***	160228.54 – 249548.77	-5.20 ***	-6.86 – -3.58
?G?A-?T?C	258256.40 ***	180412.19 – 336100.60	-3221.46 ***	-3221.46 – -3221.45	155923.45 ***	116772.61 – 195074.29	-4.24 ***	-5.94 – -2.35
?G?A-?T?G	239335.73 ***	145967.18 – 332704.27	550.67 ***	550.67 – 550.68	155450.41 ***	108661.08 – 202239.75	-4.61 ***	-6.40 – -2.66
?G?C-?A?A	161083.24 *	27829.09 – 294337.39	-5760.22 ***	-5760.23 – -5760.22	90080.17 **	23091.71 – 157068.62	-3.55	-6.23 – 1.43
?G?C-?A?C	32424.59	-86549.81 – 151398.98	-5955.13 ***	-5955.14 – -5955.12	65213.67 *	5861.04 – 124566.30	-5.16 ***	-7.22 – -3.12
?G?C-?A?T	105603.79	-27650.36 – 238857.94	-5343.54 ***	-5343.55 – -5343.53	49769.95	-17218.51 – 116758.41	-3.55	-6.23 – 1.43
?G?C-?G?A	162668.62 ***	94308.12 – 231029.11	-20261.05 ***	-20261.05 – -20261.04	28438.04	-5918.73 – 62794.81	-2.33	-4.65 – 2.59

?G?C-?G?C	103330.96 **	39770.15 – 166891.77	-17596.49 ***	-17596.50 – -17596.49	61456.25 ***	29557.80 – 93354.70	-3.31 *	-5.07 – -0.96
?G?C-?G?T	257008.76 ***	200473.80 – 313543.73	-12183.01 ***	-12183.02 – -12183.01	190346.46 ***	161940.91 – 218752.02	-1.94	-4.22 – 2.96
?G?C-?T?A	355688.05 ***	278913.29 – 432462.82	-8812.53 ***	-8812.54 – -8812.53	188093.52 ***	149617.19 – 226569.86	-4.24 ***	-5.94 – -2.35
?G?C-?T?C	339234.24 ***	234346.24 – 444122.24	-5566.94 ***	-5566.95 – -5566.93	199524.96 ***	146988.09 – 252061.83	-5.19 ***	-7.01 – -3.41
?G?C-?T?T	298618.51 ***	215354.95 – 381882.06	-7941.67 ***	-7941.67 – -7941.66	148468.81 ***	106719.85 – 190217.77	-3.89 **	-5.79 – -1.48
?G?G-?A?A	-91795.4	-186124.86 – 2534.05	-484.70 ***	-484.71 – -484.70	-2242.86	-49644.75 – 45159.03	-2.96	-5.41 – 1.97
?G?G-?A?G	54122.79	-79901.64 – 188147.22	-6029.16 ***	-6029.17 – -6029.16	58250.35	-9138.35 – 125639.04	-3.55	-6.23 – 1.43
?G?G-?A?T	156254.69 *	23000.54 – 289508.84	-4692.72 ***	-4692.73 – -4692.72	86948.59 *	19960.13 – 153937.04	-3.55	-6.23 – 1.43
?G?G-?G?A	320793.47 ***	237755.64 – 403831.30	-14313.97 ***	-14313.97 – -14313.96	183249.01 ***	141599.41 – 224898.61	-3.86 **	-5.73 – -1.46
?G?G-?G?G	115330.95 ***	53726.00 – 176935.91	-1742.15 ***	-1742.15 – -1742.15	96217.31 ***	65270.29 – 127164.33	-2.11	-4.41 – 2.79
?G?G-?G?T	314006.71 ***	232544.02 – 395469.41	-19377.92 ***	-19377.93 – -19377.92	137429.28 ***	96681.08 – 178177.48	-4.71 ***	-6.32 – -3.04
?G?G-?T?A	254605.25 ***	171342.86 – 337867.63	-8011.07 ***	-8011.07 – -8011.06	140233.36 ***	98484.73 – 181981.99	-3.89 **	-5.79 – -1.48
?G?G-?T?G	296381.78 ***	217883.58 – 374879.99	-7987.37 ***	-7987.37 – -7987.37	177508.95 ***	138064.88 – 216953.02	-4.24 ***	-5.94 – -2.35
?G?G-?T?T	235350.68 ***	156837.97 – 313863.39	-7677.34 ***	-7677.35 – -7677.34	107783.31 ***	68334.25 – 147232.38	-2.63	-5.02 – 2.30
?G?T-?A?T	183919.22 *	16893.43 – 350945.01	-7987.91 ***	-7987.92 – -7987.90	40370.71	-42925.35 – 123666.78	-5.16 ***	-7.80 – -2.53
?G?T-?G?C	244237.34 ***	165775.00 – 322699.69	-10063.52 ***	-10063.52 – -10063.51	149908.44 ***	110475.32 – 189341.56	-2.59	-4.96 – 2.33
?G?T-?G?G	157440.94 ***	78978.60 – 235903.29	-4233.26 ***	-4233.26 – -4233.25	129379.54 ***	89946.41 – 168812.66	-2.59	-4.96 – 2.33
?G?T-?G?T	312565.37 ***	237688.23 – 387442.52	-10688.67 ***	-10688.68 – -10688.67	186698.52 ***	149193.70 – 224203.35	-4.79 ***	-6.30 – -3.27
?G?T-?T?C	212266.15 ***	116765.59 – 307766.71	-12851.64 ***	-12851.64 – -12851.63	57574.81 *	9579.04 – 105570.58	-3.01	-5.50 – 1.93
?G?T-?T?G	252294.31 ***	173839.14 – 330749.48	-5877.55 ***	-5877.56 – -5877.55	124311.63 ***	84882.20 – 163741.07	-2.63	-5.02 – 2.30
?G?T-?T?T	514164.76 ***	433152.36 – 595177.16	-12015.02 ***	-12015.02 – -12015.01	215697.71 ***	175197.56 – 256197.86	-5.19 ***	-6.74 – -3.69
?T??-?G??	-13088.22	-46556.75 – 20380.31	-1547.81 ***	-1547.82 – -1547.81	-10480.12	-27279.92 – 6319.68	-0.84	-3.01 – 4.04
?T??-?T??	6520.52	-33431.45 – 46472.50	-911.19 ***	-911.19 – -911.19	2682.24	-17355.31 – 22719.79	-2.85 **	-4.28 – -1.11
?T?A-?G?A	307704.38 ***	246140.02 – 369268.74	-10871.98 ***	-10871.98 – -10871.97	176198.75 ***	145266.09 – 207131.42	-2.11	-4.41 – 2.79
?T?A-?G?C	194599.61 ***	116137.27 – 273061.95	-11485.72 ***	-11485.73 – -11485.72	118668.46 ***	79235.34 – 158101.59	-2.59	-4.96 – 2.33
?T?A-?G?G	159214.27 **	63672.86 – 254755.68	-9244.55 ***	-9244.56 – -9244.55	91234.43 ***	43222.89 – 139245.97	-4.57 ***	-6.34 – -2.65
?T?A-?T?A	209851.84 ***	138487.99 – 281215.69	-6705.52 ***	-6705.52 – -6705.51	120234.61 ***	84429.22 – 156040.01	-3.59 **	-5.42 – -1.19
?T?A-?T?C	229917.26 ***	159930.43 – 299904.08	-5818.50 ***	-5818.51 – -5818.50	135972.83 ***	100657.48 – 171288.19	-2.36	-4.72 – 2.55
?T?A-?T?G	222003.87 ***	138515.27 – 305492.46	-5033.99 ***	-5034.00 – -5033.99	136221.81 ***	94339.12 – 178104.50	-3.89 **	-5.79 – -1.48
?T?C-?G?A	194960.53 ***	126600.03 – 263321.02	-7491.81 ***	-7491.82 – -7491.81	85922.00 ***	51565.24 – 120278.77	-2.33	-4.65 – 2.59
?T?C-?G?C	168053.93 ***	106501.51 – 229606.34	-5404.94 ***	-5404.94 – -5404.94	94178.88 ***	63250.18 – 125107.59	-2.11	-4.41 – 2.79
?T?C-?G?T	146034.88 ***	89499.92 – 202569.84	-3975.08 ***	-3975.08 – -3975.08	95375.61 ***	66970.06 – 123781.17	-1.94	-4.22 – 2.96
?T?C-?T?A	223839.31 ***	152508.12 – 295170.49	-5029.95 ***	-5029.95 – -5029.95	113596.43 ***	77814.01 – 149378.85	-3.59 **	-5.42 – -1.19
?T?C-?T?C	230592.61 ***	104553.48 – 356631.73	-6038.61 ***	-6038.62 – -6038.61	174001.74 ***	111098.82 – 236904.65	-5.78 ***	-7.75 – -4.03
?T?C-?T?T	181440.47 ***	102927.76 – 259953.18	-3077.17 ***	-3077.17 – -3077.16	107326.60 ***	67877.54 – 146775.67	-2.63	-5.02 – 2.30
?T?G-?G?A	251784.65 ***	173368.61 – 330200.69	-13128.82 ***	-13128.83 – -13128.82	156385.11 ***	116967.94 – 195802.28	-2.59	-4.96 – 2.33
?T?G-?G?G	102120.02 **	40515.07 – 163724.98	-3444.82 ***	-3444.82 – -3444.82	90936.48 ***	59989.46 – 121883.50	-2.11	-4.41 – 2.79

?T?G-?G?T	267197.27 ***	198808.69 – 335585.84	-10348.91 ***	-10348.92 – -10348.91	87456.75 ***	53090.26 – 121823.24	-2.33	-4.65 – 2.59
?T?G-?T?A	181389.45 ***	102876.74 – 259902.16	-8110.65 ***	-8110.66 – -8110.65	115845.47 ***	76396.41 – 155294.54	-2.63	-5.02 – 2.30
?T?G-?T?G	244606.64 ***	166108.44 – 323104.85	-9107.79 ***	-9107.79 – -9107.78	175291.86 ***	135847.79 – 214735.93	-4.24 ***	-5.94 – -2.35
?T?G-?T?T	129431.02 **	46167.46 – 212694.58	-3407.71 ***	-3407.71 – -3407.70	72320.58 ***	30571.62 – 114069.53	-3.89 **	-5.79 – -1.48
?T?T-?G?C	302098.51 ***	223636.17 – 380560.85	-10819.91 ***	-10819.92 – -10819.91	188365.33 ***	148932.21 – 227798.46	-2.59	-4.96 – 2.33
?T?T-?G?G	144231.22 ***	65768.88 – 222693.57	-3773.03 ***	-3773.03 – -3773.03	118650.14 ***	79217.02 – 158083.27	-2.59	-4.96 – 2.33
?T?T-?G?T	256780.49 ***	188396.95 – 325164.02	-9838.02 ***	-9838.02 – -9838.02	164857.76 ***	130493.13 – 199222.39	-3.94 ***	-5.54 – -2.09
?T?T-?T?G	123083.67 **	31698.39 – 214468.94	-2448.90 ***	-2448.90 – -2448.89	84698.47 ***	38977.10 – 130419.83	-4.61 ***	-6.40 – -2.66
?T?T-?T?T	288626.18 ***	220208.14 – 357044.23	-11315.60 ***	-11315.61 – -11315.60	152982.11 ***	118606.68 – 187357.54	-3.97 ***	-5.60 – -2.10
A???-A???	58902.95 **	20996.56 – 96809.35	-4866.83 ***	-4866.83 – -4866.82	15518.95	-3482.54 – 34520.44	-3.33 ***	-4.59 – -1.99
A???-G???	-2250.97	-35200.48 – 30698.53	-73.97 ***	-73.97 – -73.97	-2513.79	-19034.33 – 14006.76	-3.07 ***	-4.31 – -1.74
A??A-A??A	404388.52 ***	338103.05 – 470673.99	-7323.15 ***	-7323.16 – -7323.15	144588.31 ***	111356.31 – 177820.32	-4.53 ***	-5.99 – -3.04
A??A-A??C	319965.46 ***	261902.63 – 378028.28	-6937.09 ***	-6937.09 – -6937.09	208768.95 ***	179613.15 – 237924.74	-3.13 *	-4.87 – -0.78
A??A-A??G	204340.41 ***	101426.78 – 307254.05	-8045.03 ***	-8045.03 – -8045.02	94774.37 ***	43350.50 – 146198.24	-5.18 ***	-6.98 – -3.41
A??A-G??A	306098.47 ***	232588.75 – 379608.19	-3208.19 ***	-3208.19 – -3208.18	181241.36 ***	144472.23 – 218010.49	-4.81 ***	-6.34 – -3.27
A??A-G??C	173879.62 ***	107244.09 – 240515.14	522.72 ***	522.71 – 522.72	111525.84 ***	78120.35 – 144931.32	-3.96 ***	-5.58 – -2.10
A??A-G??G	144368.47 ***	73170.84 – 215566.10	-3576.88 ***	-3576.89 – -3576.88	102902.58 ***	67164.05 – 138641.11	-3.56 **	-5.37 – -1.18
A??C-A??A	342186.82 ***	273293.88 – 411079.77	-8277.49 ***	-8277.50 – -8277.49	148691.02 ***	113948.65 – 183433.39	-4.53 ***	-5.99 – -3.04
A??C-A??C	323246.42 ***	233926.95 – 412565.89	-4743.43 ***	-4743.44 – -4743.43	217901.73 ***	173360.33 – 262443.13	-5.54 ***	-7.11 – -4.07
A??C-A??T	65185.1	-11424.32 – 141794.52	-3106.76 ***	-3106.77 – -3106.76	40789.12 *	2379.94 – 79198.30	-4.22 ***	-5.88 – -2.34
A??C-G??A	223040.27 ***	165559.34 – 280521.19	-4047.70 ***	-4047.71 – -4047.70	141343.40 ***	112502.89 – 170183.92	-3.14 *	-4.89 – -0.78
A??C-G??C	192518.97 ***	103558.75 – 281479.20	-3983.66 ***	-3983.66 – -3983.65	112471.52 ***	68104.97 – 156838.07	-5.54 ***	-7.11 – -4.07
A??C-G??T	108472.00 **	40773.78 – 176170.22	-2193.90 ***	-2193.90 – -2193.89	69668.27 ***	35656.43 – 103680.11	-2.34	-4.67 – 2.57
A??G-A??A	220113.34 ***	124655.49 – 315571.18	-6379.48 ***	-6379.49 – -6379.48	131568.41 ***	83586.31 – 179550.51	-2.99	-5.45 – 1.95
A??G-A??G	159109.32 ***	66571.57 – 251647.07	2875.73 ***	2875.72 – 2875.73	120400.44 ***	74060.14 – 166740.75	-4.59 ***	-6.37 – -2.66
A??G-A??T	295906.45 ***	214604.37 – 377208.54	-11548.64 ***	-11548.65 – -11548.64	116338.88 ***	75670.30 – 157007.46	-4.72 ***	-6.34 – -3.04
A??G-G??A	210398.22 ***	153569.61 – 267226.82	-9642.07 ***	-9642.07 – -9642.06	144519.03 ***	116057.46 – 172980.60	-3.98 ***	-5.46 – -2.39
A??G-G??G	71827.43 *	5196.83 – 138458.03	479.23 ***	479.23 – 479.24	43732.47 *	10328.11 – 77136.82	-3.95 ***	-5.56 – -2.10
A??G-G??T	124906.05 ***	58553.81 – 191258.29	-8620.03 ***	-8620.03 – -8620.02	39463.75 *	6223.66 – 72703.84	-3.96 ***	-5.58 – -2.10
A??T-A??C	279766.45 ***	211451.95 – 348080.95	-5707.80 ***	-5707.81 – -5707.80	193873.94 ***	159533.00 – 228214.88	-2.33	-4.65 – 2.59
A??T-A??G	307860.00 ***	229433.90 – 386286.10	-8922.05 ***	-8922.05 – -8922.04	113852.79 ***	74432.68 – 153272.90	-2.61	-4.99 – 2.31
A??T-A??T	560443.72 ***	467042.01 – 653845.42	-19571.37 ***	-19571.38 – -19571.37	309681.16 ***	262879.50 – 356482.82	-5.17 ***	-6.81 – -3.58
A??T-G??C	269736.40 ***	208362.61 – 331110.19	-10635.06 ***	-10635.06 – -10635.05	124820.09 ***	93954.13 – 155686.06	-2.13	-4.43 – 2.78
A??T-G??G	237786.56 ***	158565.05 – 317008.06	-7725.20 ***	-7725.20 – -7725.19	132887.32 ***	93351.02 – 172423.63	-5.17 ***	-6.70 – -3.69
A??T-G??T	224819.18 ***	162892.41 – 286745.94	-8310.78 ***	-8310.78 – -8310.78	119788.02 ***	88781.56 – 150794.48	-4.18 ***	-5.68 – -2.56
C???-A???	76583.50 ***	40847.30 – 112319.69	-4442.35 ***	-4442.35 – -4442.35	27110.54 **	9165.54 – 45055.54	-0.98	-3.16 – 3.90
C??A-A??A	319993.22 ***	198200.30 – 441786.14	-10773.08 ***	-10773.08 – -10773.07	169458.33 ***	108936.00 – 229980.65	-6.60 ***	-8.46 – -5.08

C??A-A??C	542229.95 ***	477249.66 – 607210.23	-11723.55 ***	-11723.55 – -11723.55	291423.84 ***	258903.52 – 323944.16	-4.53 ***	-5.99 – -3.04
C??A-A??G	286352.18 ***	160626.77 – 412077.58	-11968.98 ***	-11968.99 – -11968.98	163381.21 ***	100700.95 – 226061.47	-5.76 ***	-7.72 – -4.03
C??C-A??A	288613.70 ***	225936.47 – 351290.92	-6697.07 ***	-6697.07 – -6697.06	147824.94 ***	116165.34 – 179484.55	-3.73 ***	-5.31 – -1.90
C??C-A??C	361776.36 ***	294657.38 – 428895.34	-8223.75 ***	-8223.76 – -8223.75	209992.21 ***	176342.11 – 243642.30	-3.95 ***	-5.56 – -2.10
C??C-A??T	280380.14 ***	188131.75 – 372628.53	-11299.90 ***	-11299.91 – -11299.89	157273.42 ***	111111.51 – 203435.32	-5.17 ***	-6.81 – -3.58
C??G-A??A	225825.67 ***	130367.83 – 321283.52	-5908.13 ***	-5908.14 – -5908.13	137977.42 ***	89995.32 – 185959.52	-2.99	-5.45 – 1.95
C??G-A??G	103684.22 *	11146.47 – 196221.97	-5347.76 ***	-5347.77 – -5347.76	82285.38 ***	35945.07 – 128625.69	-4.59 ***	-6.37 – -2.66
C??G-A??T	453583.90 ***	376629.76 – 530538.05	-18693.96 ***	-18693.96 – -18693.95	79580.34 ***	41000.02 – 118160.67	-4.22 ***	-5.88 – -2.34
C??T-A??C	294899.78 ***	226585.27 – 363214.28	-5713.24 ***	-5713.24 – -5713.24	190595.44 ***	156254.50 – 224936.38	-2.33	-4.65 – 2.59
C??T-A??G	187930.07 ***	96150.73 – 279709.40	-7688.34 ***	-7688.35 – -7688.34	79559.41 ***	33625.11 – 125493.72	-4.59 ***	-6.37 – -2.66
C??T-A??T	325952.36 ***	242452.39 – 409452.34	-12959.73 ***	-12959.74 – -12959.73	155302.16 ***	113414.44 – 197189.88	-4.72 ***	-6.34 – -3.04
G???-A???	57348.08 **	21176.80 – 93519.37	-4649.75 ***	-4649.76 – -4649.75	23334.94 *	5178.91 – 41490.98	-2.12	-3.72 – 0.17
G???-G???	-6893.69	-38419.97 – 24632.60	-1515.14 ***	-1515.14 – -1515.14	-10228.86	-26057.86 – 5600.14	-0.75	-2.92 – 4.12
G??A-A??A	240022.95 ***	149222.44 – 330823.47	-6633.40 ***	-6633.41 – -6633.40	163152.95 ***	117814.75 – 208491.14	-5.80 ***	-7.34 – -4.41
G??A-A??C	326056.78 ***	267993.95 – 384119.61	-8248.87 ***	-8248.87 – -8248.86	219067.64 ***	189911.84 – 248223.43	-3.13 *	-4.87 – -0.78
G??A-A??G	238968.60 ***	136041.59 – 341895.61	-6291.14 ***	-6291.15 – -6291.14	109235.34 ***	57807.06 – 160663.62	-5.18 ***	-6.98 – -3.41
G??A-G??A	220560.30 ***	150572.96 – 290547.65	-3357.40 ***	-3357.40 – -3357.39	173926.35 ***	138859.50 – 208993.21	-4.42 ***	-5.97 – -2.77
G??A-G??C	167613.27 ***	106158.38 – 229068.16	-2638.73 ***	-2638.73 – -2638.73	109665.26 ***	78771.50 – 140559.02	-2.14	-4.45 – 2.77
G??A-G??G	181719.93 ***	110540.94 – 252898.91	-4921.00 ***	-4921.00 – -4920.99	125739.78 ***	90007.36 – 161472.20	-3.56 **	-5.37 – -1.18
G??C-A??A	247235.80 ***	171556.73 – 322914.87	-9773.56 ***	-9773.57 – -9773.56	121324.46 ***	83271.68 – 159377.25	-5.17 ***	-6.62 – -3.77
G??C-A??C	375053.59 ***	311364.03 – 438743.15	-12910.94 ***	-12910.94 – -12910.94	204366.28 ***	172402.25 – 236330.31	-3.32 *	-5.09 – -0.96
G??C-A??T	343150.60 ***	250902.21 – 435398.99	-19124.06 ***	-19124.06 – -19124.05	143124.51 ***	96962.60 – 189286.41	-5.17 ***	-6.81 – -3.58
G??C-G??A	187265.94 ***	129766.65 – 244765.23	-4902.60 ***	-4902.61 – -4902.60	135992.07 ***	107142.26 – 164841.88	-3.14 *	-4.89 – -0.78
G??C-G??C	293417.40 ***	227409.73 – 359425.07	-6189.12 ***	-6189.12 – -6189.11	122807.54 ***	89692.28 – 155922.80	-3.95 ***	-5.56 – -2.10
G??C-G??T	88821.26 *	21123.04 – 156519.48	-3238.58 ***	-3238.58 – -3238.57	48461.13 **	14449.29 – 82472.98	-2.34	-4.67 – 2.57
G??G-A??A	227566.19 ***	132108.34 – 323024.03	-11556.70 ***	-11556.70 – -11556.69	128986.28 ***	81004.18 – 176968.38	-2.99	-5.45 – 1.95
G??G-A??G	210914.75 ***	132488.65 – 289340.85	-10318.78 ***	-10318.79 – -10318.78	61754.78 **	22334.67 – 101174.89	-2.61	-4.99 – 2.31
G??G-A??T	275848.53 ***	198894.39 – 352802.68	-17180.58 ***	-17180.58 – -17180.57	-65219.82 ***	-103800.15 – -26639.49	-4.22 ***	-5.88 – -2.34
G??G-G??A	266655.27 ***	212004.40 – 321306.14	-10719.64 ***	-10719.64 – -10719.63	161488.53 ***	134090.02 – 188887.04	-3.56 ***	-5.12 – -1.75
G??G-G??G	85423.40 **	23973.52 – 146873.29	570.09 ***	570.08 – 570.09	44214.59 **	13322.03 – 75107.15	-2.13	-4.43 – 2.78
G??G-G??T	115708.93 ***	54755.94 – 176661.92	-6145.81 ***	-6145.82 – -6145.81	51406.39 ***	20791.55 – 82021.23	-2.14	-4.45 – 2.77
G??T-A??C	282824.65 ***	214510.15 – 351139.15	-7215.11 ***	-7215.12 – -7215.11	197595.33 ***	163254.39 – 231936.27	-2.33	-4.65 – 2.59
G??T-A??G	211643.69 ***	122467.91 – 300819.47	-6998.02 ***	-6998.03 – -6998.02	112445.51 ***	67831.84 – 157059.17	-4.59 ***	-6.37 – -2.66
G??T-A??T	554939.25 ***	452203.18 – 657675.33	-21439.42 ***	-21439.43 – -21439.42	248878.08 ***	197555.50 – 300200.65	-5.62 ***	-7.34 – -4.05
G??T-G??C	147869.47 ***	86495.68 – 209243.26	-7045.36 ***	-7045.37 – -7045.36	110019.30 ***	79153.34 – 140885.27	-2.13	-4.43 – 2.78
G??T-G??G	199972.87 ***	138595.32 – 261350.43	-5105.79 ***	-5105.80 – -5105.79	123291.04 ***	92423.93 – 154158.15	-2.13	-4.43 – 2.78
G??T-G??T	203810.46 ***	147818.28 – 259802.64	-7803.53 ***	-7803.53 – -7803.53	101356.10 ***	73236.65 – 129475.54	-1.96	-4.25 – 2.95

T???-G???	-13703.31	-45906.98 – 18500.37	-1143.13 ***	-1143.13 – -1143.13	-14063.18	-30221.22 – 2094.86	-2.42 **	-3.83 – -0.70
T??A-G??A	214882.87 ***	148720.69 – 281045.05	-3397.78 ***	-3397.79 – -3397.78	144643.96 ***	111412.29 – 177875.62	-3.96 ***	-5.58 – -2.10
T??A-G??C	138183.07 ***	76728.18 – 199637.96	-1502.25 ***	-1502.26 – -1502.25	85413.38 ***	54519.62 – 116307.14	-2.14	-4.45 – 2.77
T??A-G??G	133099.89 ***	58388.49 – 207811.30	-6027.72 ***	-6027.73 – -6027.72	89579.75 ***	52144.52 – 127014.99	-4.22 ***	-5.88 – -2.34
T??C-G??A	198931.84 ***	141432.54 – 256431.13	-6256.62 ***	-6256.62 – -6256.61	93704.91 ***	64855.10 – 122554.72	-3.14 *	-4.89 – -0.78
T??C-G??C	181524.50 ***	111103.36 – 251945.65	-6090.18 ***	-6090.18 – -6090.17	106002.02 ***	70696.92 – 141307.11	-4.41 ***	-5.95 – -2.77
T??C-G??T	65691.3	-2006.92 – 133389.52	-4671.14 ***	-4671.15 – -4671.14	28542.32	-5469.53 – 62554.16	-2.34	-4.67 – 2.57
T??G-G??A	204325.86 ***	149674.85 – 258976.88	-9805.23 ***	-9805.23 – -9805.23	130879.42 ***	103480.87 – 158277.96	-3.56 ***	-5.12 – -1.75
T??G-G??G	59309.17	-2140.72 – 120759.06	-6615.75 ***	-6615.76 – -6615.75	32476.60 *	1584.04 – 63369.16	-2.13	-4.43 – 2.78
T??G-G??T	86385.40 **	25432.41 – 147338.39	-5600.34 ***	-5600.34 – -5600.34	37988.57 *	7373.73 – 68603.41	-2.14	-4.45 – 2.77
T??T-G??C	113543.47 ***	52169.68 – 174917.26	-4401.07 ***	-4401.07 – -4401.07	89139.61 ***	58273.64 – 120005.57	-2.13	-4.43 – 2.78
T??T-G??G	267858.55 ***	206481.00 – 329236.10	-6367.14 ***	-6367.14 – -6367.14	144730.62 ***	113863.51 – 175597.73	-2.13	-4.43 – 2.78
T??T-G??T	135488.82 ***	76245.01 – 194732.64	-4629.86 ***	-4629.86 – -4629.85	66727.31 ***	37034.01 – 96420.61	-3.74 ***	-5.33 – -1.91
log(Copy_Number)	-29806.08 ***	-33111.53 – -26500.62	815.61 ***	815.61 – 815.61	-18679.63 ***	-20333.60 – -17025.65	0.03	-0.07 – 0.14
FAMDuty	22803.42 ***	15870.89 – 29735.95	-504.44 ***	-504.44 – -504.44	8060.37 ***	3829.30 – 12291.45		
??CA-??CC							-6.77 ***	-11.75 – -4.11
??CC-??CA							-6.77 ***	-11.75 – -4.11
??TC-??CC							-6.77 ***	-11.75 – -4.11
??TT-??CG							-7.36 ***	-12.29 – -4.94
??TT-??TC							-6.82 ***	-11.81 – -4.13
?G?A-?A?C							-6.77 ***	-11.75 – -4.11
?G?A-?A?G							-6.77 ***	-11.75 – -4.11
?G?T-?A?C							-6.77 ***	-11.75 – -4.11
?G?T-?A?G							-7.36 ***	-12.29 – -4.94
?T?T-?T?C							-7.41 ***	-12.36 – -4.95
ICC	0.69		0.84		0.81			
N	501 $_{ExpID}$		501 $_{ExpID}$		501 $_{ExpID}$			
	7 $_{Target}$		7 $_{Target}$		7 $_{Target}$			
Observations	3543		3543		3543		4008	
Marginal R^2 / Conditional R^2	0.359 / 0.800		0.115 / 0.863		0.287 / 0.867			
Adjusted R^2	0.781		0.850		0.854			
** p<0.05 ** p<0.01 *** p<0.001*								

Supplementary Table 2. Coefficients obtained when generalising model according to mismatch position.

	TMG		MG		TP		RS	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Log-Odds*	*CI*
(Intercept)	220740.32 **	85877.87 – 355602.77	17489.42 ***	17489.42 – 17489.43	357041.97 ***	265180.41 – 448903.54	5.03 ***	4.02 – 6.14
1n	29225.06 **	7301.31 – 51148.81	-1690.54 ***	-1690.54 – -1690.54	6278.24	-5365.47 – 17921.96	-1.71 **	-2.83 – -0.58
2n	586.56	-21048.88 – 22222.00	-2283.58 ***	-2283.58 – -2283.58	-2523.11	-13988.81 – 8942.58	-1.82 **	-2.92 – -0.75
3n	24546.77 *	2894.46 – 46199.08	-2680.51 ***	-2680.51 – -2680.51	3826.58	-7656.66 – 15309.83	-2.27 ***	-3.31 – -1.32
T	114189.59 ***	97550.69 – 130828.48	-3386.97 ***	-3386.97 – -3386.97	63249.46 ***	54498.80 – 72000.11	-3.07 ***	-4.00 – -2.29
T1n	221225.63 ***	201461.35 – 240989.91	-7021.01 ***	-7021.01 – -7021.01	122009.89 ***	111731.49 – 132288.29	-3.57 ***	-4.55 – -2.74
T2n	177187.10 ***	157376.37 – 196997.84	-5738.28 ***	-5738.28 – -5738.27	103152.77 ***	92869.84 – 113435.69	-3.48 ***	-4.47 – -2.65
T3n	183061.16 ***	163400.19 – 202722.12	-5614.04 ***	-5614.04 – -5614.03	97658.36 ***	87433.89 – 107882.83	-3.75 ***	-4.73 – -2.93
log(Copy_Number)	-21668.12 ***	-25163.63 – -18172.62	526.65 ***	526.65 – 526.65	-13531.39 ***	-15336.46 – -11726.31	0	-0.09 – 0.10
FAMDuty	25362.69 ***	18187.66 – 32537.71	-486.37 ***	-486.37 – -486.36	8281.33 ***	3648.21 – 12914.45		
ICC	0.58		0.82		0.77			
N	501 ExpID		501 ExpID		501 ExpID			
	7 Target		7 Target		7 Target			
Observations	3543		3543		3543		4008	
Marginal R^2 / Conditional R^2	0.171 / 0.656		0.032 / 0.826		0.102 / 0.792			
Adjusted R^2	0.655		0.826		0.791			
** p<0.05 ** p<0.01 *** p<0.001*								

Supplementary Table 3. Coefficients obtained when generalising model according to gibbs free energy.

	TMG		MG		TP		RS	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Log-Odds*	*CI*
(Intercept)	205819.55 **	73647.21 – 337991.90	18278.09 ***	18278.09 – 18278.10	314381.00 ***	224607.12 – 404154.89	3.83 ***	3.07 – 4.61
Terminal_Gibbs_Free	13954.18 ***	11040.98 – 16867.38	-818.75 ***	-818.75 – -818.75	6278.93 ***	4645.58 – 7912.27	-0.29 ***	-0.38 – -0.21
log(Copy_Number)	-2070.35	-4675.60 – 534.89	30.92 ***	30.92 – 30.92	-3195.22 ***	-4662.50 – -1727.94	-0.19 ***	-0.26 – -0.12
FAMDuty	21406.66 ***	14366.91 – 28446.40	-517.20 ***	-517.20 – -517.20	7561.13 **	2880.08 – 12242.19		
ICC	0.51		0.85		0.65			
N	501 ExpID		501 ExpID		501 ExpID			
	7 Target		7 Target		7 Target			
Observations	3384		3384		3384		3384	
Marginal R^2 / Conditional R^2	0.088 / 0.550		0.033 / 0.857		0.029 / 0.658			
Adjusted R^2	0.550		0.857		0.658			
** p<0.05 ** p<0.01 *** p<0.001*								

Supplementary Table 4. Coefficients obtained when generalising model according to mismatch position, excluding T1n.

	TMG		MG		TP		RS	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Log-Odds*	*CI*
(Intercept)	297606.75 ***	164783.08 – 430430.42	14717.33 ***	14717.33 – 14717.34	395188.92 ***	304405.17 – 485972.66	4.43 ***	3.39 – 5.57
1n	26440.55 *	3650.16 – 49230.94	-1751.02 ***	-1751.02 – -1751.02	5006.42	-8044.00 – 18056.84	-1.72 **	-2.84 – -0.59
2n	-362.42	-22904.22 – 22179.38	-2519.40 ***	-2519.40 – -2519.40	-2937.95	-15829.18 – 9953.27	-1.84 **	-2.94 – -0.76
3n	17401.49	-5053.89 – 39856.88	-2796.31 ***	-2796.31 – -2796.31	-2615.83	-15470.65 – 10238.98	-2.28 ***	-3.32 – -1.33
T	113339.65 ***	96805.20 – 129874.09	-4771.54 ***	-4771.54 – -4771.54	55713.73 ***	46317.79 – 65109.67	-3.23 ***	-4.17 – -2.46
T2n	168720.71 ***	148424.01 – 189017.41	-7543.29 ***	-7543.29 – -7543.28	89269.36 ***	77888.56 – 100650.17	-3.84 ***	-4.84 – -2.98
T3n	179435.52 ***	159452.18 – 199418.85	-7492.47 ***	-7492.47 – -7492.47	79508.90 ***	68278.18 – 90739.62	-4.10 ***	-5.10 – -3.25
log(Copy_Number)	-24176.70 ***	-27784.98 – -20568.42	827.83 ***	827.83 – 827.83	-14050.24 ***	-16065.31 – -12035.17	0.08	-0.02 – 0.18
FAMDuty	22205.56 ***	15196.57 – 29214.55	-464.33 ***	-464.33 – -464.33	6297.21 **	1676.21 – 10918.21		
ICC	0.54		0.87		0.69			
N	501 ExpID		501 ExpID		501 ExpID			
	7 Target		7 Target		7 Target			
Observations	3384		3384		3384		3384	
Marginal R^2 / Conditional R^2	0.133 / 0.600		0.038 / 0.874		0.063 / 0.711			
Adjusted R^2	0.599		0.874		0.710			
** p<0.05 ** p<0.01 *** p<0.001*								

Chapter 5. PrimedInclusivity. A programmable framework to assess the presence and impact of primer binding site nucleotide variation on NAAT-based assay performance.

RESEARCH PAPER COVER SHEET

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

SECTION A – Student Details

Student ID Number	1702842	**Title**	Mr
First Name(s)	Matthew		
Surname/Family Name	Higgins		
Thesis Title	Developing an RPA-based Molecular Barcoding Tool for Plasmodium Malaria		
Primary Supervisor	Prof. Taane Clark		

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?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	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?	Bioinformatics
Please list the paper's authors in the intended authorship order:	Matthew Higgins, Susana Campino, Taane G. Clark
Stage of publication	**Submitted**

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 designed and built the PrimedInclusivity software. I also performed all wet-lab experiments associated with software validation and building of custom RPA and Taq PCR modules. I wrote the first draft of the manuscript which was then circulated to supervisors and co-authors.

SECTION E

Student Signature	[redacted]
Date	26/06/2022

Supervisor Signature	[redacted]
Date	26/6/2022

Title: PrimedInclusivity. A programmable framework to assess the presence and impact of primer binding site nucleotide variation on NAAT-based assay performance.

Authors: Matthew Higgins [1,**], Susana Campino [1,*], Taane G. Clark [1,2,*]

1. Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK
2. Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* Joint authors

** Correspondence: matthew.higgins@lshtm.ac.uk, Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Abstract

Motivation: Nucleic acid amplification technologies (NAATs) have become fundamental to biological research, including diagnostic development and whole genome sequencing preparation. PrimedInclusivity is a python-based programmable framework, enabling researchers to detect and infer the impact of primer binding site genetic variation on NAAT-based assay performance.

Results: Here we demonstrate the framework's utility when designing novel NAAT-assays for *Plasmodium falciparum,* the deadliest malaria parasite, utilising publicly available whole genome sequence data from 5,668 samples, covering 28 countries. PrimedInclusivity enables users to improve assay-design and avoid non-specific amplification.

Availability and Implementation: The framework is available at {{https://github.com/MatthewHiggins2017/Primed_Inclusivity}} and supported on Linux and Macintosh operating systems.

Contact: matthew.higgins@lshtm.ac.uk

Introduction

Nucleic acid amplification technologies (NAATs) form a cornerstone of molecular biology and diagnostics. Polymerase chain reaction (PCR), developed in the 1980s remains the most popular technique to date and has been ubiquitously implemented in the detection of many pathogens, including HIV-1 and SARS-CoV-2 viruses [1,2,3]. Other NAATs have been continually developed, such as Loop mediated Isothermal Amplification (LAMP) and Recombinase Polymerase Amplification (RPA) pioneered in 2000 and 2006 respectively [4,5]. A fundamental strength shared by all NAATs is their plasticity stemming from their ease of design and exchange of primers. Whilst primers provide NAATs with their greatest strength, this reliance acts as a shared "achilles heel". It is common practice to design primers against a conserved sequence, ensuring full complementarity; however, this assumes that no genetic variation exists in the primer binding site. Population diversity is intrinsic to species survival and evolution [6,7], and the assumption of primer binding site conservation can be violated, leading to the formation of nucleotide mismatches within the primer-template complex [8,9]. Across NAATs, such mismatches are known to be detrimental to amplification and, when not accounted for, can invalidate several NAAT-based analyses, including nucleic acid quantification and single nucleotide polymorphism (SNP) genotyping [8,10]. Historically, this issue has remained unaddressed due to a lack of genomic data. However, reductions in the costs of next-generation sequencing have led to an abundance of available genomic data, compared to when NAATs were first pioneered [11]. As such it is possible to incorporate this data into the design of next generation NAAT-based diagnostics to combat infectious diseases such as malaria, more effectively. [12]

Malaria is a global disease affecting 241 million people annually and resulting in an estimated 627,000 deaths in 2020 [13]. Nearly half the world's population is at risk of infection

and the malaria burden is set to increase due to the ongoing SARS-CoV-2 pandemic and climate change [14,15]. Six *Plasmodium* species have been identified to infect humans. *P. falciparum* is associated with the highest mortality whilst *P. vivax* is the most widely distributed [16]. Historically NAATs have played a key role in combating malaria from fundamental speciation diagnostics to targeted DNA enrichment for whole genome sequencing [17,18]. As countries push towards achieving malaria free status, the utilisation of NAAT-based diagnostics is of the utmost importance due their ability to detect sub-microscopic infections compared to other methodologies such as microscopy or immunochromatic rapid diagnostic tests (RDTs) [19,20]. In addition, earlier diagnosis and subsequently early treatment improves clinical outcomes, reducing the risk of complicated malaria and subsequently mortality [21]. Existing malaria diagnostics have been shown to exert strong selective pressures on *Plasmodium* populations, leading to the emergence and spread of diagnostic-evasive strains, including those which lack pfDHR2/3 antigen targeted by commonly used RDTs [22]. As such when designing and assessing the next generation of diagnostics, researchers need to factor in the genomic data available to ensure diagnostics inclusivity and specificity. By incorporating all available data from across the globe, a robust diagnostic can be created which is important when the origin of infection may not be known. To achieve this, we have developed the PrimedInclusivity software tool, enabling researchers to account for possible binding site diversity and capture the subsequent impact on NAAT performance. Here we demonstrate the utility of the PrimedInclusivity tool through improving NAAT-based assays for the pathogen *P. falciparum*.

System and Methods

PrimedInclusivity

PrimedInclusivity (https://github.com/MatthewHiggins2017/Primed_Inclusivity) is a python-based programmable framework, enabling users to assess the conservation of primer/probes binding sites, for a predefined assay set, when considering the genetic diversity of the target organism, represented by a multi-sample VCF file. Primers are typically designed using a single genomic reference to facilitate complete Watson-Crick complementary binding [23]. Under default settings, PrimedInclusivity will quantify the proportion of samples, defined in the VCF, with reference matching, conserved complementary binding sites. In the presence of binding site variants, this analysis can be expanded to gauge their subsequent impact on variables associated with assay performance, such as the probability of reaction success or reaction efficiency. An overview of the framework schema is shown (**Figure 1**).

When assessing a given primer set, genetic variants which fall within each primer binding site are first identified, including single nucleotide polymorphisms (SNPs) or insertions / deletions (Indels). The presence of one or more variants will result in binding site heterogeneity, infringing Watson-Crick complementary primer binding. Each binding site sequence variant is subsequently extracted and classified via the customisable classification engine, **(Supplementary Information)**. Under default settings the classification engine simply characterises the presence or absence of nucleotide mismatches, however this functionality is expandable. Once all primers have been successfully classified, analysis continues on a sample-by-sample basis. For each sample, the associated binding site variant and classification string is identified for each primer. The classification strings are subsequently interpreted by the customisable output engine and sample-specific output values derived for each variable of interest for the entire primer set **(Supplementary Information)**.

Under default settings the output engine determines if one or more primer binding sites within a sample contains mismatches and if so the set is subsequently marked to fail Watson-Crick complementary primer binding.

After assessing *in-silico* the performance of a primer set against each individual sample, summary values are derived. PrimedInclusivity, allows users to assign samples into categories, for example, clustering samples according to country of origin. For each category, summary output values are calculated, which under default settings is the ratio of samples within the category with complementary binding, across all primers in a given set, hereafter referred to as the complementary binding ratio. The category-specific output summary values are then combined to generate population-wide summary values. During this consolidation, PrimedInclusivity, allows users to account for any underlying biases within the genomic dataset, according to differences in the distribution of samples within assigned categories and the known true distribution, via an iterative proportional fitting algorithm. For example, this allows us to adjust for the difference between the known spatial distribution of *P. falciparum* infections compared to the distribution of whole genome sequence data available **(Supplementary Table 3).**

Classification Engine

When assessing variant primer binding sites, PrimedInclusivity will extract two fixed variables; the target binding sequence and a Boolean array representing the presence or absence of Watson-Crick complementary binding per nucleotide position in the binding sequence. This information is used by the classification engine to generate a classification string. Each string can consist of multiple entries, with each entry composed of two components, a key and value. Users can add python-based custom modules to the

classification engine to interpret these two fixed variables and expand the classification string generated, which is subsequently linked to the output engine.

Output Engine

The output engine utilises the classification strings generated along with the user-defined output guide. Within the output guide, the user can define variables of interest and how these should be handled throughout the analysis. For example, if dealing with the probability of reaction success for a given sample, the product probability of all primers present in the set is required. In comparison, if you are simply looking at the presence or absence of perfect binding for a given primer set, the minimum perfect binding key value would be extracted from all primers in the set, as a Boolean value of 0, indicates the presence of mismatches in one or more primers in the set. The classification string is processed by the output engine to generate values according to the output variables of interest, defined in the output guide. Users can again customise the output engine by adding python-based modules to interpret any new classification string entries added.

Plug and Play System

The impact of mismatches within the primer-template complex will depend on the NAAT. Through the PrimedInclusivity platform, users can account for NAAT specific differences by adding custom modules to the classification and output engines as well as modifying the output guide file, which describes how the output engine should handle variables for interest. For example, the presence of a 3' terminal cytosine-cytosine mismatch in the primer-template complex has been shown to significantly affect the probability of RPA reaction success which utilises Bsu polymerase [5]. However, a polymerase with 3'-5' exonuclease activity would potentially be more tolerant to the same mismatch as theoretically it could be removed

prior to extension [24]. More information regarding the classification and output engines is provided (**Supplementary Information)**.

Implementation

Whilst PrimedInclusivity can be used to target any organism the following examples are associated with the detection of *Plasmodium spp*.

Optimising RPA Assay Inclusivity

The following RPA primer sets were designed in house, for a fluorescence-based assay to detect *P. falciparum* (PfRPA Set 1: FP1: 5'-CTATTTTGTCTATTTTGTATATTATAACCA, RP1: 5'-AAAAAATAATTTACAAAATGGTAATATCAG; PfRPA Set 2: FP2: 5'-CTGTTTGAGCATTAAATGAACAAATATCAT, RP2: 5'-CTTTGGATTTTTTAAAATTAAATT GTTCTG). Prior to the inclusion of a fluorescent probe, each primer set was successfully validated against the *P. falciparum 3D7* laboratory strain in line with standard practice (**Supplementary Figure 1)**. PrimedInclusivity was subsequently used to assess each set utilising publicly available *P. falciparum* whole genome sequence data, consisting of 5,668 samples in total from 28 countries **(Supplementary Information)**.

Through the addition of custom RPA modules, the *in-silico* assessment was expanded to not only quantify the abundance of binding site variants, but to gauge their subsequent impact on the probability of RPA reaction success, reaction onset time and efficiency, the latter metrics of which are essential for accurate fluorescent-based quantification [10]. This analysis was performed on a Macintosh operating system (2.3 GHz Dual-Core Intel Core i5 with 8GB of

RAM). The binding site variants identified within the *P. falciparum* population are presented (**Table 1)**. Binding site heterogeneity was identified for three out of the four primers screened. This included variants which would result in 3' terminal mismatches in the primer-template complex which are known to be detrimental to NAATs, including RPA [8].

The primers in PfRPA sets 1 and 2 were predicted to be complementary to 4,071 and 3,135 *P. falciparum* samples, respectively. When accounting for all 5,668 samples incorporated in this analysis, we can express these values as complementary binding ratios of 0.72 and 0.55, respectively (**Table 2)**. When adjusting our analysis to account for the known malaria burden across the 28 countries where the samples were collected (**Supplementary Tables 3)**, PfRPA Set 2 became the preferred primer set with an adjusted complementary binding ratio of 0.94, +0.44 greater than PfRPA set 1 which became 0.50 post adjustment. As expected, when considering the predicted probability of RPA reaction success, accounting for the position of the polymorphisms in the binding site, the PfRPA set 2 (P: 0.99) emerges as the preferred candidate compared to PfRPA set 1 (P: 0.89) (**Table 2)**, in line with the higher complementary binding ratio and as such should be prioritised for further optimisation. When breaking down the *in-silico* assessment on a country-by-country basis, the heterogeneity associated with RPA kinetics follows the pattern established by the complementary binding ratio within each country (**Supplementary Information).**

Optimisation of Taq PCR Specificity

PrimedInclusivity can also be used to augment NAAT assay optimization. The Taq PCR primer set (PfTaq FP: CCATTATCATGGATATCTGGATTGAT, PfTaq RP: GCATAGAATGCACACATAAACC) was designed with full complementarity to *P. falciparum 3D7* reference genome. Upon preliminary validation at the recommended annealing

temperature 49ºC derived by the NEB Tm Calculator, the primers were able to not only amplify *P. falciparum* but also *P. vivax* **(Supplementary Figure 2)**. Aligning the primer to *P. vivax P01* reference genome, revealed that mismatch sites existed in both the forward and reverse primer binding sites.

Theoretically, if we desired to make our assay specific to *P. falciparum*, we would systematically increase the annealing temperature used in the PCR cycle through a trial-and-error process until undesired amplification of *P.vivax* is eradicated. Increasing the annealing temperature decreases the primer-template binding fraction, which subsequently inhibits amplification. The primer binding fraction, at a given temperature, is underpinned by its thermodynamic stability, whereby the presence of nucleotide mismatches decreases the stability of the complex [31]. Therefore, assuming an off-target complex contains mismatches, the annealing temperature can be optimised such that the primer binding fraction for the off-target complex is reduced below the threshold required for successful amplification; whilst the binding fraction for the on-target complex with complementary binding, remains above this threshold ensuring specificity. However, this typical wet-lab approach is labour intensive, and a successful outcome is not guaranteed. In addition, optimisation is typically performed on a single target sample and, as such, the presence of primer binding site heterogeneity is not accounted for. To overcome these limitations, PrimedInclusivity can be used to guide assay optimization first *in-silico* by assessing the performance of the primer set against *P. falciparum* and *P. vivax* across a range of annealing temperatures, whilst accounting for binding site diversity. Once an optimal annealing temperature is identified, maximising the probability of reaction success for *P. falciparum* whilst minimising it for *P. vivax*, the optimal annealing temperature can be validated experimentally. The link between the primer binding fraction and the probability of reaction success was successfully characterised for NEB Taq2x PCR

(**Supplementary Information**) and custom Taq modules added to the PrimedInclusivity classification and output engines.

PrimedInclusivity was then utilised to gauge the probability of reaction success for the PfTaq primer set, against both *P. falciparum* and *P. vivax* across an annealing temperature range of 46-67ºC. Target and off-target population diversity was accounted for by including 5,668 and 846 publicly available samples for *P. falciparum* and *P. vivax* respectively (**Supplementary Information**). The binding site variants for *P. falciparum* and *P. vivax* are presented (**Table 3).** For *P. falciparum* the primer binding sites were conserved in 5,663 samples, indicating robust assay design with regards to inclusivity. As expected, no *P. vivax* samples were found to be complementary to the primer set.

The PrimedInclusivity derived probability of reaction success was extracted for both *Plasmodium spp.* across the annealing temperature range **(Figure 2)**. From the analysis, at the recommended NEB annealing temperature of 49ºC, amplification of both *P. falciparum* (P:0.96) and *P. vivax* (P:0.54) was to be expected. However, between the annealing temperatures of 55ºC and 60ºC the probability of reaction success for *P. vivax* is minimised, whilst the probability for *P. falciparum* ranged from P:0.90 to P:0.21. Guided by this analysis, the PfTaq primer set was screened *in-vitro* at 55ºC, 58ºC and 60ºC and *P. falciparum* specificity was successfully obtained when using annealing temperatures of 58ºC and 60ºC (**Figure 3).**

Discussion

As the shift towards NAAT-based malaria diagnostics has begun, the examples provided demonstrate the advantages of accounting for population diversity during assay design. The country-specific breakdown of primer set performance highlights that a one size fits all

approach may not always be practical and instead a tailored framework may be needed to account for other species and country specific diversity. This aligns with current understanding of disease transmission dynamics, whereby dominant strains becoming fixed once introduced to a new area. The country-specific approach will not only help laboratories in malaria endemic countries but also those dealing with imported cases when the country of origin is known such as tourists and migrant workers returning from endemic areas. The need to account for binding site diversity is exacerbated in the detection of other pathogens such as HIV and Ebola, which are known to have higher mutation rates, resulting in a higher-level of population diversity [32,33]. The versatility of the PrimedInclusivity framework was demonstrated in guiding laboratory optimisation to overcome issues of specificity. Such issues arise when dealing with closely related and cryptic species, which are common in neglected pathogens and vectors (e.g., *Anopheles* mosquitoes) [34]. Through making PrimedInclusivity software open-source, the framework can be continuously expanded and improved by the user-base, through the addition of custom modules. In summary, researchers now have a user-friendly medium to incorporate genomic surveillance data into next generation diagnostic design empowering those in infectious disease control and elimination.

Funding

This work was supported by a BBSRC LIDO PHD studentship (M.H). T.G.C was funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1). S.C was funded by Medical Research Council UK grants (ref. MR/M01360X/1, MR/R025576/1, and MR/R020973/1). The authors declare no conflicts of interest.

References

1. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51 Pt 1**, 263–273 (1986).
2. Jangam, S. R., Agarwal, A. K., Sur, K. & Kelso, D. M. A point-of-care PCR test for HIV-1 detection in resource-limitedsettings. *Biosensors and Bioelectronics* **42**, 69–75 (2013).
3. Treibel, T. A. *et al.* COVID-19: PCR screening of asymptomatic health-care workers at London hospital. *Lancet* **395**, 1608–1610 (2020).
4. Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res.* **28**, E63 (2000).
5. Piepenburg, O., Williams, C. H., Stemple, D. L. & Armes, N. A. DNA detection using recombination proteins. *PLoS Biol.* **4**, e204 (2006).
6. Reed, D. H. & Frankham, R. Correlation between Fitness and Genetic Diversity. *Conserv. Biol.* **17**, 230–237 (2003).
7. Ward, D. *et al.* An integrated in silico immuno-genetic analytical platform provides insights into COVID-19 serological and vaccine targets. *Genome Med.* **13**, 4 (2021).
8. Stadhouders, R. *et al.* The Effect of Primer-Template Mismatches on the Detection and Quantification of Nucleic Acids Using the 5′ Nuclease Assay. *J. Mol. Diagn.* **12**, 109–117 (2010).
9. Brown, K. A. *et al.* S-Gene Target Failure as a Marker of Variant B.1.1.7 Among SARS-CoV-2 Isolates in the Greater Toronto Area, December 2020 to March 2021. *JAMA* **325**, 2115–2116 (2021).
10. Crannell, Z. A., Rohrman, B. & Richards-Kortum, R. Development of a quantitative recombinase polymerase amplification assay with an internal positive control. *J. Vis. Exp.* (2015) doi:10.3791/52620.
11. Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 (2016).
12. Volkman, S. K., Neafsey, D. E., Schaffner, S. F., Park, D. J. & Wirth, D. F. Harnessing genomics and genome biology to understand malaria biology. *Nat. Rev. Genet.* **13**, 315–328 (2012).

13. Organization, W. H. & Others. World malaria report 2021. (2021).
14. Caminade, C. *et al.* Impact of climate change on global malaria distribution. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 3286–3291 (2014).
15. Weiss, D. J. *et al.* Indirect effects of the COVID-19 pandemic on malaria intervention coverage, morbidity, and mortality in Africa: a geospatial modelling analysis. *Lancet Infect. Dis.* **21**, 59–69 (2021).
16. Phillips, M. A. *et al.* Malaria. *Nat Rev Dis Primers* **3**, 17050 (2017).
17. Ibrahim, A. *et al.* Selective whole genome amplification of Plasmodium malariae DNA from clinical samples reveals insights into population structure. *Sci. Rep.* **10**, 10832 (2020).
18. Li, P. *et al.* Nested PCR detection of malaria directly using blood filter paper samples from epidemiological surveys. *Malar. J.* **13**, 175 (2014).
19. Hofmann, N. *et al.* Ultra-sensitive detection of Plasmodium falciparum by amplification of multi-copy subtelomeric targets. *PLoS Med.* **12**, e1001788 (2015).
20. Landier, J. *et al.* Effect of generalised access to early diagnosis and treatment and targeted mass drug administration on Plasmodium falciparum malaria in Eastern Myanmar: an observational study of a regional elimination programme. *Lancet* **391**, 1916–1926 (2018).
21. Mousa, A. *et al.* The impact of delayed treatment of uncomplicated P. falciparum malaria on progression to severe malaria: A systematic review and a pooled multicentre individual-patient meta-analysis. *PLoS Med.* **17**, e1003359 (2020).
22. Berzosa, P. *et al.* First evidence of the deletion in the pfhrp2 and pfhrp3 genes in Plasmodium falciparum from Equatorial Guinea. *Malar. J.* **19**, 99 (2020).
23. Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71–4 (2007).
24. Ricardo, P. C., Françoso, E. & Arias, M. C. Fidelity of DNA polymerases in the detection of intraindividual variation of mitochondrial DNA. *Mitochondrial DNA B Resour* **5**, 108–112 (2019).
25. 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).

26. Benavente, E. D. *et al.* Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. *Nat. Commun.* **12**, 3160 (2021).
27. MalariaGEN *et al.* An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. *Wellcome Open Res* **6**, 42 (2021).
28. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
29. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
30. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
31. SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1460–1465 (1998).
32. Andrews, S. M. & Rowland-Jones, S. Recent advances in understanding HIV evolution. *F1000Res.* **6**, 597 (2017).
33. Whitfield, Z. J. *et al.* Species-Specific Evolution of Ebola Virus during Replication in Human and Bat Cells. *Cell Rep.* **32**, 108028 (2020).
34. Tennessen, J. A. *et al.* A population genomic unveiling of a new cryptic mosquito taxon within the malaria-transmitting Anopheles gambiae complex. *Mol. Ecol.* **30**, 775–790 (2021).

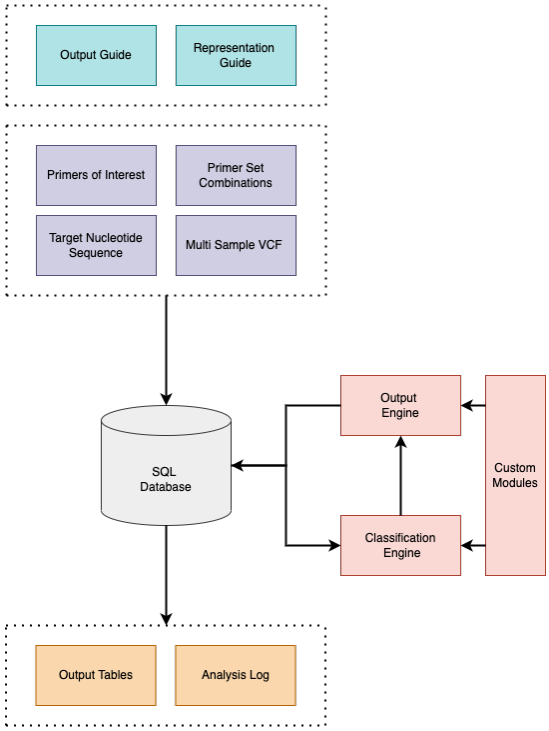


Figure 1. The schema of PrimedInclusivity.

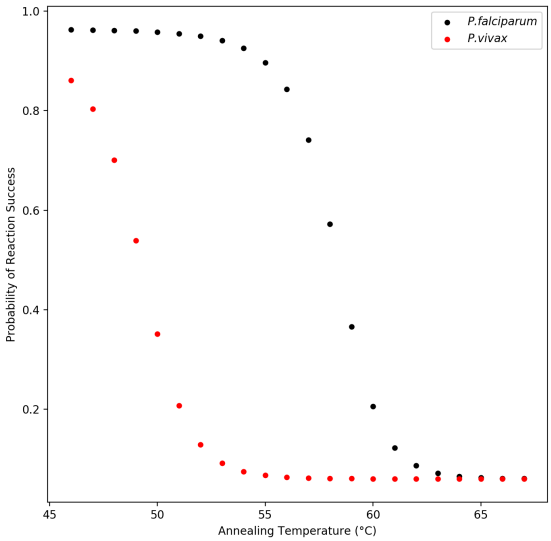


Figure 2. Predicted probability of reaction success (P) for Pf.Taq primer set against *P. falciparum 3D7* (Pf) and *P. vivax P01* (Pv).

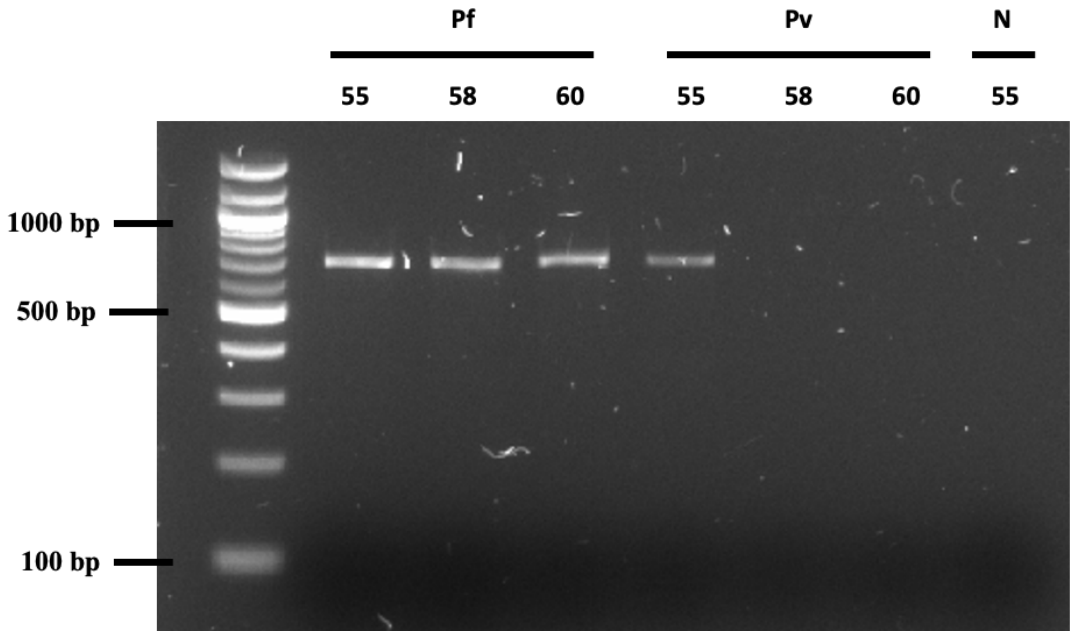


Figure 3. Screening of PfTaq primer set across annealing temperatures of 55°C, 58°C and 60°C, against *P. falciparum 3D7* (Pf) and *P. vivax P01* (Pv).

Table 1. Highlights the variant binding sites identified within the *P. falciparum* population. Variant positions resulting in a primer-template mismatch are shown in bold and underlined. Forward Primer (FP1), Reverse Primer (RP1).

Set	Primer	Binding Site Variant (5'-3')	
PfRPA Set 1	FP1	0	GATAAAACAGATAAAACATATAATATTGGT
		1	GATAAAACAGATAAAACATATAATATTGG**A**
		2	GATAAAA**A**AGATAAAACATATAATATTGGT
		3	GATAAAACAGATAAAACATATAATATT**A**G**A**
		4	GATAAAACAG**G**TAAAACATATAATATTGG**A**
		5	G**G**TAAAACAGATAAAACATATAATATTGGT
	RP1	0	TTTTTTATTAAATGTTTTACCATTATAGTC
		1	TT**C**TTTATTAAATGTTTTACCATTATAGTC
		2	TTT**A**TTATTAAATGTTTTACCATTATAGTC
		3	TTTTTTATT**T**AATGTTTTACCATTATAGTC
		4	TTTTTTATTAAATGTTTTACCATTATA**C**TC
		5	TTTTTTATTAAATGTTTTACCATTATAGT**G**
PfRPA Set 2	FP2	0	GACAAACTCGTAATTTACTTGTTTATAGTA
		1	GACAAACTCGTAATTTACTTGTTTATAGT**T**

		2	GACAAACTCGTAATTTACTTGTTTATA**T**TA
		3	GACAAACTCGTAATTTACTTGTTTATA**T**T**T**
		4	GACAAACTCGTAATTTACTTGTT**A**ATAGTA
		5	GA**A**AAACTCGTAATTTACTTGTTTATAGTA
	RP2	0	GAAACCTAAAAAATTTTAATTTAACAAGAC

Table 2. Outcome for PrimedInclusivity assessment of PfRPA Set 1 and 2. Corrected values are shown in brackets after adjusting for the distribution of malaria disease burden across the 28 countries included in this analysis (**Supplementary Table 3).**

	PfRPA Set 1	**PfRPA Set 2**
Ratio of samples with complementary binding	0.72 (0.50)	0.55 (0.94)
Probability of reaction success (P)	0.94 (0.89)	0.96 (0.99)
Mean decrease in reaction efficiency (%)	11.12 (19.65)	13.20 (1.63)
Mean increase in reaction onset time (%)	7.87 (13.91)	7.14 (0.88)

Table 3. PfTaq set binding sites variants in the *P. falciparum* population. Forward Primer (FP), Reverse Primer (RP).

Set	Target	Primer	Binding Sequence	
PfTaq	Pf	FP	0	GGTAATAGTACCTATAGACCTAACTA
			1	GGTAATAGTACCTATAGACC**C**AACTA
			2	GGT**G**ATAGTACCTATAGACCTAACTA
		RP	0	CGTATCTTACGTGTGTATTTGG
			1	CGTATCTTAC**A**TGTGTATTTGG
	Pv	FP	0	GGTAATAGTACCTATAG**T**CCTAA**T**TA
		RP	0	CGTATCTTAC**T**TGTGTATTTGG
			1	CGTATCTTAC**T**TGTGTA**C**TTGG

Supplementary Information

Plasmodium Samples

Publicly available whole genome sequencing data for 5,668 *P. falciparum* and 846 *P. vivax* samples were included in our investigation, made available from MalariaGEN (https://www.malariagen.net/) [1–3] . The samples were obtained from a total of 38 countries including Papua New Guinea, Colombia, Malawi, Benin, Côte d'Ivoire, Indonesia, Myanmar, India, China, Bhutan, Kenya, Gabon, Ethiopia, Mauritania, Vietnam, Burkina Faso, Gambia, Cameroon, Bangladesh, Mali, Madagascar, Thailand, Laos, Senegal, Nicaragua, Cambodia, Tanzania, Mexico, Peru, Brazil, Democratic Republic of the Congo, Guinea, Panama, Nigeria, Ghana, Malaysia, Uganda and Sri Lanka.

Mapping and Variant Calling

Raw illumina reads were mapped against the respective reference, *P. falciparum 3D7* or *P. vivax P01,* using bwa-mem software under default settings and indexed using samtools [4,5]. Variants were subsequently called using GATK software and filtered as defined previously[6]. Mixed infections were identified and excluded from downstream analysis using a custom python script.

Taq Polymerase DNA Amplification

PCR was performed using the NEB Taq 2x master mix. Briefly, reactions consisted of 6.5ul Taq2x Master Mix, 0.25ul of 100uM of each primer, 5ul of nuclease free water and 1ul of template, buffered in nuclease free water. Once set up, reactions were carried out in a G-Storm Thermocycler, following the standard thermocycling conditions for Taq Polymerase. On reaction completion successful amplification was determined by agarose gel electrophoresis.

Recombinase Polymerase DNA Amplification.

RPA was performed according to the manufacturer's protocol using the TwistAmp Basic kit. Reactions were incubated at 39ºC for 20 minutes followed by a denaturing step of 99ºC for 30 minutes. On reaction completion successful amplification was determined by agarose gel electrophoresis.

Thermodynamic Calculations

The ratio of annealed primer at a specified temperature was determined according to the Nearest-Neighbour thermodynamic model, utilising entropy and enthalpy metrics derived

previously for each given nucleotide pair [7]. Typically, the Nearest-Neighbour thermodynamic model is used to calculate the melting temperature for a given primer-target complex which corresponds to the temperature at which the ratio of bound to free primer is 0.5. With this understanding we can reverse the equation to determine at any given temperature what is the subsequent ratio of annealed primer. In addition, the stability and therefore the melting temperature of a given primer-target complex will be influenced by buffer conditions, including but not limited to the presence and concentration of cations which stabilise primer-target complex formation. As such buffer conditions were defined in line with the NEB Taq2x product notes: 0.2uM primer, 0.8mM dNTPs, 1.5mM divalent cations and 50mM monovalent cations.

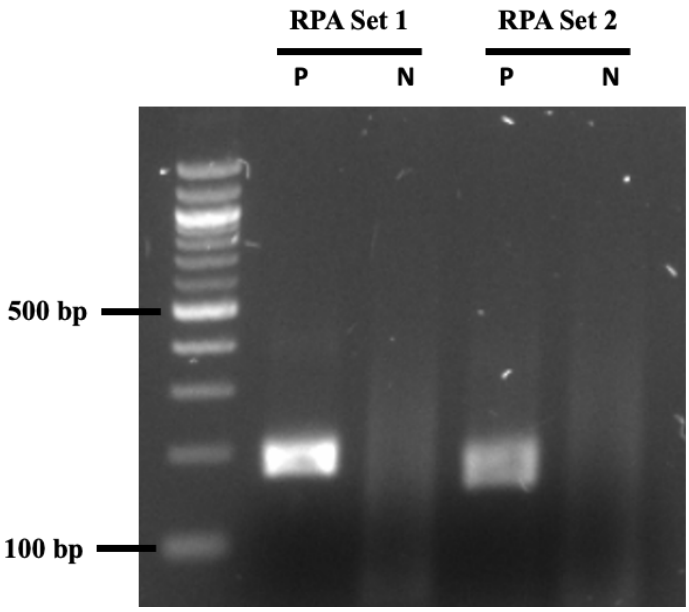
Country Specific Performance of RPA Primer Set

When breaking down the *in-silico* assessment of PfRPA Set 1 and 2 on a country-by-country basis, the heterogeneity associated with RPA kinetics follows the pattern established by the complementary binding ratio of each individual country (**Supplementary Figures 2,3).** For PfRPA Set 1, the complementary binding ratio was highest for countries in Southeast Asia, including Cambodia. In comparison for PfRPA Set 2, countries with the highest complementary binding ratios were found in West Africa, including Burkina Faso and Mali. Consequently a ≥20% difference in reaction efficiency is predicted for PfRPA Set 1 when used on samples from West Africa compared to Southeast Asia. If carried forward for fluorescence-based quantification, such a difference will affect downstream analysis which relies on the reaction kinetic profile and lead to system under-quantification of *P. falciparum* and could reduce assay-sensitivity to low-parasitemia infections. This highlights the importance of accounting for the target organism's regional population diversity.

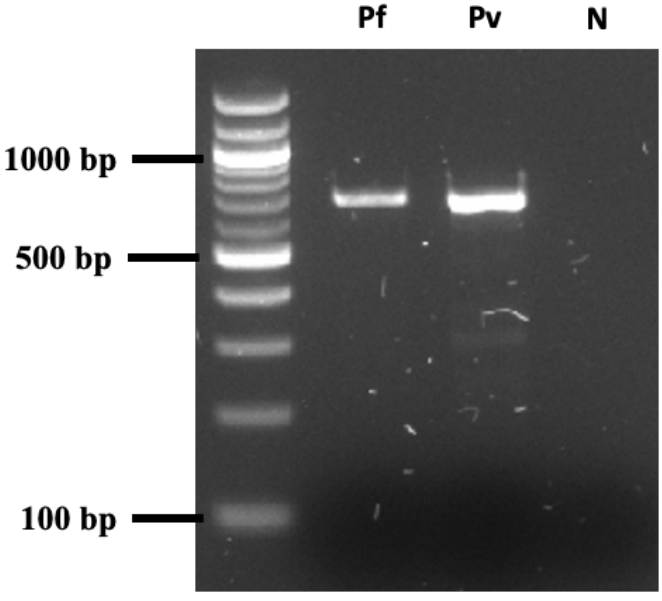
Characterising Primer Binding Ratio and Probability of Reaction Success

To characterise the link between the primer binding fraction and the probability of reaction success for Taq-based NAATs (TQ), 7 *P. falciparum* target primer sets were designed and screened across a range of annealing temperatures (45-70ºC) and template copy numbers (1, 10, 1000) (**Supplementary Table 1)**. Each reaction was assigned a binary outcome depending on the presence of an amplicon band when assessed via gel-electrophoresis (**Supplementary Table 2).**

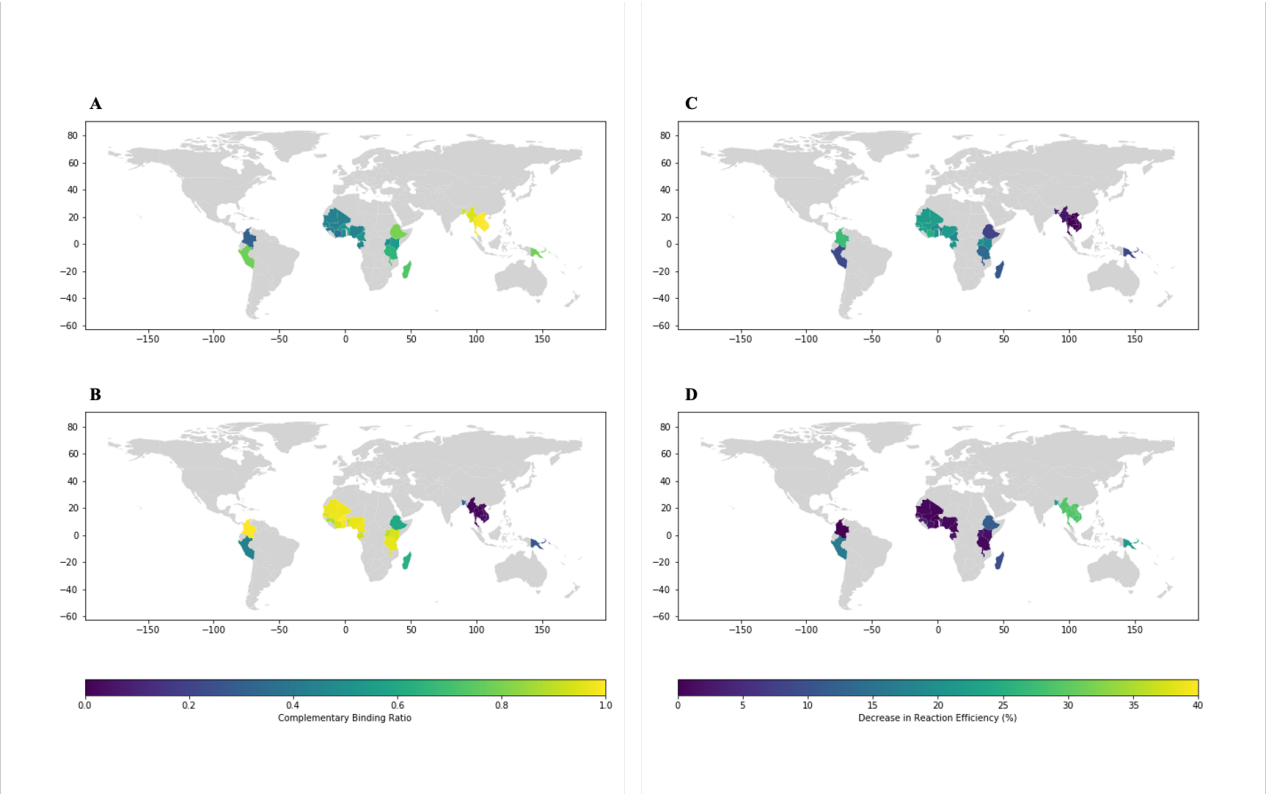
To estimate the primer binding fraction for a primer-template complex at a given temperature we can use the Nearest-Neighbour model[7]. The predicted binding fraction for each TQ primer (**Supplementary Table 1)** across the annealing temperature gradient is shown (**Supplementary Figure 5)**. The primer sets were specifically designed to maximise the binding fraction fluctuation of the forward primer in each set across the annealing temperature range whilst keeping the binding fraction for the reverse primer relatively constant. The predicted binding fraction and reaction outcome classification (**Supplementary Table 2)**, was used to fit a logistic model (Area under the ROC Curve (AUC): 0.865) via the Scipy python package (**Supplementary Figure 6).** The model was subsequently incorporated into custom Taq modules for PrimedInclusivities classification and output engines.



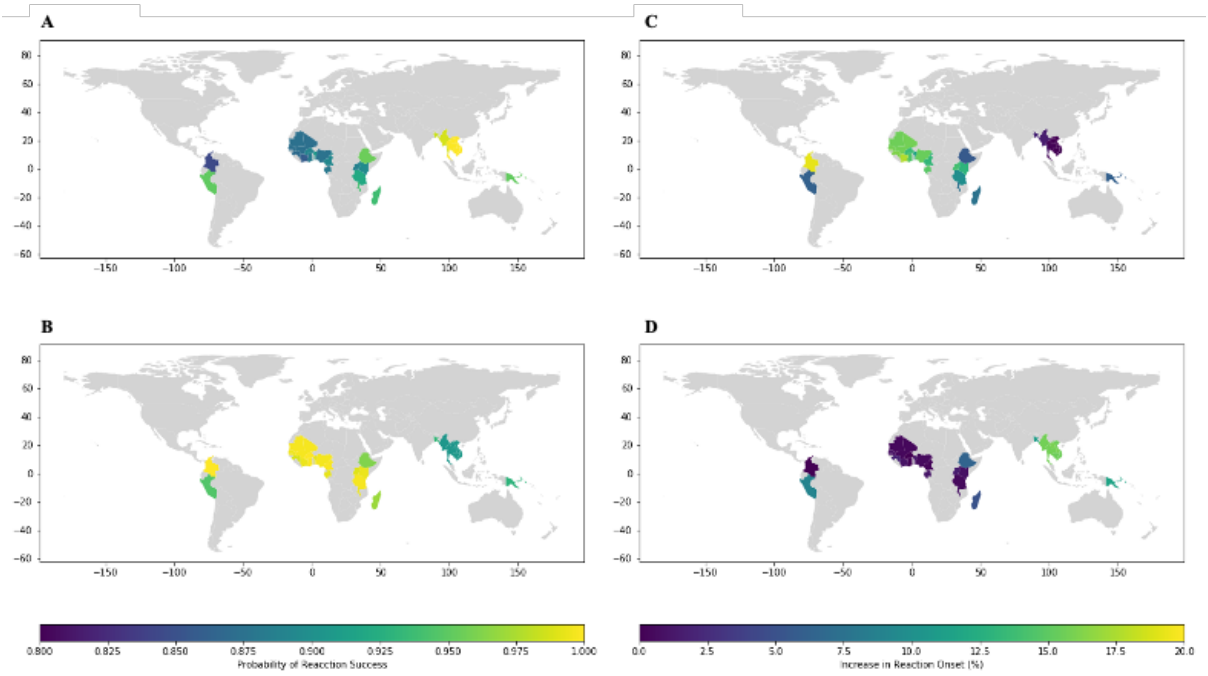
Supplementary Figure 1. Successful screening of PfRPA primer sets one and two against *P. falciparum* 3D7 (P). Nuclease free water was used as a template in negative control reactions (N).



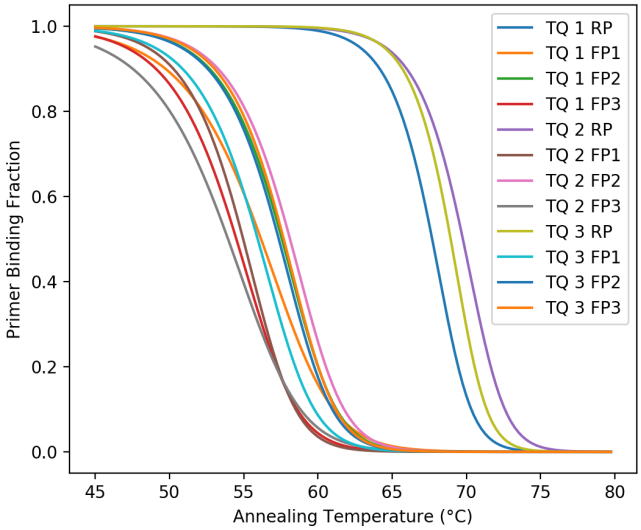
Supplementary Figure 2. Screening of Pf. Taq primer set on *P. falciparum 3D7* (Pf) and *P. vivax P01* (Pv). Negative control (N).



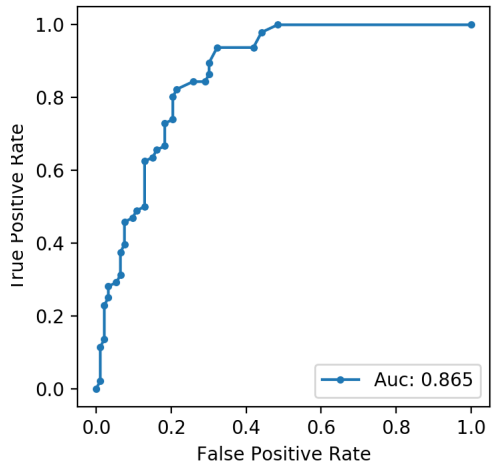
Supplementary Figure 3. Country-specific heatmaps for the proportion of samples with complementary binding ratio (A, B) and expected decrease in reaction efficiency (C, D). Plots (A,C) and (B, D) correspond to PfRPA Set 1 and 2 respectively. Countries shown in grey were not included in the analysis.



Supplementary Figure 4. Country-specific heatmaps for the probability of reaction success (A, B) and expected increase in reaction onset (C, D). Plots (A, C) and (B, D) correspond to PfRPA Set 1 and 2 respectively. Countries shown in grey were not included in the analysis.



Supplementary Figure 5. The predicted primer binding fraction across the annealing temperature range for each Taq (TQ) screening primer according to John Santa Lucia Nearest Neighbour model.



Supplementary Figure 6. Receiver operating characteristic (ROC) analysis to assess reaction success prediction model performance, Area Under Curve (AUC).

Supplementary Table 1. Primers used to investigate the link between primer binding fraction and the probability of reaction success for Taq (TQ) Polymerase PCR. Forward Primer (FP) Reverse Primer (RP).

Primer ID	Sequence (5'-3')	Set
TQ 1 RP	CCCCAATAACTCATTTGACCCCATGGTAAGAC	1,2,3
TQ 1 FP1	CGCAACAGGTGCTTCT	1
TQ 1 FP2	GAGAATTATGGAGTGGATGGTG	2
TQ 1 FP3	GCAAGTCGATATACACCAG	3
TQ 2 RP	CAGTCCCAGCGACAGCGGTTATACTTTGG	4,5
TQ 2 FP2	CGCCCTTAACGTAAAGATCATT	4
TQ 2 FP3	GAAACAGCCGGAAAGG	5
TQ 3 RP	CCTTACGGTCTGATTTGTTCCGCTCAATACTCAG	6,7
TQ 3 FP2	GGTTTATGTGTGCATTCTATGC	6
TQ 3 FP3	GGTGCTAGAGATTATTCTGTTCC	7

Supplementary Table 2. Assessment of Taq (TQ) primer sets amplification outcome and predicted primer binding fraction.

Primer	Annealing Temp (°C)	Template Copy Number log10	Successful Amplification	Binding Fraction
TQ_1_FP1	45	3	1	0.9755
TQ_1_FP2	45	3	1	0.9954
TQ_1_FP3	45	3	1	0.9762
TQ_2_FP2	45	3	1	0.9964
TQ_2_FP3	45	3	1	0.9523
TQ_3_FP2	45	3	1	0.9954
TQ_3_FP3	45	3	1	0.9965
TQ_1_FP1	45	2	1	0.9755
TQ_1_FP2	45	2	1	0.9954
TQ_1_FP3	45	2	1	0.9762
TQ_2_FP2	45	2	1	0.9964
TQ_2_FP3	45	2	1	0.9523
TQ_3_FP2	45	2	1	0.9954
TQ_3_FP3	45	2	1	0.9965
TQ_1_FP1	45	0	0	0.9755
TQ_1_FP2	45	0	0	0.9954
TQ_1_FP3	45	0	1	0.9762
TQ_2_FP2	45	0	0	0.9964
TQ_2_FP3	45	0	0	0.9523
TQ_3_FP2	45	0	1	0.9954
TQ_3_FP3	45	0	1	0.9965
TQ_1_FP1	50	3	1	0.8921

TQ_1_FP2	50	3	1	0.9647
TQ_1_FP3	50	3	1	0.8642
TQ_2_FP2	50	3	1	0.9722
TQ_2_FP3	50	3	1	0.8016
TQ_3_FP2	50	3	1	0.9636
TQ_3_FP3	50	3	1	0.9707
TQ_1_FP1	50	2	1	0.8921
TQ_1_FP2	50	2	1	0.9647
TQ_1_FP3	50	2	1	0.8642
TQ_2_FP2	50	2	1	0.9722
TQ_2_FP3	50	2	1	0.8016
TQ_3_FP2	50	2	1	0.9636
TQ_3_FP3	50	2	1	0.9707
TQ_1_FP1	50	0	0	0.8921
TQ_1_FP2	50	0	1	0.9647
TQ_1_FP3	50	0	1	0.8642
TQ_2_FP2	50	0	0	0.9722
TQ_2_FP3	50	0	0	0.8016
TQ_3_FP2	50	0	1	0.9636
TQ_3_FP3	50	0	1	0.9707
TQ_1_FP1	52.5	3	1	0.7862
TQ_1_FP2	52.5	3	1	0.9071
TQ_1_FP3	52.5	3	1	0.7044
TQ_2_FP2	52.5	3	1	0.9256
TQ_2_FP3	52.5	3	1	0.6315

TQ_3_FP2	52.5	3	1	0.9027
TQ_3_FP3	52.5	3	1	0.9190
TQ_1_FP1	52.5	2	1	0.7862
TQ_1_FP2	52.5	2	1	0.9071
TQ_1_FP3	52.5	2	1	0.7044
TQ_2_FP2	52.5	2	1	0.9256
TQ_2_FP3	52.5	2	0	0.6315
TQ_3_FP2	52.5	2	1	0.9027
TQ_3_FP3	52.5	2	1	0.9190
TQ_1_FP1	52.5	0	1	0.7862
TQ_1_FP2	52.5	0	1	0.9071
TQ_1_FP3	52.5	0	1	0.7044
TQ_2_FP2	52.5	0	0	0.9256
TQ_2_FP3	52.5	0	0	0.6315
TQ_3_FP2	52.5	0	1	0.9027
TQ_3_FP3	52.5	0	1	0.9190
TQ_1_FP1	55	3	1	0.6079
TQ_1_FP2	55	3	1	0.7711
TQ_1_FP3	55	3	1	0.4424
TQ_2_FP2	55	3	1	0.8118
TQ_2_FP3	55	3	1	0.3962
TQ_3_FP2	55	3	1	0.7576
TQ_3_FP3	55	3	1	0.7896
TQ_1_FP1	55	2	1	0.6079
TQ_1_FP2	55	2	1	0.7711

TQ_1_FP3	55	2	1	0.4424
TQ_2_FP2	55	2	0	0.8118
TQ_2_FP3	55	2	0	0.3962
TQ_3_FP2	55	2	1	0.7576
TQ_3_FP3	55	2	1	0.7896
TQ_1_FP1	55	0	0	0.6079
TQ_1_FP2	55	0	0	0.7711
TQ_1_FP3	55	0	1	0.4424
TQ_2_FP2	55	0	0	0.8118
TQ_2_FP3	55	0	0	0.3962
TQ_3_FP2	55	0	1	0.7576
TQ_3_FP3	55	0	1	0.7896
TQ_1_FP1	57.5	3	1	0.3694
TQ_1_FP2	57.5	3	1	0.5099
TQ_1_FP3	57.5	3	1	0.1754
TQ_2_FP2	57.5	3	1	0.5778
TQ_2_FP3	57.5	3	1	0.1761
TQ_3_FP2	57.5	3	1	0.4822
TQ_3_FP3	57.5	3	1	0.5262
TQ_1_FP1	57.5	2	1	0.3694
TQ_1_FP2	57.5	2	1	0.5099
TQ_1_FP3	57.5	2	1	0.1754
TQ_2_FP2	57.5	2	0	0.5778
TQ_2_FP3	57.5	2	0	0.1761
TQ_3_FP2	57.5	2	1	0.4822

TQ_3_FP3	57.5	2	1	0.5262
TQ_1_FP1	57.5	0	0	0.3694
TQ_1_FP2	57.5	0	1	0.5099
TQ_1_FP3	57.5	0	1	0.1754
TQ_2_FP2	57.5	0	0	0.5778
TQ_2_FP3	57.5	0	0	0.1761
TQ_3_FP2	57.5	0	1	0.4822
TQ_3_FP3	57.5	0	1	0.5262
TQ_1_FP1	60	3	1	0.1581
TQ_1_FP2	60	3	1	0.2011
TQ_1_FP3	60	3	1	0.0439
TQ_2_FP2	60	3	1	0.2591
TQ_2_FP3	60	3	1	0.0564
TQ_3_FP2	60	3	1	0.1756
TQ_3_FP3	60	3	1	0.2022
TQ_1_FP1	60	2	1	0.1581
TQ_1_FP2	60	2	1	0.2011
TQ_1_FP3	60	2	1	0.0439
TQ_2_FP2	60	2	0	0.2591
TQ_2_FP3	60	2	0	0.0564
TQ_3_FP2	60	2	1	0.1756
TQ_3_FP3	60	2	1	0.2022
TQ_1_FP1	60	0	0	0.1581
TQ_1_FP2	60	0	1	0.2011
TQ_1_FP3	60	0	1	0.0439

TQ_2_FP2	60	0	0	0.2591
TQ_2_FP3	60	0	0	0.0564
TQ_3_FP2	60	0	1	0.1756
TQ_3_FP3	60	0	1	0.2022
TQ_1_FP1	62.5	3	0	0.0495
TQ_1_FP2	62.5	3	1	0.0440
TQ_1_FP3	62.5	3	1	0.0090
TQ_2_FP2	62.5	3	1	0.0622
TQ_2_FP3	62.5	3	0	0.0153
TQ_3_FP2	62.5	3	1	0.0355
TQ_3_FP3	62.5	3	1	0.0408
TQ_1_FP1	62.5	2	0	0.0495
TQ_1_FP2	62.5	2	1	0.0440
TQ_1_FP3	62.5	2	1	0.0090
TQ_2_FP2	62.5	2	0	0.0622
TQ_2_FP3	62.5	2	0	0.0153
TQ_3_FP2	62.5	2	1	0.0355
TQ_3_FP3	62.5	2	1	0.0408
TQ_1_FP1	62.5	0	0	0.0495
TQ_1_FP2	62.5	0	0	0.0440
TQ_1_FP3	62.5	0	0	0.0090
TQ_2_FP2	62.5	0	0	0.0622
TQ_2_FP3	62.5	0	0	0.0153
TQ_3_FP2	62.5	0	0	0.0355
TQ_3_FP3	62.5	0	0	0.0408

TQ_1_FP1	65	3	0	0.0133
TQ_1_FP2	65	3	1	0.0075
TQ_1_FP3	65	3	0	0.0018
TQ_2_FP2	65	3	0	0.0107
TQ_2_FP3	65	3	0	0.0040
TQ_3_FP2	65	3	1	0.0057
TQ_3_FP3	65	3	1	0.0063
TQ_1_FP1	65	2	0	0.0133
TQ_1_FP2	65	2	1	0.0075
TQ_1_FP3	65	2	0	0.0018
TQ_2_FP2	65	2	0	0.0107
TQ_2_FP3	65	2	0	0.0040
TQ_3_FP2	65	2	0	0.0057
TQ_3_FP3	65	2	0	0.0063
TQ_1_FP1	65	0	0	0.0133
TQ_1_FP2	65	0	0	0.0075
TQ_1_FP3	65	0	0	0.0018
TQ_2_FP2	65	0	0	0.0107
TQ_2_FP3	65	0	0	0.0040
TQ_3_FP2	65	0	0	0.0057
TQ_3_FP3	65	0	0	0.0063
TQ_1_FP1	67.5	3	0	0.0035
TQ_1_FP2	67.5	3	0	0.0012
TQ_1_FP3	67.5	3	0	0.0004
TQ_2_FP2	67.5	3	0	0.0017

TQ_2_FP3	67.5	3	0	0.0010
TQ_3_FP2	67.5	3	0	0.0009
TQ_3_FP3	67.5	3	0	0.0009
TQ_1_FP1	67.5	2	0	0.0035
TQ_1_FP2	67.5	2	0	0.0012
TQ_1_FP3	67.5	2	0	0.0004
TQ_2_FP2	67.5	2	0	0.0017
TQ_2_FP3	67.5	2	0	0.0010
TQ_3_FP2	67.5	2	0	0.0009
TQ_3_FP3	67.5	2	0	0.0009
TQ_1_FP1	67.5	0	0	0.0035
TQ_1_FP2	67.5	0	0	0.0012
TQ_1_FP3	67.5	0	0	0.0004
TQ_2_FP2	67.5	0	0	0.0017
TQ_2_FP3	67.5	0	0	0.0010
TQ_3_FP2	67.5	0	0	0.0009
TQ_3_FP3	67.5	0	0	0.0009
TQ_1_FP1	70	3	0	0.0009
TQ_1_FP2	70	3	0	0.0002
TQ_1_FP3	70	3	0	0.0001
TQ_2_FP2	70	3	0	0.0003
TQ_2_FP3	70	3	0	0.0003
TQ_3_FP2	70	3	0	0.0001
TQ_3_FP3	70	3	0	0.0001
TQ_1_FP1	70	2	0	0.0009

TQ_1_FP2	70	2	0	0.0002
TQ_1_FP3	70	2	0	0.0001
TQ_2_FP2	70	2	0	0.0003
TQ_2_FP3	70	2	0	0.0003
TQ_3_FP2	70	2	0	0.0001
TQ_3_FP3	70	2	0	0.0001
TQ_1_FP1	70	0	0	0.0009
TQ_1_FP2	70	0	0	0.0002
TQ_1_FP3	70	0	0	0.0001
TQ_2_FP2	70	0	0	0.0003
TQ_2_FP3	70	0	0	0.0003
TQ_3_FP2	70	0	0	0.0001
TQ_3_FP3	70	0	0	0.0001

Supplementary Table 3. Highlights *P. falciparum* associated malaria burden and whole genome sequencing sample availability across the 28 countries of interest. Metrics derived from the Malaria Atlas Program historic 2015-2019 data[8]

Country	WGS Samples	Burden (%)
Nigeria	29	37.08
Democratic Republic of the Congo	301	16.07
Uganda	10	6.59
Cote d'Ivoire	65	5.12
Cameroon	207	4.25
Ghana	783	4.25
Tanzania	285	4.12
Burkina Faso	37	3.75
Mali	352	2.87
Benin	71	2.84
Kenya	97	2.81
Ethiopia	20	2.51
Malawi	205	2.30
Guinea	135	2.29
Madagascar	22	1.38
Senegal	161	0.67
Papua New Guinea	107	0.27
Mauritania	72	0.27
Gabon	53	0.26
Gambia	230	0.16
Myanmar	195	0.07

Colombia	16	0.02
Cambodia	959	0.02
Peru	23	0.01
Bangladesh	69	0.01
Laos	115	0.01
Vietnam	225	0.01
Thailand	824	8.99E-04

Supplementary Table 4. Highlights *P. vivax* associated Malaria Burden and WGS sample availability across the 21 countries of interest. Metrics derived from the Malaria Atlas Program historic 2015-2019 data[8]

Country	**WGS Samples**	**Burden (%)**
India	13	64.80
Ethiopia	60	18.93
Papua New Guinea	58	5.12
Indonesia	12	4.17
Brazil	91	2.96
Madagascar	4	1.13
Peru	164	0.77
Myanmar	17	0.74
Colombia	85	0.48
Cambodia	130	0.35

Laos	2	0.17
Nicaragua	3	0.14
Vietnam	26	0.08
Bangladesh	28	0.05
Malaysia	93	0.04
Thailand	171	0.02
Panama	3	0.02
Mexico	40	0.01
Bhutan	9	6.51E-04
China	12	1.52E-04
Sri Lanka	1	0

References

1. 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).
2. Benavente, E. D. *et al.* Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. *Nat. Commun.* **12**, 3160 (2021).
3. MalariaGEN *et al.* An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. *Wellcome Open Res* **6**, 42 (2021).

4. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
5. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
6. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
7. SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbour thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1460–1465 (1998).
8. Pfeffer, D. A. *et al.* malariaAtlas: an R interface to global malariometric data hosted by the Malaria Atlas Project. *Malar. J.* **17**, 352 (2018).

Chapter 6. Genomic variation of *Plasmodium ovale spp* in Sub Saharan Africa and the creation of two new reference genomes for *P. ovale curtisi* and *P. ovale walkeri.*

RESEARCH PAPER COVER SHEET

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

SECTION A – Student Details

Student ID Number	1702842	**Title**	Mr
First Name(s)	Matthew		
Surname/Family Name	Higgins		
Thesis Title	Developing an RPA-based Molecular Barcoding Tool for Plasmodium Malaria		
Primary Supervisor	Prof. Taane Clark		

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?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	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?	Scientific Reports

Please list the paper's authors in the intended authorship order:	Matthew Higgins, Daniel Ward, Debbie Nolder, Colin Sutherland, Taane G. Clark, Susana Campino
Stage of publication	**Submitted**

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 designed the SWGA primer design set and completed all bioinformatic analysis highlighted. I wrote the first draft of the manuscript which was then circulated to supervisors and co-authors.

SECTION E

Student Signature	[redacted]
Date	26/06/2022

Supervisor Signature	[redacted]
Date	26/6/2022

Title: Genomic variation of *Plasmodium ovale spp* in Sub Saharan Africa and the creation of two new reference genomes for *P. ovale curtisi* and *P. ovale walkeri.*

Authors: Matthew Higgins [1], Daniel Ward [1], Debbie Nolder [1], Colin Sutherland [1], Taane G. Clark [1,2,*], Susana Campino [1,*,]

1. Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK
2. Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* Joint corresponding authors

Prof. Susana Campino and Prof. Taane Clark, Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Abstract

Genomic characterisation of the neglected malaria parasites, *P. ovale curtisi (Poc)* and *P. ovale walkeri (Pow)* remains limited. With the incidence of *P. ovale spp* mixed infections on the rise here we seek to address this issue. We present a Selective Whole Genome Sequencing (SWGA) primer set for *P. ovale spp* specific enrichment, demonstrating its use in generating high quality long and short read whole genome sequencing data for 11 *Poc* and 8 *Pow* samples collected from the African continent. For both *Poc* and *Pow* we present new reference genomes, improving on existing assemblies. A total of 449,399 and 371,291 SNPs were identified, for *Poc* and *Pow* respectively, of which 68,608 (15.3%) and 71,477 (19.3%) were unique. Further evidence is present to support the dimorphic separation hypothesis confirming *Poc* and *Pow* are indeed two separate sympatric species. Antimalarial resistance orthologs were identified in both species, including ones associated with pyrimethamine resistance, located within the dihydrofolate reductase (*DHFR*) gene.

Introduction

Plasmodium ovale curtisi (Poc) and *Plasmodium ovale walkeri (Pow)* are the least studied human infecting *Plasmodium* parasites. Large gaps remain in our understanding of these elusive parasites, from their full geographic distribution to antimalarial susceptibility. *P. ovale spp* was first discovered in 1922 as the causative agent of tertian malaria[1]. In 2010, it was demonstrated that two non-recombining sympatric species of *P. ovale spp* exist, *Poc* and *Pow* [2]. Historically, *P. ovale spp* has been associated with a benign form of malaria but severe disease can arise, including jaundice, anaemia and fatal pulmonary impairments [3]. It has been suggested that *P. ovale spp* can cause a relapse in infection similar to *P. vivax*, however the presence of hypnozoites representing the dormant stage of the parasite have not

been clearly distinguished. Treatment of *P. ovale spp* typically follows the same path as *P. vivax,* however evidence for the efficacy of antimalarials against *P. ovale spp* is lacking [4].

Current estimates predict the majority of *P. ovale spp* infections occur in Africa (94.5%) followed by Asia (5.3%) [3]. A cluster of *P. ovale spp* infections were detected in South America, however these are believed to be the result of imported cases instead of local transmission [5]. *P. ovale spp* infections are known to have a low parasitemia resulting in most individuals remaining asymptomatic. As such these cases have historically passed under the radar, leading to an under-estimation of the prevalence of *P. ovale spp*. This issue is compounded by the fact that *P. ovale spp* has been systematically mischaracterized as *P. vivax* when diagnosed by microscopy, and commonly used pan-*Plasmodium* immunochromatic rapid diagnostic test have poor sensitivity (22.2%) for *P. ovale spp* infections [6,7]. Cases of *P. ovale spp* are commonly detected as co-infections with another *Plasmodium* parasite (e.g., *P. falciparum*) causing the infected individual to become symptomatic and in turn reach out for medical support [8]. In countries such as Kenya the prevalence of *P. ovale spp* co-infections is increasing, the cause of which is unknown [9].

To date, the genomic characterisation of *P. ovale curtisi* and *P. ovale walkeri* remains limited, with incomplete reference genomes, and no studies of genome diversity. Overcoming this is important for both *P. ovale spp* treatment development and surveillance efforts. Sequence data is crucial to assist the assessment of vaccine targets and drug resistance orthologs, and to improve diagnostic design to ensure parasite inclusivity. To aid this, we designed, tested and implemented a *P. ovale spp* specific selective whole genome amplification (SWGA) primer set, for parasite specific genomic enrichment, to generate whole genome sequencing data for 32 isolates from different regions in Africa. In addition,

long-read sequences were used to generate much-needed *Poc* and *Pow* reference genomes and provide a first assessment of *P. ovale spp* population diversity across the African continent.

Methods

P. ovale sample collection

This project incorporated 32 *P. ovale spp* samples, obtained from 13 countries in the African continent, including Congo, South Sudan, Nigeria, Ghana, Sierra Leone, Guinea, Cameroon, Kenya, Sudan, DRC, Tanzania, Uganda, Mozambique (**Supplementary Table 1).** The DNA samples were extracted from blood samples from returning travellers to the UK, who were diagnosed with malaria between 2019 and 2020, confirmed by the UK Health Security Agency-Malaria Reference Laboratory at the London School of Hygiene and Tropical Medicine (LSHTM). Samples were initially designated as *P. ovale spp* infections by nested PCR and qPCR according to standard practice [10,11]. The UK National Research Ethics Service (Ref: 18/LO/0738) and LSHTM Research Ethics Committee (Ref: 14710) provided approval for the project "Drug susceptibility and genetic diversity of imported malaria parasites from UK travellers". Informed consent was obtained from all UK traveller study participants.

P. ovale spp. Selective Whole Genome Amplification (SWGA)

Similar to previous *Plasmodium* genomic studies, we utilised SWGA to optimise our genomic investigation [12]. Candidate primer sets were first identified using the SWGA tool (https://github.com/eclarke/swga) [13], and designed to preferentially amplify *P. ovale curtisi* (PocGH01) and *P. ovale walkeri* (PowCR01) (https://plasmodb.org) over the human genome (GRCh38) to facilitate parasite enrichment [14]. The top primer sets identified were

subsequently extracted and overlapping primers combined to form a final set of 7 primers: CGAAAAA*A*C, CGAAAT*T*G, TCGTAAA*A*A, CGTAAT*A*A, TTTACGT*A*T, ATTTTCG*A*T, and TATCGT*T*A, where an asterisk (*) represents the presence of a phosphorothioate bond which minimises primer degradation by the 3' exonuclease activity of Phi29. *In-silico* assessment of the candidate primers demonstrate a minimum 11-fold preferential enrichment for both *Poc* and *Pow* compared to the human genome (**Supplementary Table 2).**

Samples were subject to SWGA following previously published protocols [12]. All SWGA reactions were carried out in a UV Cabinet for PCR Operations (UV-B-AR, Grant-Bio) to eliminate potential contamination. A maximum of 80 ng of gDNA (minimum of 5 ng) was added to a total 50 µl reaction alongside 5 µl of 10 × Phi29 DNA Polymerase Reaction Buffer (New England BioLabs), 0.5 µl of Purified 100 × BSA (New England BioLabs), 0.5 µl of 250 µM Primer mix, 5 µl 10 mM dNTP (Roche), 30 units Phi29 DNA Polymerase (New England BioLabs) and Nuclease-Free Water (Ambion, The RNA Company) to reach a final reaction volume of 50 µl. The reaction was carried out on a thermocycler with the following step-down program: 5 min at 35 °C, 10 min at 34 °C, 15 min at 33 °C, 20 min at 32 °C, 25 min 31 °C, 16 h at 30 °C and 10 min at 65 °C. After SWGA, samples were purified using a 1:1 ratio of AMPure XP beads (Beckman-Coulter), following manufacturer's instructions and quantified via Qubit assay.

Library Preparation and whole genome sequencing

Short read sequencing for all 32 *P. ovale spp* samples was performed using Illumina technology with paired end 150bp reads, made available through The Applied Genomics Centre, LSHTM. For long-read sequencing, we multiplexed two *P. ovale spp* samples

(Poc_SSD_001, Pow_NGA_001) using the Oxford Nanopore Technology's LSK-109 and EXP-NBD104 barcoding kit as per manufacturer's instruction. Due to the hyperbranched structure of WGA products, we firstly treated the SWGA samples with T7 endonuclease (NEB-M0302S), as per manufacturer's protocol (WAL_9070_v109_ revQ_14Aug2019). To select for fragments of greater mass during the library preparation procedure, we used LSB buffer (ONT) during magnetic bead clean-up. We loaded 120 ng on the R10 flow cell. The total yield from three wash-run cycles (EXP-WSH004) was 2.5 GB (N50 5KB). The greatest sample yield was 1.37 Gb, which resulted in a 26% on target sequencing yield of 356 Mb (74% Human background). Resulting fast5 files were base called using Bonito (ONT) (model: dna_r10.3@v3.3). Reads were then trimmed and demultiplexed using Porechop v0.2.4.

Sequence data quality control

Trimmomatic software was applied under default settings to all short-read data [15]. Human contamination was subsequently removed using bwa-mem and samtools software through mapping to the human reference (GRCh38) [16,17]. Similarly, human contamination was removed from long-read data using minimap2 and samtools software [18]. For the two samples used for reference creation (Poc_SSD_001, Pow_NGA_001), 5.5% and 5.4% of short reads were filtered out via trimming and 75.9% and 80.8% were subsequently removed by mapping to the human genome, leaving >9 and >6 million read pairs, respectively. Processed short-read data for each sample was subsequently mapped to the PocGH01 mitochondria genome. Under the dimorphic separation hypothesis, haplotypes in the mitochondria *cytB* gene can be used for *P.ovale spp* species classification, *Poc* and *Pow* [2]. Point mutation differences within the *cytB* gene were found (**Supplementary Table 3**). Utilising the mitochondria is desirable due to the traditionally high coverage of circular genomes obtained through SWGA, allowing

accurate variant calling [12]. If a sample contained >90% of species-specific sites they were classified accordingly.

Hybrid Genome Assembly

Hybrid Spades software [19] was implemented under default settings to perform the initial assembly utilising *Plasmodium* specific minion and paired-end Illumina reads. Contigs derived by Spades software were subsequently scaffolded via RagTag using the existing *PocGH01* reference genome as a guide [19,20]. GapFiller, Pilon and Abyss-Sealer tools were all subsequently applied to polish and close gaps in the assembly all of which utilised short-read data available [21–23]. Coverage-based missembly corrections were made using a custom python script and the final assemblies were assessed using Busco software (plasmodium_odb10 lineage database) [24]. Assemblies were annotated via Companion software whilst the OrthoMCL tool was used under default settings to identify gene orthologs [25,26]. Amino acid sequences from the following 15 Plasmodium references were included in the ortholog analysis; *P. berghei ANKA*, *P. chabaudi chabaudi*, *P. cynomolgi M*, *P. falciparum 3D7, P. gallinaceum 8A*, *P. knowlesi H*, *P. malariae UG01, Poc GH01*, *P. reichenowi CDC*, *P. vinckei brucechwatti DA*, *P. vinckei lentum DE*, *P. vinckei vinckei CY, P. vivax P01*, *P. yoelii yoelii 17X*, *Pow CR01*. Only complete protein sequences were included in the ortholog analysis. Expanded gene family analysis was performed through the identification of OrthoMCL clusters with known *PIR* and *Surfin* genes. Subsequently all genes in the clusters identified were carried forward for the species-specific gene family final count. Chromosomal structural variant analysis between *Poc* and *Pow* was completed using progressive Mauve 2.4.0 software [27]. The program was run with default "seed families" and default values for all other parameters. Candidate centromeric DNA regions were identified using a custom python script.

Phylogenetic and population genetic analysis

A total of 3,239 clusters of 1:1 orthologs were identified across all 15 *Plasmodium* references and formed the foundation of our phylogenetic analysis. The proteins belonging to each cluster were extracted and aligned using Mafft software under default settings [28]. Each alignment was then processed using the GBlocks software under default settings to remove gapped and uninformative positions [29]. All alignments were subsequently combined to form a sequence covering 1,456,931 amino acids. This combined sequence was used to construct multiple maximum likelihood phylogenetic trees via bootstrapping with RAXML-ng software utilising the following substitution models, LG, LG4X, LG4M, JTT, JTT-DCMut, PROTGTR all applied with gamma distribution [30]. The conserved tree structure identified was subsequently visualised and plotted via ITOL [31]. Filtered short-read data was mapped to the appropriate *P. ovale spp* assembly, depending on *cytB* classification, using bwa-mem software under default settings. Samtools was subsequently used to extract coverage statistics for each sample. Poor quality samples were removed (<40% of the genome with ≥5-fold coverage) leaving 19 high quality samples. SNPs were identified using GATK and filtered according to previously defined methods [12,32]. Gene specific consensus sequences for *Poc* and *Pow* samples that passed quality control steps were derived using bcftools software for each of the 3239 1:1 ortholog clusters. Mafft software was subsequently used to align the sequence under default settings and a custom python script used to extract dimorphic sites in each alignment. These dimorphic sites were subsequently used to investigate the population structure of the *Poc* and *Pow* isolates through the creation of a distance matrix based on the pairwise identity of each sample given the allele present. The 'ape' R package was used to create a neighbour joining tree which was subsequently visualised in iTOL [33].

Results

New *P. ovale spp* reference genomes

The new reference genomes, for *Poc (*Poc_221) and *Pow (*Pow_222), were assembled following a hybrid approach combining both Illumina and Minion next generation sequencing data. The new assemblies were subsequently benchmarked against the existing reference genomes, PocGH01 and PowCR01 [14], respectively (**Table 1).** Gains in core genome contiguity were made for both species (~3.3 Mbp), reflected by 3.9- and 4.4-fold improvements in N50 and a 68% and 73% reduction in the number of gaps for *Poc* and *Pow,* respectively. A BUSCO analysis revealed that the new assemblies outperformed the existing assemblies, improving the number of complete single copy core orthologs by 17 and 13, shared across all *Plasmodium*, whilst decreasing the number of missing orthologs by 2 and 22 for *Poc* and *Pow,* respectively (**Supplementary Table 4)**.

Expanding the comparative analysis into core genome contiguity, the new and existing references were compared on a chromosomal level (**Figure 1).** For *Poc,* gains in chromosome length were made across all nuclear chromosomes, with a maximum increase of 696 kbp in chromosome 7 of and minimum increase of 0.698 kbp in chromosome 10. Whilst overall gains were made for *Pow* nuclear genome contiguity (~3.3 Mbp), some chromosomes experienced a decrease in length (e.g., chr. 2, 5 and 6). However, a large increase in length for chromosome 10 was observed of ~1.3 Mbp. Mitochondrial and apicoplast organelle genomes were successfully obtained for Poc_221 (5974 bp and 33,482bp) and Pow_222 (5975 bp and 34310 bp), respectively. Both mitochondrial genomes were successfully circularised, improving on the historic PowCR01 reference. In addition, gains in contiguity were made for both apicoplast of +5 kb and +3 kb for *Poc* and *Pow*, respectively.

The annotation features for both Poc_221 and Pow_222 were assessed (**Table 2).** A total of 6,821 and 6,564 genes were identified for Poc_221 and Pow_222 respectively and complete genes were clustered against 15 other *Plasmodium* references via the OrthoMCL tool. A total of 6,140 clusters were identified which contained at least one Poc_221 or Pow_222 feature (**Supplementary Table 5)**. Of these, a total of 4487 and 4505 features for Poc_221 and Pow_222 respectively, were clustered with at least one other feature from another *Plasmodium* species. A total of 3,239 clusters were identified containing 1:1 orthologs found across all 15 *Plasmodium* reference genomes included in this analysis. Previous work has highlighted that expanded gene families are one of the driving factors behind expansion in the *P. ovale spp* genome size [34]. Based on OrthoMCL cluster analysis, utilising known PIR and STP1 gene family members in non-ovale *Plasmodium* species, 1,205 and 932 PIR and 56 and 68 STP1 genes were identified in the new assemblies for *Poc* (Poc_221) and *Pow* (Pow_222), respectively. These findings are in line with previous reports [34]. When comparing the Poc_221 and Pow_222 assemblies, 150 1:1 orthologs between *Poc* and *Pow* were identified to be located on different chromosomes, indicating chromosomal rearrangement (**Supplementary Table 6)**.

When comparing chromosomal structural variants between Poc_221 and Pow_222, inversions greater than 1 kbp were identified in all chromosomes except chromosome 2 (**Supplementary Table 7)**. For each assembly, candidate centromeric DNA regions were identified for each nuclear chromosome, based on AT enrichment and absence of coding-DNA, in line with previous reports (**Supplementary Table 8).** Sole candidate regions were identified in 13 and 12 *Poc* and *Pow* chromosomes, respectively. For *Pow* chromosome 14, two candidate regions were identified, one of which aligned with the sole candidate centromeric region in *Poc* chromosome 14. However, when comparing these regions, it

appears that the aligned *Pow* region is located inside an inversion, making the inversion pericentric (**Figure 2)**. Pericentric inversions are hypothesised to contribute to speciation and as such its presence may have contributed to the rise of an additional centromeric DNA candidate region in *Pow* chromosome 14.

Utilising the 1:1 ortholog clusters, maximum likelihood trees were constructed for the 15 selected *Plasmodium* reference genomes included in this analysis. The conserved tree topology identified was in line with previous reports (**Figure 3)** [14]. As expected, both *Poc* and *Pow* cluster together as highlighted in blue. As previously reported, the *P.ovale spp*, clade share a most recent common ancestor with rodent infecting *Plasmodium* species, including *P. berghei ANKA, P. chabaudi chabaudi, P. vinckei brucechwatti, P. vinckei lentum, P. vinckei vinckei,* and *P. yoelii yoelii.* [14]

SWGA Enrichment of *P. ovale spp*.

A total of 32 samples underwent SWGA, sequencing and mapping. Subsequently, 19 samples carried forward for analysis averaged 83% and 84% genome coverage ≥5-fold for *Poc* and *Pow* respectively, facilitating reliable variant calling. Samples which were unsuccessfully enriched we suspect were due to low starting parasitaemias, which has been previously correlated to enrichment success [12]. However, our findings demonstrate that the primer set designed in this investigation can be successfully used for the enrichment of both *P. ovale spp* species. For the 11 *Poc* and 8 *Pow* isolates carried forward a total of 449,399 and 371,291 SNPs were identified, respectively, of which 68,608 (15.3%) and 71,477 (19.3%) were unique. Subsequently the nucleotide diversity of the *Poc* and *Pow* isolates were determined to be 2.98×10^{-3} and 2.70×10^{-3}, respectively.

Confirmation that *Poc* and *Pow* are separate species

It was hypothesised that *P. ovale spp* as two separate non-recombining sympatric species based on the abundance of dimorphic sites across several conserved genes. We expanded this approach to investigate the 3239 1:1 protein clusters identified through ortholog analysis [2]. After aligning the *Poc* and *Pow* conserved genes in each cluster, biallelic loci were identified and extracted from the 11 *Poc* and 8 *Pow* isolates respectively. A total of 283,794 biallelic loci were identified. Population differentiation fixation index (F_{ST}) values were calculated for each loci separating samples based on the assigned *P.ovale spp* species. A total of 96.7% (274,441) loci had an F_{ST} of 1, representing complete dimorphic separation of both species. A total of 9,320 loci were found to have mixed calls occurring in only one of the *P.ovale spp* species, representing 3.3% of all loci identified. Only 33 loci (0.01%) were found to have major and minor alleles prevalent in both *Poc* and *Pow* populations. A neighbour-joining tree was constructed using the loci identified, and a clear separation between each *P. ovale spp* species is observed (**Figure 4).** This provides further evidence that *Poc* and *Pow* are separate sympatric species as samples from the same country of origin, such as Nigeria, cluster first according to species and then geographical origin.

P. ovale spp. gamete linked gene assessment

The absence of recombination between the sympatric *P.ovale spp* species, would be one driving factor behind dimorphic separation. As such we investigated the conservation of three key proteins, associated with *Plasmodium* gamete recognition and fusion, P47, P230 and P48/45 [35]. Gamete recognition is fundamental to the sexual stages of the *Plasmodium* life cycle and essential for recombination [35]. The divergence of each protein sequence between each *P. ovale spp.* species is shown (**Table 3)**. Of the three proteins, P47 appeared the most divergent (15.4%), whilst P48/45 was the most conserved (94.6%). With the isolates

available, no divergence of the P48/45 protein orthologs was found in either *P. ovale spp* species. The mean sequence divergence for the other two proteins remained below 0.1% across both *P. ovale spp* species. Whilst the full role of P47 protein ortholog is disputed across *Plasmodium* species, knock-out studies in *P. bergi* have demonstrated that the protein is essential for female gamete fertility under *in-vitro* conditions and significantly impaired fertilisation *in vivo* [36]. When benchmarked against the divergence of other *Plasmodium* species comparisons, *P. ovale spp* P47 protein divergence was in line with *P. yoelii* versus *P. berghei* and greater than *P. chabaudi* versus *P. vinckei vinkei* whose reference genome orthologs had sequence identities of 85.2% and 89.8%, respectively. In comparison the P230 and P48/45 *P. ovale spp* orthologs were more conserved compared to *P. yoelii vs P. berghei* and *P. chabaudi* vs *P.vinckei vincke*i orthologs which had an identity of 89.4%, 93.4% and 89.7%, 92.3% respectively. Of the three proteins investigated, P47 appears to be the most likely candidate for our hypothesis of gamete incompatibility, but further work is required to investigate this *in-vitro*.

<u>*P. ovale spp.* resistance orthologs</u>

The presence of antimalarial resistance is well characterised amongst other human infecting *Plasmodium* species, including *P. falciparum* and *P. vivax*. However, clinical investigations of the efficacy of existing antimalarials against *P.ovale spp* infections are rare. We sought to identify the presence of known antimalarial resistance orthologs in *P. ovale spp* whose mechanisms have been well characterised in other *Plasmodium. P. ovale spp* orthologs to 14 genes were screened *in-silico* (**Table 4),** and antimalarial resistance markers were identified in four (*PF3D7_0417200 (dhfr), PF3D7_0112200 (mrp1), PF3D7_1447900 (mdr2) and PVP01_1010900 (mdr1)*). No antimalarial resistance markers associated with artemisinin were identified, including orthologs to those found in *PfKelch13*. Several known markers

associated with pyrimethamine resistance were identified including in the known drug target, dihydrofolate reductase (DHFR) [42]. Orthologs to the C59R (3/11) and S108N (1/11) mutations were observed in the *Poc* population. The C59R (3/8) mutation was also identified in *Pow* samples in addition to the I164L (1/11) mutation. In both *P. ovale* species, other mutations of interest were identified within the *dhfr* gene, whose orthologs have not been phenotypically characterised in other *Plasmodium*, **(Supplementary Figure 1)**, including I164T which was identified in the *Poc* population. The F423Y mutation in the *Pfmdr2* gene, previously linked *in-vitro* to pyrimethamine resistance, appeared to be fixed in both *Poc* (11/11) and *Pow* populations (8/8) [48].

For the *P.ovale spp* orthologs to *Pvmdr1*, the F1076L resistance conferring mutation was found to be fixed in both *P. ovale spp* species. The F1076L mutation has previously been associated as a potential marker for chloroquine resistance and warrants further investigation [50]. However, other chloroquine-resistance orthologs were absent in all *P. ovale spp* samples, including the well documented K76 mutation in orthologs to the PfCRT protein. Finally, orthologs to the PfMRP1 protein were found to harbour the I876V mutation in both the *Poc* (6/11) and *Pow* populations (8/8) which has been previously associated with pyronaridine resistance [51].

Discussion

Poc and *Pow* are both neglected malaria parasites whose incidence is rising. Genetic characterisation of *P. ovale spp* may provide clues as to what is driving this increase, from the identification and surveillance of antimalarial resistance markers, to assessing the conservation of vaccine candidate and diagnostic associated genes. Here we present an SWGA primer set to aid in *P. ovale spp* whole genome sequencing efforts and confirm it can

be successfully applied via next-generation short and long-read sequencing platforms. Utilising a hybrid assembly approach, we successfully constructed two high quality *P. ovale spp* reference genomes, for *Poc* and *Pow,* with improvements in contiguity and completeness compared to previous work. Phylogenetic analysis positioned the *P. ovale spp* clade in line with previous reports, when comparing the new *P. ovale spp* reference genomes against other human, rodent and avain infecting *Plasmodium* species. Further investigation is needed to determine if the pericentric inversion identified in *Pow* (Chromosome 14)*,* is fixed within the population, where such a large structural variant could have contributed to *P. ovale spp* speciation. Our analysis provides further evidence to support the dimorphic separation hypothesis, indicating that *Pow* and *Poc* are two separate non-recombining species [2]. Investigation of the key genes associated with *Plasmodium* gamete recognition and fertilisation, including P47, demonstrated divergence between *Poc* and *Pow* to be consistent and exceed other closely related *Plasmodium* species. To the best of our knowledge, we are the first to report the presence of antimalarial resistance markers in both *Poc* and *Pow* populations linked to both pyrimethamine and pyronaridine resistance. The identification of such resistance haplotypes provides further evidence that a *P. ovale spp* specific assessment of antimalarial treatments is warranted, especially as this could be a contributing factor to the rise in *P. ovale spp* infections. It is crucial for malaria elimination that *P. ovale spp* does not exploit any niches left behind from the successful treatment of other *Plasmodium* species.

Acknowledgements

M.H is a recipient of a BBSRC LiDO PhD studentship. S.C is funded by Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and MR/R020973/1) and Bloomsbury SET. T.G.C is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no.

BB/R013063/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

S.C and T.G.C conceived and directed the project. D.N and C.J.S organised isolate collection and processing. M.H and D.W undertook laboratory work including sequencing. M.H performed bioinformatic analysis under the supervision of S.C and T.G.C, and together they interpreted the results. Additional advice from C.J.S was sought during analysis. M.H wrote the first draft of the manuscript with guidance from S.C and T.G.C. All authors commented on versions of the manuscript and approved the final manuscript. M.H, C.J.S, S.C and T.G.C compiled the final manuscript.

Competing interest statement

The authors have declared that no competing interests exist.

References

1. Stephens, J. W. W. A New Malaria Parasite of Man. *Ann. Trop. Med. Parasitol.* **16**, 383–388 (1922).
2. Sutherland, C. J. *et al.* Two nonrecombining sympatric forms of the human malaria parasite Plasmodium ovale occur globally. *J. Infect. Dis.* **201**, 1544–1550 (2010).
3. Mahittikorn, A., Masangkay, F. R., Kotepui, K. U., Milanez, G. D. J. & Kotepui, M. Comparison of Plasmodium ovale curtisi and Plasmodium ovale wallikeri infections by a meta-analysis approach. *Sci. Rep.* **11**, 6409 (2021).
4. Groger, M., Fischer, H. S., Veletzky, L., Lalremruata, A. & Ramharter, M. A systematic review of the clinical presentation, treatment and relapse characteristics of human

Plasmodium ovale malaria. *Malar. J.* **16**, 112 (2017).

5. Guerra, R. I. *et al.* A cluster of the first reported Plasmodium ovale spp. infections in Peru occuring among returning UN peace-keepers, a review of epidemiology, prevention and diagnostic challenges in nonendemic regions. *Malaria Journal* vol. 18 (2019).
6. Tanizaki, R. *et al.* Performance of Rapid Diagnostic Tests for Plasmodium ovale Malaria in Japanese Travellers. *Trop. Med. Health* **42**, 149–153 (2014).
7. Chavatte, J.-M., Tan, S. B. H., Snounou, G. & Lin, R. T. P. V. Molecular characterization of misidentified Plasmodium ovale imported cases in Singapore. *Malar. J.* **14**, 454 (2015).
8. Mitchell, C. L. *et al.* Under the Radar: Epidemiology of Plasmodium ovale in the Democratic Republic of the Congo. *J. Infect. Dis.* **223**, 1005–1014 (2021).
9. Akala, H. M. *et al.* Plasmodium interspecies interactions during a period of increasing prevalence of Plasmodium ovale in symptomatic individuals seeking treatment: an observational study. *The Lancet Microbe* **2**, e141–e150 (2021).
10. Mangold, K. A. *et al.* Real-time PCR for detection and identification of Plasmodium spp. *J. Clin. Microbiol.* **43**, 2435–2440 (2005).
11. Isozumi, R. *et al.* Improved detection of malaria cases in island settings of Vanuatu and Kenya by PCR that targets the Plasmodium mitochondrial cytochrome c oxidase III (cox3) gene. *Parasitol. Int.* **64**, 304–308 (2015).
12. Ibrahim, A. *et al.* Selective whole genome amplification of Plasmodium malariae DNA from clinical samples reveals insights into population structure. *Sci. Rep.* **10**, 10832 (2020).
13. Clarke, E. L. *et al.* swga: a primer design toolkit for selective whole genome amplification. *Bioinformatics* **33**, 2071–2077 (2017).
14. Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into

malaria parasite evolution. *Nature* **542**, 101–104 (2017).

15. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
16. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
18. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
19. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
20. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
21. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13 Suppl 14**, S8 (2012).
22. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
23. Paulino, D. *et al.* Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* **16**, 230 (2015).
24. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
25. Steinbiss, S. *et al.* Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* **44**, W29–34 (2016).
26. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for

eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

27. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
28. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
29. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
30. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
31. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
32. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
33. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
34. Ansari, H. R. *et al.* Genome-scale comparison of expanded gene families in Plasmodium ovale wallikeri and Plasmodium ovale curtisi with Plasmodium malariae and with other Plasmodium species. *Int. J. Parasitol.* **46**, 685–696 (2016).
35. van Dijk, M. R. *et al.* Three members of the 6-cys protein family of Plasmodium play a role in gamete fertility. *PLoS Pathog.* **6**, e1000853 (2010).
36. Molina-Cruz, A., Canepa, G. E. & Barillas-Mury, C. Plasmodium P47: a key gene for malaria transmission by mosquito vectors. *Curr. Opin. Microbiol.* **40**, 168–174 (2017).
37. Henrici, R. C., van Schalkwyk, D. A. & Sutherland, C. J. Modification of pfap2µ and pfubp1 Markedly Reduces Ring-Stage Susceptibility of Plasmodium falciparum to

Artemisinin In Vitro. *Antimicrob. Agents Chemother.* **64**, (2019).

38. Miotto, O. *et al.* Genetic architecture of artemisinin-resistant Plasmodium falciparum. *Nat. Genet.* **47**, 226–234 (2015).
39. Breglio, K. F. *et al.* A single nucleotide polymorphism in the Plasmodium falciparum atg18 gene associates with artemisinin resistance and confers enhanced parasite survival under nutrient deprivation. *Malar. J.* **17**, 391 (2018).
40. Demas, A. R. *et al.* Mutations in Plasmodium falciparum actin-binding protein coronin confer reduced artemisinin susceptibility. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12799–12804 (2018).
41. Foguim, F. T. *et al.* Prevalence of mutations in the Plasmodium falciparum chloroquine resistance transporter, PfCRT, and association with ex vivo susceptibility to common anti-malarial drugs against African Plasmodium falciparum isolates. *Malar. J.* **19**, 201 (2020).
42. Lynch, C. *et al.* Emergence of a dhfr mutation conferring high-level drug resistance in Plasmodium falciparum populations from southwest Uganda. *J. Infect. Dis.* **197**, 1598–1604 (2008).
43. Torrevillas, B. K. *et al.* Plasmodium falciparum DHFR and DHPS Mutations Are Associated With HIV-1 Co-Infection and a Novel DHPS Mutation I504T Is Identified in Western Kenya. *Front. Cell. Infect. Microbiol.* **10**, 600112 (2020).
44. Balikagala, B. *et al.* Absence of in vivo selection for K13 mutations after artemether-lumefantrine treatment in Uganda. *Malar. J.* **16**, 23 (2017).
45. Straimer, J. *et al.* Drug resistance. K13-propeller mutations confer artemisinin resistance in Plasmodium falciparum clinical isolates. *Science* **347**, 428–431 (2015).
46. Li, J. *et al.* High prevalence of pfmdr1 N86Y and Y184F mutations in Plasmodium falciparum isolates from Bioko Island, Equatorial Guinea. *Pathog. Glob. Health* **108**,

339–343 (2014).

47. Zeng, W. *et al.* Significant Divergence in Sensitivity to Antimalarial Drugs between Neighboring Plasmodium falciparum Populations along the Eastern Border of Myanmar. *Antimicrob. Agents Chemother.* **61**, (2017).
48. Briolant, S. *et al.* The F423Y mutation in the pfmdr2 gene and mutations N51I, C59R, and S108N in the pfdhfr gene are independently associated with pyrimethamine resistance in Plasmodium falciparum isolates. *Antimicrob. Agents Chemother.* **56**, 2750–2752 (2012).
49. Ngassa Mbenda, H. G. *et al.* Evolution of the Plasmodium vivax multidrug resistance 1 gene in the Greater Mekong Subregion during malaria elimination. *Parasit. Vectors* **13**, 67 (2020).
50. Brega, S. *et al.* Identification of the Plasmodium vivax mdr-like gene (pvmdr1) and analysis of single-nucleotide polymorphisms among isolates from different areas of endemicity. *J. Infect. Dis.* **191**, 272–277 (2005).
51. Gupta, B. *et al.* Plasmodium falciparum multidrug resistance protein 1 (pfmrp1) gene and its association with in vitro drug susceptibility of parasite isolates from north-east Myanmar. *J. Antimicrob. Chemother.* **69**, 2110–2117 (2014).

Table 1. Comparison of assembly metrics for existing and new *P. ovale curtisi (Poc)* and *P. ovale walkeri (Pow)* reference genomes.

	Assembly Size	Core Genome Size	N50	Number of Gaps	N Per 100kb	BUSCO Score
PocGH01	33,485,483	20,744,212	38,586	883	271.3	96.1
Poc_221	34,734,807	24,049,073	151,239	277	87.6	96.6
PowCR01	33,529,622	20,953,308	30,524	1206	777.8	95.5
Pow_222	33,017,020	24,319,118	134,314	325	113.5	95.9

Table 2. Annotation summary for the new reference genomes (Poc_221 and Pow_222).

	Poc_221	Pow_222
Scaffolds*	16 (832)	16 (627)
Number of genes	6,821	6,564
Gene density (genes/megabase)	194.13	196.33
Number of coding genes	6737	6475
Number of pseudogenes	310	286
Number of non-coding genes	84	89
Overall GC%	28.53	29.53

* Unassigned contigs indicated in parentheses.

Table 3. Investigating protein divergence for key proteins associated with *Plasmodium* gamete recognition and fusion.

	Sequence Divergence (%)		
Protein	Poc vs. Pow	Poc Population	Pow Population
P47	15.4	0.065	0.089
P230	7.9	0.042	0.035
P48/45	3.6	0	0

Poc = *P. ovale curtisi;* Pow = *P. ovale walkeri*

Table 4. Fourteen antimalarial resistance genes of interest and corresponding non-synonymous mutations.

Gene of Interest	PlasmoDB ID	Mutations
PfAP2-MU	PF3D7_1218300	I592T [37]
PfARPS10	PF3D7_1460900	V127M [38]
PfATG18	PF3D7_1012900	T381[39]
PfCoronin	PF3D7_1251200	G50E:R100K:E107V [40]
PfCRT	PF3D7_0709000	K76T:K76A:K76N:K76I:T93S:H97Y:C101F:F145I:M343L:C350R:G353V [41]
PfDHFR-TS	PF3D7_0417200	N51I:C59R:S108N:I164L [42]
PfDHPS	PF3D7_0810800	A437G:K540E:I504T:A581G [43]
PfFD	PF3D7_1318100	D193Y [44]
PfKelch13	PF3D7_1343700	C580Y:Y493H:R539T [45]
PfMDR1	PF3D7_0523000	N86Y:Y184F:S1034C:N1042D:D1246Y [46]
PfUBP1	PF3D7_0104300	V3275F [37]
PfMRP1	PF3D7_0112200	H191Y:N325S:S437A:H785N:I876V:T1007M:F1390I [47]
PfMDR2	PF3D7_1447900	F423Y [48]
PvMDR1	PVP01_1010900	Y976F:F1076L [49]

.

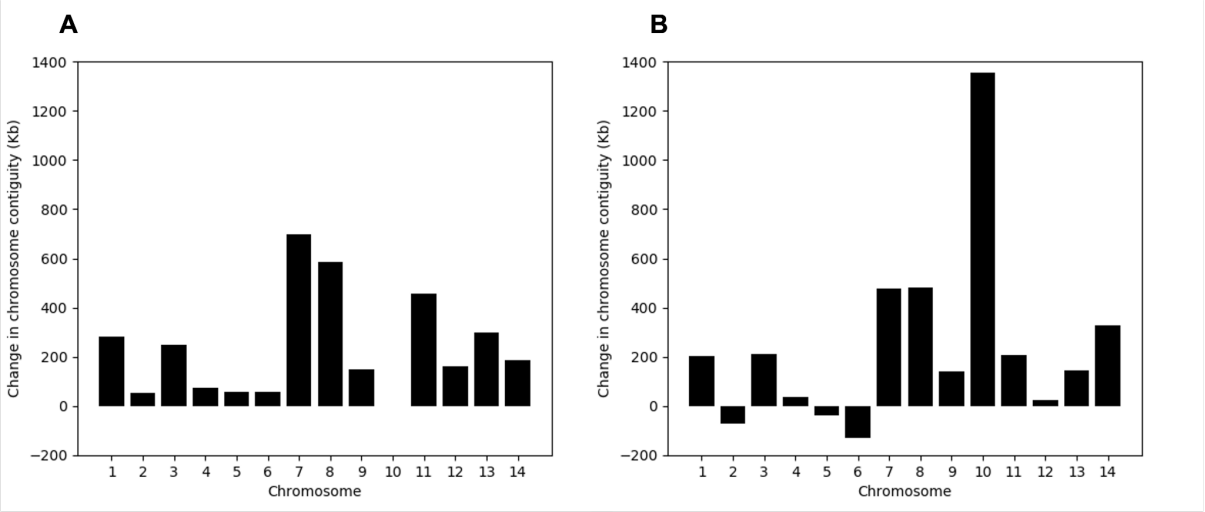


Figure 1. Differences in chromosome length for (A) *P. ovale curtisi (Poc)* and (B) *P. ovale walkeri (Pow)* when comparing the new assemblies against existing references.

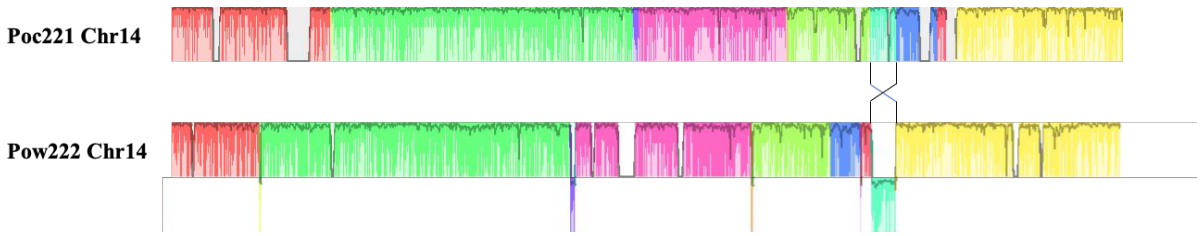


Figure 2. Alignment of *P. ovale spp* chromosome 14 using Mauve software. Homologous regions are coloured accordingly, along with the teal coloured pericentric inversion which contains the predicted centromeric region.

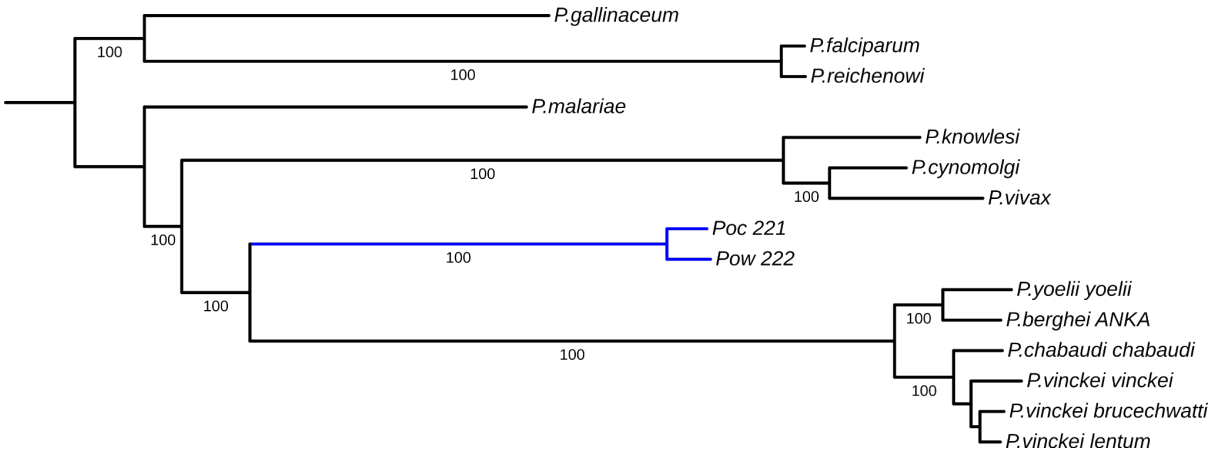


Figure 3. The conserved maximum-likelihood phylogenetic tree consists of 15 *Plasmodium* references including those which infect humans, rodents and poultry. Topology was identical across all derived trees. Branch lengths and Felsenstein bootstrap proportion (FBP) values were derived using the LG+G model.

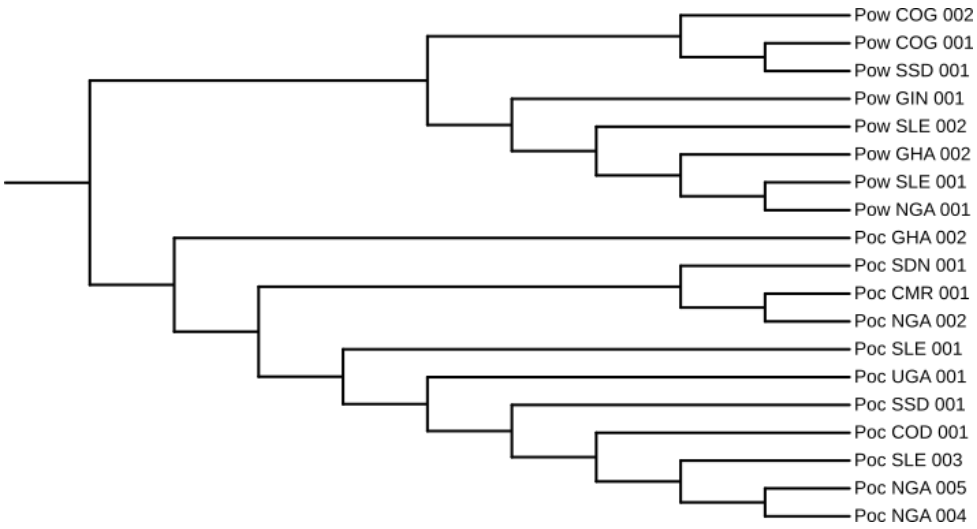


Figure 4. Neighbour-joining tree consisting of *P. ovale curtisi* (*Poc*) and *P. ovale walkeri* (*Pow*) isolates using dimorphic loci identified within conserved gene orthologs.

Supplementary Table 1. Metadata for *P. ovale spp* samples included in this investigation.

Sample ID	Location	Collection Year
Poc_CMR_001	Cameroon	2020
Poc_CMR_002	Cameroon	2019
Poc_COD_001	DRC	2019
Poc_GHA_001	Ghana	2019
Poc_GHA_002	Ghana	2019
Poc_KEN_001	kenya	2020
Poc_NGA_001	Nigeria	2020
Poc_NGA_002	Nigeria	2020
Poc_NGA_003	Nigeria	2020
Poc_NGA_004	Nigeria	2020
Poc_NGA_005	Nigeria	2020
Poc_SDN_001	Sudan	2019
Poc_SLE_001	Sierra leon	2020
Poc_SLE_002	Sierra leon	2020
Poc_SLE_003	Sierra Leon	2019
Poc_SSD_001	South Sudan	2020
Poc_UGA_001	Uganda	2019
Pow_CMR_001	Cameroon	2019
Pow_COD_001	DRC	2019
Pow_COG_001	Congo	2020
Pow_COG_002	Congo	2020
Pow_GHA_001	Ghana	2020
Pow_GHA_002	Ghana	2020

Pow_GIN_001	Guinea	2020
Pow_MOZ_001	Mozambique	2019
Pow_NGA_001	Nigeria	2020
Pow_NGA_002	Nigeria	2020
Pow_SLE_001	Sierra leon	2019
Pow_SLE_002	Sierra Leon	2019
Pow_SSD_001	South Sudan	2019
Pow_TZA_001	Tanzania	2019
Pow_UGA_001	Uganda	2019

Supplementary Table 2. *In-silico* evaluation of *P. ovale spp.* targeted SWGA primers, against *P. ovale curtisi* (PocGH01), *P. ovale walkeri* (PowCR01), Human (GRCh38), *P. falciparum* (Pf3D7) and *P. vivax* (PvP01). Reference genomes obtained from PlasmoDB (https://plasmodb.org/) and NCBI (https://www.ncbi.nlm.nih.gov/genome/guide/human**)** databases

	Binding Sites Per 100 kb				
Primer	**PocGH01**	**PowCR01**	**GRCh38**	**Pf3D7**	**PvP01**
ATTTTCG*A*T	2.0486	2.0967	0.102	1.611	1.674
CGAAAT*T*G	4.9156	5.2431	0.261	1.547	6.271
TATCGT*T*A	5.0798	4.9449	0.289	4.226	2.562
CGAAAAA*A*C	1.8695	1.9177	0.064	0.930	1.880
TCGTAAA*A*A	4.2765	4.3573	0.092	1.569	3.054
TTTACGT*A*T	3.9450	3.8652	0.195	1.650	2.125
CGTAAT*A*A	5.9877	6.0036	0.344	3.870	3.616

Supplementary Table 3. *P. ovale spp* dimorphisms in the *cytB* mitochondrial gene.

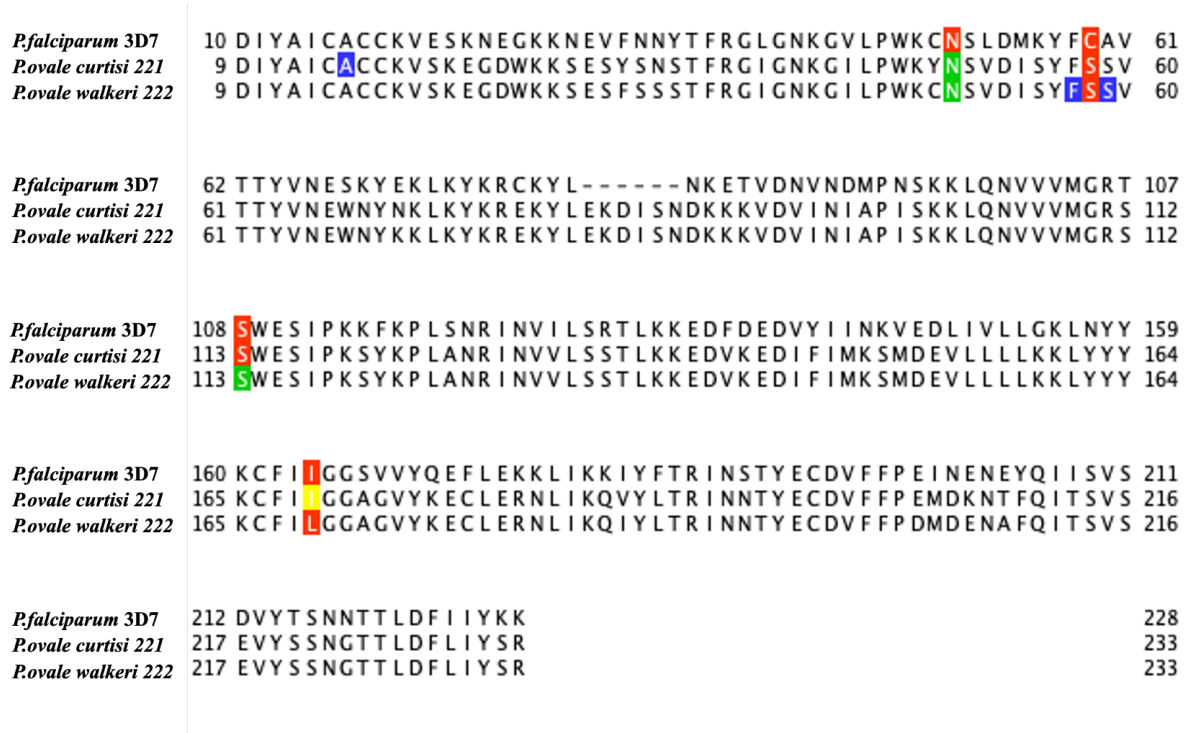
Position	Nucleotide	
	Poc	***Pow***
162	a	c
201	a	g
375	t	a
402	c	t
492	t	a
534	c	t
744	g	t
756	c	t
774	t	a
885	c	t
903	c	t
948	t	a

P. ovale curtisi (*Poc*) and *P. ovale walkeri* (*Pow*)

Supplementary Table 4. BUSCO-based assessment of existing and new *P. ovale curtisi* and *P. ovale walkeri* assemblies.

	Poc_221	**PocGH01**	**Pow_222**	**PowCR01**
Complete BUSCOs (C)	3517	3500	3491	3479
Complete and single-copy BUSCOs (S)	3517	3500	3491	3478
Complete and duplicated BUSCOs (D)	0	0	0	1
Fragmented BUSCOs (F)	33	48	44	34
Missing BUSCOs (M)	92	94	107	129

Supplementary Table 5-8. Please find the relevant supplementary tables at https://github.com/MatthewHiggins2017/Thesis/blob/9ae206ce17c860b3aeb31985580584be0a253e13/Povale%20Supplementary%20Tables.xlsx



Supplementary Figure 1. Alignment of the PfDHFR domain 10-228 with corresponding *P. ovale spp* orthologs. Amino acids are coloured as follows: Red, known resistance position and resistance-associated amino acid present in species population. Yellow, known resistance position but uncharacterised amino acid present in species population. Green, known resistant position and susceptible amino present across all samples. Blue, Unknown position and variant amino acid in population. For *P. falciparum* 3D7 only known resistance sites were highlighted.

Chapter 7. Discussion & Conclusion

Discussion and Conclusion

Discussion

Diagnostics play an essential role in infectious disease control with early malaria diagnosis resulting in a reduced mortality risk and break in transmission chains [1]. "Diagnostic-resistant" parasites have now been observed across multiple sites in Africa, conferred by deletions in *pfhrp2/3* genes targeted by commonly used RDTs [2–4]. Therefore, new efficacious and still cost-effective diagnostics are needed. Nucleic acid amplification technologies (NAATs), including Recombinase Polymerase Amplification (RPA), have the potential to underpin this next-generation of diagnostics, with known improvements in sensitivity and specificity, compared to microscopy or immunochromatic-based malaria RDTs [5–7]. In addition, the wide-scale implementation of NAATs has been successfully demonstrated in the UK and other countries in response to the SARS-CoV-2 pandemic, when used for population level screening. Out of the NAATs available, RPA has the most potential for adaptation to in-field use, due to its isothermal reaction temperature which can be powered by body heat, enabling it to be used ubiquitously in low-resource settings [8]. However, even though the technology was first published in 2006, it remains in the research and development stage due the presence of several hurdles which I sought to address through my thesis [9].

Chapter Two, outlines a bioinformatics tool, PrimedRPA, which I created to optimise RPA assay design, accounting for both assay specificity and target inclusivity [10]. This tool was used to create an RPA assay specifically targeting *P. vivax.* The development of a platform to aid assay design brought RPA in line with other commonly used NAATs such as PCR and LAMP [11]. In addition, prior to PrimedRPA's release, members of the research community were having to design RPA assays by hand, which is a time consuming and lengthy process,

especially when including probe oligos for fluorescence or lateral flow-based detection. Since its release, PrimedRPA has been used globally in the design of numerous pathogen detection assays from SARS-CoV-2 to *Salmonella* [12,13] and has more than 35 scientific citations to date. One future enhancement of PrimedRPA would be converting it from a command-line interface tool to a web-based tool, increasing its accessibility to all members of the research community.

Having a robust tool to enhance RPA assay design, the next step was to explore cost-effective alternatives to lateral-flow based RPA end-point detection to enhance assay sensitivity and align the unit economics of a potential RPA malaria diagnostic with existing immunochromatic RDTs. This was explored in Chapter Three, highlighting several colorimetric approaches including the use of SYBR Green, Malachite Green or a pH-based indicator such as Cresol Red. Whilst this work was cut-short due to restricted laboratory access, a result of the SARS-CoV-2 pandemic, I was able to uncover the difficulties related to this approach and possible routes forward to address such issues and create a low-cost one-step colorimetric assay. Given more time, I would have liked to validate the *in-silico* model experimentally and look to generalise it such that it could be used for the adaptation of any existing and emerging NAATs for colorimetric detection. During my pursuit of a colorimetric RPA-assay, I simultaneously gathered preliminary evidence that RPA could be adapted for SNP genotyping, in line with other NAATs. The deliberate inclusion of primer-template mismatches resulted in the ability to detect several antimalarial resistance markers in the *Pf.kelch13* gene associated with Artemisinin, a core component of the existing front-line therapies for *P. falciparum* in the form of Artemisinin combinatorial therapies (ACT) [14]. The ability for RPA-based diagnostics to move beyond conventional pathogen detection in the

field, to the simultaneous detection of clinically relevant biomarkers that could inform treatment or surveillance efforts, is key to creating the next generation of malaria diagnostics.

The ceaseless arms-race between the development of new treatments and emergence of resistance has been well characterised across infectious diseases [15,16]. As we push towards malaria eradication, it can be safely assumed new biomarkers of interest will emerge, whose monitoring and surveillance will be essential. As such in Chapter Four, I sought to improve on my findings of Chapter Three and characterise the full impact of primer-template mismatch combinations on RPA reaction kinetics to determine if this approach could be generalised into a robust methodology enabling the design of RPA assays to detect any emerging biomarkers of interest. I successfully characterised the impact of 315 mismatch combinations on RPA reaction kinetics and success. I demonstrated that like other NAATs, RPA was susceptible to the detrimental impact of mismatches when positioned towards the 3' terminus of the primer-template complex. In addition, this work was the first to model the RPA reaction profile using a generalised logistic function and present a statistical framework to characterise the impact of mismatches across targets, highlighting that both the position and nucleotides involved in the mismatch are associated with the detrimental impact. Based on the discovered differences in the reaction kinetics, the incorporation of a fluorescent probe could enable RPA-based SNP genotyping, in line with other genotyping NAATs such as TaqMan SNP Assays (Themo Fisher) [17]. However, this is not suitable for low-resource in-field settings, due to the reliance on a fluorescence reader to track reaction progress. Instead, the enzymes within the RPA assay, specifically the polymerase, should be optimised such that in the presence of a single 3' terminal mismatch, covering the SNP site of interest, amplification is completely inhibited. This would create a binary assay, removing the need for a probe and allow the genotyping

method to be combined with the colorimetric end-point detection options explored earlier, enabling in-field use.

The wide-scale use of existing malaria diagnostics has been shown to exert a selective pressure on *Plasmodium spp* populations. This was demonstrated first-hand in the emergence of *PfHRP2* deletions resulting in a rise of false-negatives diagnosis from HRP2 antigen-based immunochromatic RDTs [18]. As such the development and deployment of the next generation of malaria diagnostics, needs to account for possible genetic variants which could impede assay performance through the introduction of mismatches, as highlighted in Chapter Four, or the deletion of a NAAT oligo binding target site. The PrimedInclusivity tool outlined in Chapter Five, enables researchers to achieve this, combining available whole genome sequence data with NAAT specific *in-silico* models to predict assay performance. Through its use, PrimedInclusivity will not only improve assay design, but provide a solution to incorporate WGS surveillance data to ensure assay efficacy is maintained. Compared to previous solutions, such as the use of primer-Blast [19], the PrimedInclusivity software enables users to not only infer binding site diversity but the subsequent impact of such diversity on assay performance and take action as deemed necessary. In addition, the generalised design of PrimedInclusivity allows it to be used for RPA as well as other NAATs that may emerge.

For PrimedInclusivity to achieve its potential in facilitating robust assay design, ongoing whole genome sequencing (WGS) is required to maintain a good inflow of surveillance data to guide assay assessment. Whilst wide-scale WGS datasets exist for *P. falciparum* and *P. vivax*, this is not the case for other human infecting *Plasmodium* species such as *P. ovale curtisi* and *P. ovale walkeri* [20]. Whilst not associated with a high rate of mortality, the understudied *P. ovale spp* parasites have been increasing in prevalence and historically have passed under the radar due

to diagnostic misclassification [21]. As such in Chapter Six, I outlined a primer set to be used in the selective whole genome amplification (SWGA) of *P. ovale spp* from human-blood samples, overcoming the fundamental barrier to *Plasmodium* WGS regarding the high concentration of parasite DNA required and minimising sequencing of background human DNA from clinical samples [22]. I successfully demonstrated *P. ovale spp*, enrichment across 19 samples and the DNA was subsequently sequenced via two platforms, Oxford Nanopore Minion and Illumina Miseq. Through this enrichment I was able to assemble two new reference genomes, via a hybrid approach, improving on both existing *P. ovale spp* genomes available. Prior to this investigation genomic data for six *P.ovale spp* samples was available and as such our contribution increased the availability of this data >3-fold. With the additional data to hand, it provided further evidence that *P. ovale curtisi* and *P. ovale walkeri* are two separate species occurring sympatrically and revealed the presence of previously unreported antimalarial resistance orthologs, which warrant further investigation through phenotypic susceptibility assays. Not only will this improvement in *P. ovale spp* WGS data enhance assay design, but empower other effectors in malaria eradication, for pan-vaccine candidate assessment to antimalarial development.

Conclusion

This thesis presents a wet and dry lab exploration of the RPA technology for the development of the next generation of malaria diagnostics. NAATs, including RPA, have several advantages over existing immunochromatic RDTs, including increased specificity and sensitivity as reported widely in literature [23–25]. This thesis highlights that RPA has the potential to be adapted for end-point colorimetric detection, ensuring the proposed diagnostic is economically viable and in line with existing RDTs, considering the highest malaria burden is associated with low-income countries. Whilst further work is needed to transition

the RPA technology from research setting to a functional in-field diagnostic, this thesis presents a first stepping stone along this journey. As the world refocuses on malaria eradication, improved diagnostics are essential to ensure no parasite reservoirs remain in asymptomatic individuals, as well as for the rapid detection of key biomarkers to inform clinical decision makers to improved patient outcomes.

Future of Malaria Diagnostics, Genomics and Control Efforts.

Lessons learned from the SARS-Cov-2 pandemic should be applied to the global malaria eradication strategy moving forward. Initiatives should target a unified global funding approach with endemic and non-endemic countries contributing regardless of disease burden, given the impact malaria has on the global economy and theoretical geographical expansion of disease burden with climate change [26–28]. To reverse the recent stagnation and rise in malaria incidence, fundamentally more investment is required in implementation and R&D efforts. In the 2016 WHO Global Technical Strategy for Malaria report, an annual funding target of US$ 7.7 billion by 2025 was established to achieve a 75% reduction in incidence, however current estimates indicate we are falling far short of this [29,30]. In addition, the research community must consider the economics of new technologies developed to combat malaria, ensuring they are accessible to all countries, including those with the highest malaria burden which typically are classified as low or middle income.

Whilst the next generation of field-use NAAT-based malaria RDTs are still in the research and development phase, efforts should be made to facilitate their full transition to commercialisation. If implemented correctly, not only would their use improve patient outcomes and break transmission chains as previously mentioned, but enable an accurate quantification of the prevalence of each individual *Plasmodium* species, in turn reflecting a

countries or regions true malaria status. Such information is essential for guiding efficient eradication efforts regarding the distribution of resources. In addition, to ensure eradication efforts are successful, all human infecting *Plasmodium* need to be addressed. Historically, RDT performance has prioritised *P. falciparum* detection, however with the next generation of RDTs, uniform performance across all human infecting *Plasmodium spp* should be prioritised. This reduces the likelihood that a neglected *Plasmodium* species will exploit the environmental niche created by the successful treatment and eradication of another. The ability for the next generation of diagnostics to perform in-field, low-cost, ease-use genotyping will be essential for optimising malaria treatment and guiding whole genome sequencing surveillance efforts. As highlighted in Chapters Three and Four, this could include the detection of known antimalarial resistance markers but can also be used to facilitate personalised antimalarial treatment based on the presence of pharmacogenetic markers. This could include the detection of G6PD deficiency, which is known to impact primaquine dosing for *P.vivax* treatment [31,32]. Outside of a clinical setting, next-generation diagnostics can expand the capacity of wide-scale surveillance strategies which are essential for monitoring the success of eradication interventions, through accurately quantifying changes in *Plasmodium spp* prevalence, in asymptomatic humans and vectors, or the abundance of a particular genotype of interest [33,34].

For next-generation diagnostics to achieve their true potential, an increase in whole genome sequencing (WGS) is required across *non-P. falciparum* species. Enhancing *Plasmodium spp* WGS efforts, will confer a myriad of benefits, from enabling genome wide association studies (GWAS) for the detection and surveillance of new markers of interest, to the ability to track selection within *Plasmodium* populations and identify transmission chains in outbreak settings. Technological developments such as the Oxford Nanopore MinION platform, will

empower such efforts in resource limited settings, given the device is self-contained and portable. Developing frameworks to incorporate available WGS data, such as the PrimedInclusivity tool (Chapter Five), into disease control decision making is essential moving forward, improving the efficacy of eradication efforts.

References

1. Landier, J. *et al.* The role of early detection and treatment in malaria elimination. *Malar. J.* **15**, 363 (2016).
2. Nsobya, S. L. *et al.* Deletions of pfhrp2 and pfhrp3 genes were uncommon in rapid diagnostic test-negative Plasmodium falciparum isolates from Uganda. *Malar. J.* **20**, 4 (2021).
3. Funwei, R. *et al.* Molecular surveillance of pfhrp2 and pfhrp3 genes deletion in Plasmodium falciparum isolates and the implications for rapid diagnostic tests in Nigeria. *Acta Trop.* **196**, 121–125 (2019).
4. Agaba, B. B. *et al.* Systematic review of the status of pfhrp2 and pfhrp3 gene deletion, approaches and methods used for its estimation and reporting in Plasmodium falciparum populations in Africa: review of published studies 2010-2019. *Malar. J.* **18**, 355 (2019).
5. Ahmad, A. *et al.* Comparison of polymerase chain reaction, microscopy, and rapid diagnostic test in malaria detection in a high burden state (Odisha) of India. *Pathog. Glob. Health* **115**, 267–272 (2021).
6. Mogeni, P. *et al.* Detecting Malaria Hotspots: A Comparison of Rapid Diagnostic Test, Microscopy, and Polymerase Chain Reaction. *J. Infect. Dis.* **216**, 1091–1098 (2017).
7. Lalremruata, A. *et al.* Recombinase Polymerase Amplification and Lateral Flow Assay for Ultrasensitive Detection of Low-Density Plasmodium falciparum Infection from Controlled Human Malaria Infection Studies and Naturally Acquired Infections. *J. Clin. Microbiol.* **58**, (2020).
8. Crannell, Z. A., Rohrman, B. & Richards-Kortum, R. Equipment-free incubation of recombinase polymerase amplification reactions using body heat. *PLoS One* **9**, e112146 (2014).

9. Piepenburg, O., Williams, C. H., Stemple, D. L. & Armes, N. A. DNA detection using recombination proteins. *PLoS Biol.* **4**, e204 (2006).
10. Higgins, M. *et al.* PrimedRPA: primer design for recombinase polymerase amplification assays. *Bioinformatics* **35**, 682–684 (2019).
11. Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71–4 (2007).
12. Li, J. *et al.* Recombinase Polymerase Amplification (RPA) Combined with Lateral Flow Immunoassay for Rapid Detection of Salmonella in Food. *Foods* **9**, (2019).
13. Behrmann, O., Bachmann, I., Hufert, F. & Dame, G. *BIOspektrum* **26**, 624–627 (2020).
14. Uwimana, A. *et al.* Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. *Nat. Med.* **26**, 1602–1608 (2020).
15. Hede, K. Antibiotic resistance: An infectious arms race. *Nature* **509**, S2–3 (2014).
16. Ippolito, M. M., Moser, K. A., Kabuya, J.-B. B., Cunningham, C. & Juliano, J. J. Antimalarial Drug Resistance and Implications for the WHO Global Technical Strategy. *Curr Epidemiol Rep* **8**, 46–62 (2021).
17. Gaedigk, A. *et al.* SNP genotyping using TaqMan technology: the CYP2D6*17 assay conundrum. *Sci. Rep.* **5**, 9257 (2015).
18. Gatton, M. L. *et al.* Impact of Plasmodium falciparum gene deletions on malaria rapid diagnostic test performance. *Malar. J.* **19**, 392 (2020).
19. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).
20. MalariaGEN *et al.* An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. *Wellcome Open Res* **6**, 42 (2021).
21. Alemu, A., Fuehrer, H.-P., Getnet, G., Tessema, B. & Noedl, H. Plasmodium ovale curtisi and Plasmodium ovale wallikeri in North-West Ethiopia. *Malar. J.* **12**, 346 (2013).
22. Ibrahim, A. *et al.* Selective whole genome amplification of Plasmodium malariae DNA from clinical samples reveals insights into population structure. *Sci. Rep.* **10**, 10832 (2020).
23. Wardhani, P. *et al.* Performance comparison of two malaria rapid diagnostic test with real time

polymerase chain reaction and gold standard of microscopy detection method. *Infect. Dis. Rep.* **12**, 8731 (2020).

24. Berzosa, P. *et al.* Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea. *Malar. J.* **17**, 333 (2018).
25. Leski, T. A. *et al.* Use of real-time multiplex PCR, malaria rapid diagnostic test and microscopy to investigate the prevalence of Plasmodium species among febrile hospital patients in Sierra Leone. *Malar. J.* **19**, 84 (2020).
26. Caminade, C. *et al.* Impact of climate change on global malaria distribution. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 3286–3291 (2014).
27. Gallup, J. L. & Sachs, J. D. *The Economic Burden of Malaria*. (American Society of Tropical Medicine and Hygiene, 2001).
28. Micah, A. E. *et al.* Tracking development assistance for health and for COVID-19: a review of development assistance, government, out-of-pocket, and other private spending on health for 204 countries and territories, 1990–2050. *Lancet* **398**, 1317–1343 (2021).
29. White, N. J., Day, N. P. J., Ashley, E. A., Smithuis, F. M. & Nosten, F. H. Have we really failed to roll back malaria? *Lancet* **399**, 799–800 (2022).
30. Programme, G. M. Global technical strategy for malaria 2016-2030, 2021 update. https://www.who.int/publications/i/item/9789240031357 (2021).
31. Watson, J., Taylor, W. R., Menard, D., Kheng, S. & White, N. J. Modelling primaquine-induced haemolysis in G6PD deficiency. *Elife* **6**, (2017).
32. Malik, S., Zaied, R., Syed, N., Jithesh, P. & Al-Shafai, M. Seven novel glucose-6-phosphate dehydrogenase (G6PD) deficiency variants identified in the Qatari population. *Hum. Genomics* **15**, 61 (2021).
33. Gnambani, E. J., Bilgo, E., Sanou, A., Dabiré, R. K. & Diabaté, A. Infection of highly insecticide-resistant malaria vector Anopheles coluzzii with entomopathogenic bacteria Chromobacterium violaceum reduces its survival, blood feeding propensity and fecundity. *Malar. J.* **19**, 352 (2020).

34. Hancock, P. A., Sinkins, S. P. & Godfray, H. C. J. Strategies for introducing Wolbachia to reduce transmission of mosquito-borne diseases. *PLoS Negl. Trop. Dis.* **5**, e1024 (2011).