A novel metabarcoded DNA sequencing tool for the detection of *Plasmodium* species in malaria positive patients.

Abdul Wahab[a1], Ayaz Shaukat[b1], Qasim Ali[c1], Mubashir Hussain[a], Taj Ali Khan[a], M. Azmat Ullah Khan[b], Imran Rashid[d], Mushtaq A. Saleem[b], Mike Evans[e], Neil D. Sargison[e]*, Umer Chaudhry[f]*

[a] *University of Science and Technology, Kohat, Khyber Pakhtunkhwa, Pakistan*

[b] *University of Central Punjab, Lahore, Punjab, Pakistan*

[c] *Gomal University, Dera Ismail Khan, Khyber Pakhtunkhwa, Pakistan*

[d] *University of Veterinary and Animal Sciences, Lahore, Punjab, Pakistan*

[e] *Royal Dick School of Veterinary Studies, University of Edinburgh, UK*

[f] *Roslin Institute, University of Edinburgh, UK*

1 Contributed equally

*Corresponding author:
Umer Chaudhry
Email: uchaudhr@exseed.ed.ac.uk, Tel: 00441316519244
University of Edinburgh, The Roslin Institute, Easter Bush Veterinary Centre, UK, EH25 9RG

Neil Sargison
Neil.Sargison@ed.ac.uk, +44 (0)131 651 7300
University of Edinburgh, R(D)SVS, Easter Bush Veterinary Centre, UK, EH25 9RG

Abstract

Various PCR based methods have been described for the diagnosis of malaria, but most depend on the use of *Plasmodium* species-specific probes and primers; hence only the tested species are identified and there is limited available data on the true circulating species diversity. Sensitive diagnostic tools and platforms for their use are needed to detect *Plasmodium* species in both clinical cases and asymptomatic infections that contribute to disease transmission. We have been recently developed for the first time a novel high throughput 'haemoprotobiome' metabarcoded DNA sequencing method and applied it for the quantification of haemoprotozoan parasites (*Theleria* and *Babesia*) of livestock. Here, we describe a novel, high throughput method using an Illumina MiSeq platform to demonstrate the proportions of *Plasmodium* species in metabarcoded DNA samples derived from human malaria patients. *Plasmodium falciparum* and *Plasmodium vivax* positive control gDNA was used to prepare mock DNA pools of parasites to evaluate the detection threshold of the assay for each of the two species and to assess the accuracy of proportional quantification. We then applied the assay to malaria-positive human samples to show the species composition of *Plasmodium* communities in the Punjab province of Pakistan and in the Afghanistan-Pakistan tribal areas. The diagnostic performance of the deep amplicon sequencing method was compared to an immunochromatographic assay that is widely used in the region. Metabarcoded DNA sequencing showed better diagnostic performance, greatly increasing the estimated prevalence of *Plasmodium* infection. The next-generation sequencing method using metabarcoded DNA has potential applications in the diagnosis, surveillance, treatment, and control of *Plasmodium* infections, as well as to study the parasite biology.

Keywords: Malaria, *Plasmodium falciparum* and *Plasmodium vivax*, metabarcoded DNA, deep amplicon sequencing.

1. Introduction

Malaria is the most important vector-borne disease, causing high morbidity and mortality (Moody, 2002). More than 3.4 billion people are infected, resulting in an estimated 1.2 billion malaria cases every year (Poostchi et al., 2018). The disease is caused by intracellular protozoan parasites of the genus *Plasmodium* transmitted through *Anopheles* mosquitoes (Dash et al., 2007; Sinka et al., 2012). *Plasmodium* has an indirect life cycle including one stage in *Anopheles* mosquitoes and three different stages in humans, all with different rates of replication (Li et al., 1997). Gametogenesis occurs in human blood and fertilisation of male and female macro and microgametes occurs in the midgut of the mosquito after feeding. Asexual stages occur in the gut of the mosquito as sporogony, and after biting the humans, sporozoites undergo exoerythrocytic schizogeny in the hepatic cells and then erythrocytic schizogony in the blood cells (Li et al., 1994b). Five species of *Plasmodium* parasite infect humans, namely *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae* and *Plasmodium knowlesi*. *P. falciparum* is the most associated with lethal disease, but in recent years, there has been an increase in disease severity attributable to *P. vivax* (Saravu et al., 2014; Scholzen et al., 2014; William et al., 2011).

The 18S rDNA of *Plasmodium* is unique due to its genomic arrangement dispersed among different chromosomes. The copy number is limited to 4 to 8 per genome and the sequences are not identical (Li et al., 1994b). Their expression is regulated during different development stages of the life cycle (Li et al., 2014); for example, at least three types of genotypic variants have been identified between stages of *P. falciparum* and *P. vivax* laboratory isolates (Li et al., 1994b; McCutchan et al., 1988; Qari et al., 1994; Rogers et al., 1995). However, the presence of these variants has not been reported in the field studies.

The 18S rDNA of *Plasmodium* forms a mosaic of conserved and variable regions; whereby the conserved regions contribute to form a secondary structure of rRNA that appears to be associated with the universal function of the ribosomes. The variable regions are scattered among the conserved regions and contribute to major differences in gene composition and size (Li et al., 1994a). The function of variable regions is not fully understood, but determination of sequence variations can discriminate between *Plasmodium* species (Agudelo et al., 2013; Haanshuus et al., 2013; Lee et al., 2015; Lefterova et al., 2015), and overcome limitations of traditional microscopic and immunochromatographic methods for the diagnosis of this group of parasites at species level.

Significant progress has been made in the global fight against malaria through high throughput rapid diagnosis. Sensitive diagnostic tools are needed to detect clinically and subclinically infected patients (Echeverry et al., 2016). Molecular methods including qPCR, species-specific PCR, nested PCR, and multiplex PCR have been described (Canier et al., 2013; Cunha et al., 2009; Das et al., 1995; Echeverry et al., 2016; Haanshuus et al., 2013; Steenkeste et al., 2009), but these are low throughput, hence relatively expensive (Chaudhry et al., 2019). These methods depend on the use of species-specific probes and primers, meaning that only the tested species are identified, hence are limited in their ability to describe true circulating species diversity (Moody, 2002). In contrast, high throughput metabarcoded DNA sequencing using the Illumina MiSeq platform is relatively low-cost and potentially less error-prone. We have applied this 'haemprotobiome' method, to the study of tick-borne haemoprotozoan parasites of ruminants (Chaudhry et al., 2019). The method has the potential to open new areas of research in the study of *Plasmodium*, to accurately provide relative quantification of co-infecting species and to evaluate drug treatment responses (Shaukat et al., 2019). The method uses primers binding to the conserved sites and analyse of up to 600 bp sequence reads. The use of adapter and barcoded primers allows a large number of samples to be pooled and sequenced in a single MiSeq flow cell, making the assay suitable for high-throughput analysis (Shaukat et al., 2019).

Here, we report for the first time the development of a deep sequencing method using the Illumina MiSeq platform to quantify *P. falciparum* and *P. vivax* present in malaria-positive human blood samples. The results are compared with a standard immunochromatographic assay to validate the method's accuracy for species identification.

2. Materials and methods

2.1. Parasite material

Positive control samples of *P. falciparum* were kindly provided by D Jason Mooney at the Roslin Institute, University of Edinburgh, UK, previously obtained from National Institute for Biological Standards and Control (NIBSC code 04/176). *P. vivax* control samples were kindly provided by Dr Imran Rashid at the University of Veterinary and Animal Sciences, Lahore, Pakistan, previously obtained from the Biodefense and Emerging Infections Research Resources Repository (BEI code MRA-41). Four replicates each of mock pools comprising of *P. falciparum* only (Mix 1), *P. vivax* only (Mix 2), and *P. falciparum* and *P. vivax* (Mix 3) were created. These were used to test the detection threshold of the metabarcoded sequencing

method and to show the proportions of each of the *Plasmodium* species present. Four negative control of human blood samples were provided by Sana Amir and Saqib Shahzad, Chughtai Diagnostic Laboratory, Lahore Pakistan.

Malaria suspected patients referred to Basic Health Units in the tribal areas of the Afghanistan-Pakistan border and Chughtai Diagnostic Laboratory in the Punjab province of Pakistan were invited to participate in this study. Prior discussions were held with key administrative and community leaders to raise awareness of the study. Samples were taken by trained para-medical workers under the supervision of local collaborators and the Basic Health Unit or Chughtai Diagnostic Laboratory staff. The institutional review boards of the University of Central Punjab (UCP-30818), and the Kohat University of Science and Technology, Pakistan (KUST/EC/1379) approved the study. Patients of all age groups were included in this study with symptoms consistent with malaria, including vomiting, fever, headache, chills, sweats, nausea and fatigue.

Blood samples were collected by venipuncture during peak malaria transmission seasons between August to November 2017 and 2018. A total of 5 ml of intravenous blood was drawn into EDTA tubes and stored at -20 ^{0}C for gDNA isolation. Each sample was also routinely analysed by microscopic examination under oil immersion (x1000) of 4% Giemsa-stained blood smears for the diagnosis of malaria. The *Plasmodium* goes through different stages of their development cycle (48 hr), which gives the parasites a different visual appearance that can be observed under the microscope. These stages show the ring (Fig. 1A), tropnozoite, schizont, and gametocyte appearance. Malaria case identification was based on the appearance of those stages on the microscopic examination (Fig. 1A). Overall, 365 malaria suspected positive patients were identified.

2.2. Immunochromatographic assay

The 365 malaria-positive on microscopic examination blood samples, each having an unknown level of parasitemia, were analysed using a commercial immunochromatographic rapid diagnostic test (RDT) kit. The Malaria Pf/Pv Ag Rapid Test (Healgen®; Zhejiang Orient Gene Biotech Co, Ltd) RDT kit was designed to detect *P. falciparum*-specific histidine-rich protein 2 (Pf-HRP2) and *P.vivax*-specific lactate dehydrogenase (Pv-LDH). The kit was transported and maintained at the room temperature, opened just before use to avoid humidity damage, and used in accordance with the manufacturer's recommendations. During the assay, an adequate volume of the blood sample was dispensed into the sample well 'S' of the test cassette and the lysis buffer is added to the buffer well 'B'. The buffer contains a detergent

that lyses the red blood cells and releases antigens, which migrate by capillary action across the strip held in the cassette. If Pf-HRP2 binds to the HRP2 gold conjugates and the immunocomplex is then captured on the membrane by the pre-coated anti-Pf-HRP2 antibodies, forming a burgundy colored pf band, indicating *P. falciparum* positive test (Fig. 1B). If Pv-LDH binds to the LDH gold conjugates and the immunocomplex is then captured on the membrane by the pre-coated anti-Pv-LDH antibodies, forming a burgundy colored pv band, indicating *P. vivax* positive test (Fig. 1B). The absence of any band suggests a negative result. The test also contained an internal control 'C' band, exhibiting a burgundy colored band of the immunocomplex of goat anti-mouse IgG/mouse IgG (anti-Pv-LDH and anti-Pf-HRP2) gold conjugates, regardless of the color development on 'C' band.

2.3. Genomic DNA isolation, primer design and adapter/barcoded PCR amplification of rDNA 18S locus

50 µl of blood from each of the 365 samples was used as a template, to extract gDNA according to the protocols described in the TIANamp blood DNA kit (Tiangen Biotech (Beijing) Co., Ltd). Overall, 589 bp and 568 bp fragments of *P. falciparum* and *P. vivax* 18S rDNA, respectively, were amplified using newly developed adapter primer sets (Supplementary Table S1). The overall scheme of the sample preparation is described in Figure 2A. Adapters were added to these primers to allow the successive annealing of subsequent metabarcode primers and N is the number of random nucleotides included between the locus-specific primers and adapter to increase the variety of generated amplicons as previously described by Chaudhry et al. (2019). Four forward (Plasmo1_For, Plasmo1_For-1N, Plasmo1_For-2N, Plasmo1_For-3N) and four reverse (Plasmo2_Rev, Plasmo1_Rev-1N, Plasmo1_Rev-2N, Plasmo1_Rev-3N) primers were mixed in equal proportion (Supplementary Table S1) and used for first-round PCR under the following conditions: 5X KAPA buffer, 10mM dNTPs, 10 uM forward and reverse adapter primer, 0.5 U KAPA Polymerase (KAPA Biosystems, USA), and 1 ul of worm lysate. The thermocycling conditions of the PCR were 95^{o}C for 2 min, followed by 35 cycles of 98^{o}C for 20 sec, 60^{o}C for 15 sec, 72^{o}C for 15 sec and a final extension 72^{o}C for 5 min. PCR products were purified with AMPure XP Magnetic Beads (1X) (Beckman Coulter, Inc.).

After the purification, a second-round PCR was performed by using sixteen forward and twenty-four reverse barcoded primers. The barcoded forward (N501 to N516) and reverse (N701 to N724) primers (10 uM each) were previously described by Chaudhry et al. (2019). The primers were used in a manner that repetition of same forward and reverse sequences did

not occur in the different samples. The second-round PCR conditions were: 5X KAPA buffer, 10 mM dNTPs 0.5 U KAPA Polymerase (KAPA Biosystems, USA), and 2 ul of first-round PCR product as DNA template. The thermocycling conditions of the second round PCR were 98^{o}C for 45 sec, followed by 7 cycles of 98^{o}C for 20 sec, 63^{o}C for 20 sec, and 72^{o}C for 2 minutes. PCR products were purified with AMPure XP Magnetic Beads (1X) according to the protocols described by Beckman Coulter, Inc.

2.4. Sequencing of metabarcoded 18S rDNA, data handling, and bioinformatic analysis

The pooled library was measured with KAPA qPCR library quantification kit (KAPA Biosystems, USA). The prepared library was then run on Illumina MiSeq Sequencer using a 600-cycle pair-end reagent kit (MiSeq Reagent Kits v2, MS-103-2003) at a concentration of 15 nM with an addition of 25% Phix Control v3 (Illumina, FC-11-2003).

The overall scheme of the data handling and bioinformatics analysis is described in Figure 2B. MiSeq data were handled with our own bioinformatics pipeline (Chaudhry et al., 2019). Briefly, MiSeq separates all sequence reads during post-run processing using the barcoded indices and to generate FASTQ files. The raw paired read-ends were run into the 'make.contigs' command to combine the two sets of reads for each sample. The command extracts sequence and quality score data from the FASTQ files, creating the complement of the reverse and forward reads, and then joining the reads into contigs. After removing the too long, or ambiguous sequence reads, the data were then aligned with the *P. falciparum* and *P. vivax* reference sequence library (for more details Supplementary Data S1 and Result section 3.1) using the 'align.seqs' command to summarise the 589 bp and 568 bp fragments encompassing parts of the 18S DNA spanning the hyper-variable region of the *P. falciparum* and *P. vivax* ribosomal cistrons. At this stage, the 18S rDNA analysis was completed by classifying the sequences into either of the two species by using the 'classify.seqs' command and creating a taxonomy file by using the 'summary.tax' command. Overall, 762674 million reads of 18S rDNA were generated from the data set.

For the phylogenetic analysis of *P. falciparum* and *P. vivax* 18S rDNA, all the classified sequences were run on the 'screen.seqs' command and the count list of the consensus sequences of each sample was created using the 'unique.seqs' command followed by the use of the 'pre.cluster' command to look for sequences differences and to merge them in groups based on their abundance. Any chimeras were identified and removed by using the 'chimera.vsearch' command. The count list was further used to create the FASTQ files of the

consensus sequences of each sample using the 'split.groups' command (for more details Supplementary Data S2).

2.5. Split and maximum-likelihood trees of P. falciparum and P. vivax *18S rDNA sequences*

P. falciparum and *P. vivax* 18S rDNA sequence reads were analysed separately in Geneious v9.0.1 software (Kearse et al., 2012) using the MUSCLE alignment tool. The aligned sequences were then imported into the FaBox 1.5 online tool to collapse those with 100% base pair similarity after corrections into single genotypes. The split tree of *P. falciparum* and *P. vivax* 18S rDNA was created in the SplitTrees4 software by using the UPGMA method in the Jukes-Cantor model of substitution (Huson and Bryant, 2005). The maximum-likelihood tree for *P. falciparum* and *P. vivax* 18S rDNA was constructed by the HKY model of substitution in the MEGA 7 software and to select the appropriate model of nucleotide substitutions (Tamura et al., 2013).

2.6. Statistical analysis

The data were analysed using CompareTests: Correct for Verification Bias in Diagnostic Accuracy and Agreement Statistics software (R package version 1.2.). The frequency of *P. falciparum* and *P. vivax* in the samples was calculated by dividing the number of sequence reads for each sample by the total number of reads. The effect of the mock pools of *P. falciparum* and *P. vivax* positive controls were analysed by running a Kruskal-Wallis rank-sum test for each admixture. The performance of the immunochromatographic assay was compared against metabarcoded sequencing, using a Fisher's Exact test to calculate the predictive value and Category-Specific Classification Probability (CSCP) with 95% confidence interval. Kappa (k) values were calculated to express the agreement beyond chance; where values greater than 0.80 were considered to represent perfect agreement; values of 0.61 - 0.80 to represent good agreement; and values of 0.21 - 0.60 to represent moderate agreement.

3. Results

3.1. P. falciparum and P. vivax *18S rDNA reference sequence libraries*

The sequence reads generated by metabarcoded 18S rDNA sequencing of the *P. falciparum* and *P. vivax* positive controls were compared to the malaria-positive samples and to the published NCBI GenBank 18S rDNA sequences to account for any genetic diversity. A

total of 34 *P. falciparum* and 74 *P. vivax* reference sequences were identified (Fig. 3, Supplementary Data S1). The maximum-likelihood tree shows distinct clustering of the *P. falciparum* and *P. vivax* 18S rDNA regions (Fig. 3), hence these closely related *Plasmodium* species can be reliably differentiated by virtue of 18S rDNA sequence variations.

3.2. Validation of the metabarcoded sequencing assay using mock pools of P. falciparum and P. vivax

Four replicates each of *P. falciparum* only (Mix 1), *P. vivax* only (Mix 2), and *P. falciparum* and *P. vivax* (Mix 3) were created from gDNA to demonstrate the detection accuracy of the metabarcoded DNA sequencing method and to show the proportions of each of the species being present (Fig. 4; Supplementary Table S2). The mixing of different mock pools demonstrates the accurate detection ability of the metabarcoded sequencing method and to show the proportions of each of the species being present. The Mix 1 pool yielded only *P. falciparum* sequence reads and the Mix 2 pool yielded only *P. vivax* sequence reads (Fig. 4). The Mix 3 pool yielded both *P. falciparum* and *P. vivax* sequence reads, with no statistically significant variations between replicates (Kruskal-Wallis rank-sum test; Mix1: $\chi 2(1)$ 0, p=1; Mix2: $\chi 2(1)$ 0, p=1; Mix3: $\chi 2(3)$ 0.02153, p=0.5231).

3.3. Assessment of the immunochromatographic assay and metabarcoded sequencing for the identification of P. falciparum and P. vivax

The immunochromatographic assay and metabarcoded DNA sequencing methods were applied to malaria-positive blood samples to detect *P. falciparum* and *P. vivax* in the field (Fig. 5, Supplementary Table S3). The results of both assays demonstrate that the prevalence of *P. vivax* infection was higher than that of *P. falciparum* infection. In the case of the metabarcoded DNA sequencing assay, those samples yielding more than 1000 reads (implying sufficient gDNA for accurate amplification) were included in the analysis (Supplementary Table S3). *Plasmodium vivax* was present in 199 (69.8%) patients, *P. falciparum* in 84 (29.5%) and mixed infection in 2 (0.7%) patients (Fig. 5). The immunochromatographic assay showed that *Plasmodium vivax* was present in 187 (65.6%) patients, *P. falciparum* in 78 (27.4%), mixed infection in 2 (0.7%) patients (Fig. 5) and 18 (6.32%) malaria-positive cases were negative in RDT, but positive in the metabarcoded DNA sequencing assay (Fig. 5).

The degree of agreement between the immunochromatographic and metabarcoded DNA sequencing assays was high with κ = 0.893 (95% CI: 0.839-0.930). The Category-Specific

Classification Probability (CSCP) was also high in all four categories, ranging from 0.892 (95% CI: 0.807 - 0.943) to 1. The Predictive Values (PV) for 'positive RDT' results were also very high, ranging from 0.962 (0.860 - 0.990) to 1. However, the PVs for 'negative RDT' results were 0.816 (0.724 - 0.883); being significantly lower than the Predictive Value for both 'RDT *P. falciparum*' ($p=0.004$) and 'RDT *P. vivax*' ($p<0.001$) results. In samples with similar disease prevalence, there is, therefore, an increased likelihood that a negative RDT result may be incorrect.

3.4. Phylogenetic analysis of the P. falciparum and P. vivax rDNA 18S sequences

Overall, 112 different genotypes of the *P. falciparum* 18S rDNA locus were identified among 84 field samples and 12 from the NCBI GenBank sequences (Supplementary Data S2 and section 3.1. and 3.3). The split tree shows at least two distinct clades (Fig. 6A), and it sets apart that 9 genotypes in the Clade I, belongs to a 'type S' 18S rDNA region (McCutchan et al., 1988). The remaining 115 genotypes were in clade II, belonging to a 'type A' 18S rDNA region (McCutchan et al., 1988). 80 genotypes predominated in group 1 and 35 genotypes in group 2 (Fig. 6A). In contrast, 30 different genotypes of the *P. vivax* 18S rDNA locus was identified among 199 field samples and 18 from the NCBI GenBank sequences (Supplementary Data S2 and section 3.1. and 3.3). The split tree shows at least two distinct clades (Fig. 6B), and it sets apart that 3 genotypes in the Clade I, belong to the 'type S' 18S rDNA as described by Li et al. (1994). The remaining 45 genotypes were in clade II, belonging to 'type A' 18S rDNA (Li et al., 1994b). 7 genotypes predominated in groups 1 and 2, and 31 genotypes in group 3 (Fig. 6B).

4. Discussion

Microscopic examination of blood smears remains the mainstay for the diagnosis of *Plasmodium* in the field. The procedure allows the detection of levels of at least 200 parasites/ul, which is sufficient for the diagnosis of most symptomatic cases, but can result in misdiagnoses at low levels of parasitemia (Rakotonirina et al., 2008; Wongsrichanalai et al., 2007). The method is labor-intensive, time-consuming, and interpretation of results requires highly skilled microscopists (Canier et al., 2013; Echeverry et al., 2016).

The immunochromatographic assay is also utilised in case investigation and malaria surveillance programs in the field. The method depends on the incorporation of conjugated monoclonal antibodies providing the indicator of infection (Moody, 2002; Wongsrichanalai et al., 2007). The targeted antigens are abundant in the sexual and asexual stages of the parasites,

while histidine-rich protein 2 (Pf-HRP2) based RDT is specific for *P. falciparum* and lactate dehydrogenase (Pf-LDH & Pv-LDH) specific for detecting both *P. falciparum* and *P. vivax*. Pan based RDT targets specific antigens including lactate dehydrogenase (P-LDH) and aldolase proteins found in all *Plasmodium* species (Akinyi Okoth et al., 2015; Murillo Solano et al., 2015). Several factors potentially affect the accuracy and false-negative results of RDT including the interpretation of the test strip colour change, the density of malaria infection in the host, improper storage/handling of the kit and poor test performance (Echeverry et al., 2016). Besides this, other major factors such cross-reactivity of HRP2 with histidine-rich protein 3 (a structural homolog with significant sequence similarity) and deletions in the HRP2 locus in *P. falciparum* isolates may account for false-negative results (Akinyi Okoth et al., 2015; Rachid Viana et al., 2017).

Conventional PCR based molecular methods are useful in the detection of *Plasmodium* species for which the reagents and conditions have been developed, but have limitations in terms of lacking scalability (Canier et al., 2013; Cunha et al., 2009; Das et al., 1995; Echeverry et al., 2016; Haanshuus et al., 2013; Steenkeste et al., 2009). The diagnostic challenges of the disease identification have not been resolved yet (Wongsrichanalai et al., 2007), therefore the metabarcod DNA sequencing potentially provides a more accurate and reliable automated high-throughput method to detect *Plasmodium* species in blood samples. The use of a single PCR utilising primers conserved between *Plasmodium* species provides a powerful tool to measure the relative sequence representation of each species in the blood samples. In the present study, we have evaluated a metabarcoded DNA sequencing method to identify the presence of *P. falciparum* and *P. vivax* using mock parasite pools, applying the method to malaria-positive blood samples, and the detection *P. falciparum* and *P. vivax* rDNA18S genotypic variants in the field samples.

We tested the ability of the metabarcoded DNA sequencing assay to accurately determine the relative species proportions in combinations of *P. falciparum* and *P. vivax*. To do this, we generated mock pools containing different estimated proportions of both species and demonstrated no significant variations between replicates. A previous study using pools of laboratory-maintained *Theileria* and *Babesia* haemoprotozoan parasites showed that the relative sequence representation was unaffected by either the number of PCR cycles employed or the parasite species composition of the sample. This study also found no sequence representation bias in PCR products used for sequencing, arising from the number of first-round PCR cycles (Chaudhry et al., 2019).

After validating the metabarcoded DNA sequencing assay using mock pools of *Plasmodium* positive DNA, we applied the method to field samples collected from suspected malaria-positive patients in the tribal area of the Afghanistan-Pakistan border and in the Punjab province of Pakistan, where malaria caused by *P. falciparum* and *P. vivax* has been reported (Kakar et al., 2010; Khattak et al., 2013). Our findings support previous reports of the high prevalence of malaria in the tribal areas of the Afghanistan-Pakistan border, and increasing prevalence over the last few decades in the Punjab province (Kakar et al., 2010; Khattak et al., 2013). Our results support the reports suggesting that while the majority of the cases of malaria are caused by *P. vivax*, the prevalence of *P. falciparum* has increased during recent years (Khattak et al., 2013). The increased prevalence of *P. falciparum* may be an attribute to antimicrobial resistance; previous studies have shown that the pyrimethamine and chloroquine resistance mutations in *P. falciparum* are present in different cities of Pakistan (Ghanchi et al., 2011). Another explanation for the increased prevalence of *P. falciparum* may be provided by the influx of people and the movement of refugees from areas of Afganistan where the parasite species is common (Howard et al., 2011).

Three structurally distinct types of rDNA 18S genotypic variants have been reported in *P. falciparum* and *P. vivax* laboratory isolates (Li et al., 1994b; McCutchan et al., 1988; Qari et al., 1994; Rogers et al., 1995). The existence of genotypic variants in the field studies has not been described. In the present study, we have identified the two independent gene duplication events that occurred in *P. falciparum*, leading to the A and S type rDNA18S lineages; the type A lineage being the ancestor of at least two groups (Fig. 6A). In *P. vivax,* we identified two independent gene duplication events, also leading to the A and S type rDNA18S lineages; the type A lineage being the ancestor of at least three groups (Fig. 6B). In the previous reports, type A was transcribed in erythrocytic schizogony and gametocyte stages consistent with those stages that could have been represented in the present study. In these reports, type S was transcribed in the exoerythrocytic schizogeny stage, while type O was associated with oocyst development only in infected mosquitoes (Li et al., 1997). These observed differences in the 18S loci of *P. falciparum* and *P. vivax* field samples confirm the presence of genotypic variants. Better understanding is needed of the function of these structurally distinct ribosomes that are active with enhanced transcription during different stages of parasitic development in *Plasmodium,* with reference to the development of disease control strategies.

In conclusion, we describe for the first time the use of metabarcoded DNA sequencing using an Illumina MiSeq platform to quantify *P. falciparum* and *P. vivax*, and demonstrate its accuracy on malaria-positive samples. Our results provide a proof of concept for the use of

the method in disease surveillance, similar to its application in the study of haemoprotozoan parasites of livestock (Chaudhry et al., 2019) and dogs (Huggins et al., 2019). This work was undertaken to explore the possibilities for the application of this high throughput method to determine the dynamics of co-infections, disease biology and epidemiology in *Plasmodium* parasites, and has applications in monitoring the changes in parasite populations after the emergence and spread of antimicrobial drug resistance (Shaukat et al., 2019).

Acknowledgment

The study was financially supported by the Roslin Institute uses facilities funded by the Biotechnology and Biological Sciences Research Council (BBSRC). Work at the University of Veterinary and Animal Science Pakistan and Kohat University of Science and Technology Pakistan uses facilities funded by the Higher Education Commission of Pakistan.

Conflict of interest

None

References

Agudelo, O., Arango, E., Maestre, A., Carmona-Fonseca, J., 2013. Prevalence of gestational, placental and congenital malaria in north-west Colombia. Malar. J. 12, 1475-2875.

Akinyi Okoth, S., Abdallah, J.F., Ceron, N., Adhin, M.R., Chandrabose, J., Krishnalall, K., Huber, C.S., Goldman, I.F., Macedo de Oliveira, A., Barnwell, J.W., Udhayakumar, V., 2015. Variation in *Plasmodium falciparum* histidine-rich protein 2 (Pfhrp2) and *Plasmodium falciparum* histidine-rich protein 3 (Pfhrp3) gene deletions in Guyana and Suriname. PLoS One. 10, e0126805.

Canier, L., Khim, N., Kim, S., Sluydts, V., Heng, S., Dourng, D., Eam, R., Chy, S., Khean, C., Loch, K., Ken, M., Lim, H., Siv, S., Tho, S., Masse-Navette, P., Gryseels, C., Uk, S., Van Roey, K., Grietens, K.P., Sokny, M., Thavrin, B., Chuor, C.M., Deubel, V., Durnez, L., Coosemans, M., Menard, D., 2013. An innovative tool for moving malaria PCR detection of parasite reservoir into the field. Malar. J. 12, 405.

Chaudhry, U., Ali, Q., Rashid, I., Shabbir, M.Z., Ijaz, M., Abbas, M., Evans, M., Ashraf, K., Morrison, I., Morrison, L., Sargison, N.D., 2019. Development of a deep amplicon sequencing method to determine the species composition of piroplasm haemoprotozoa. Ticks Tick Borne Dis. 10, 101276.

Cunha, M.G., Medina, T.S., Oliveira, S.G., Marinho, A.N., Povoa, M.M., Ribeiro-dos-Santos, A.K., 2009. Development of a Polymerase Chain Reaction (PCR) method based on amplification of mitochondrial DNA to detect *Plasmodium falciparum* and *Plasmodium vivax*. Acta Trop. 111, 35-38.

Das, A., Holloway, B., Collins, W.E., Shama, V.P., Ghosh, S.K., Sinha, S., Hasnain, S.E., Talwar, G.P., Lal, A.A., 1995. Species-specific 18S rRNA gene amplification for the detection of *P. falciparum* and *P. vivax* malaria parasites. Mol. Cell. Probes. 9, 161-165.

Dash, A., Adak, T., Raghavendra, K., Singh, O., 2007. The biology and control of malaria vectors. Curr Sci 92, 1571-1578.

Echeverry, D.F., Deason, N.A., Davidson, J., Makuru, V., Xiao, H., Niedbalski, J., Kern, M., Russell, T.L., Burkot, T.R., Collins, F.H., Lobo, N.F., 2016. Human malaria diagnosis using a single-

step direct-PCR based on the *Plasmodium* cytochrome oxidase III gene. Malar. J. 15, 016-1185.

Ghanchi, N.K., Ursing, J., Beg, M.A., Veiga, M.I., Jafri, S., Martensson, A., 2011. Prevalence of resistance associated polymorphisms in *Plasmodium falciparum* field isolates from southern Pakistan. Malar. J. 10, 18.

Haanshuus, C.G., Mohn, S.C., Morch, K., Langeland, N., Blomberg, B., Hanevik, K., 2013. A novel, single-amplification PCR targeting mitochondrial genome highly sensitive and specific in diagnosing malaria among returned travellers in Bergen, Norway. Malar. J. 12, 1475-2875.

Howard, N., Durrani, N., Sanda, S., Beshir, K., Hallett, R., Rowland, M., 2011. Clinical trial of extended-dose chloroquine for treatment of resistant falciparum malaria among Afghan refugees in Pakistan. Malar. J. 10, 171.

Huggins, L.G., Koehler, A.V., Ng-Nguyen, D., Wilcox, S., Schunack, B., Inpankaew, T., Traub, R.J., 2019. A novel metabarcoding diagnostic tool to explore protozoan haemoparasite diversity in mammals: a proof-of-concept study using canines from the tropics. Scientific Reports 9, 12644.

Huson, D.H., Bryant, D., 2005. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254-267.

Kakar, Q., Khan, M.A., Bile, K.M., 2010. Malaria control in Pakistan: new tools at hand but challenging epidemiological realities. East Mediterr Health J. 16, S54-60.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28, 1647-1649.

Khattak, A.A., Venkatesan, M., Nadeem, M.F., Satti, H.S., Yaqoob, A., Strauss, K., Khatoon, L., Malik, S.A., Plowe, C.V., 2013. Prevalence and distribution of human *Plasmodium* infection in Pakistan. Malar. J. 12, 1475-2875.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D., 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Appl. Environ. Microbiol. 79, 5112-20.

Lee, P.C., Chong, E.T., Anderios, F., Al Lim, Y., Chew, C.H., Chua, K.H., 2015. Molecular detection of human *Plasmodium* species in Sabah using PlasmoNex multiplex PCR and hydrolysis probes real-time PCR. Malar. J. 14, 015-0542.

Lefterova, M.I., Budvytiene, I., Sandlund, J., Farnert, A., Banaei, N., 2015. Simple Real-Time PCR and Amplicon Sequencing Method for Identification of *Plasmodium* Species in Human Whole Blood. J. Clin. Microbiol. 53, 2251-2257.

Li, J., Gutell, R.R., Damberger, S.H., Wirtz, R.A., Kissinger, J.C., Rogers, M.J., Sattabongkot, J., McCutchan, T.F., 1997. Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite. J. Mol. Biol. 269, 203-213.

Li, J., McConkey, G.A., Rogers, M.J., Waters, A.P., McCutchan, T.R., 1994a. *Plasmodium*: the developmentally regulated ribosome. Exp. Parasitol. 78, 437-441.

Li, J., Wirtz, R.A., McConkey, G.A., Sattabongkot, J., McCutchan, T.F., 1994. Transition of Plasmodium vivax ribosome types corresponds to sporozoite differentiation in the mosquito. Mol. Biochem. Parasitol. 65, 283-289.

Li, M., Xia, Z., Yan, H., 2014. New type of SSUrDNA sequence was detected from both *Plasmodium ovale curtisi* and *Plasmodium ovale wallikeri* samples. Malar. J. 13, 216.

McCutchan, T.F., de la Cruz, V.F., Lal, A.A., Gunderson, J.H., Elwood, H.J., Sogin, M.L., 1988. Primary sequences of two small subunit ribosomal RNA genes from *Plasmodium falciparum*. Mol. Biochem. Parasitol. 28, 63-68.

Moody, A., 2002. Rapid diagnostic tests for malaria parasites. Clin. Microbiol. Rev. 15, 66-78.

Murillo Solano, C., Akinyi Okoth, S., Abdallah, J.F., Pava, Z., Dorado, E., Incardona, S., Huber, C.S., Macedo de Oliveira, A., Bell, D., Udhayakumar, V., Barnwell, J.W., 2015. Deletion of *Plasmodium falciparum* histidine-rich protein 2 (pfhrp2) and histidine-rich protein 3 (pfhrp3) genes in Colombian Parasites. PLoS One. 10, e0131576.

Poostchi, M., Silamut, K., Maude, R.J., Jaeger, S., Thoma, G., 2018. Image analysis and machine learning for detecting malaria. Transl Res. 194, 36-55.

Qari, S.H., Goldman, I.F., Pieniazek, N.J., Collins, W.E., Lal, A.A., 1994. Blood and sporozoite stage-specific small subunit ribosomal RNA-encoding genes of the human malaria parasite *Plasmodium vivax*. Gene. 150, 43-49.

Rachid Viana, G.M., Akinyi Okoth, S., Silva-Flannery, L., Lima Barbosa, D.R., Macedo de Oliveira, A., Goldman, I.F., Morton, L.C., Huber, C., Anez, A., Dantas Machado, R.L., Aranha Camargo, L.M., Costa Negreiros do Valle, S., Marins Povoa, M., Udhayakumar, V., Barnwell, J.W., 2017. Histidine-rich protein 2 (pfhrp2) and pfhrp3 gene deletions in *Plasmodium falciparum* isolates from select sites in Brazil and Bolivia. PLoS One. 12, e0171150.

Rakotonirina, H., Barnadas, C., Raherijafy, R., Andrianantenaina, H., Ratsimbasoa, A., Randrianasolo, L., Jahevitra, M., Andriantsoanirina, V., Menard, D., 2008. Accuracy and reliability of malaria diagnostic techniques for guiding febrile outpatient treatment in malaria-endemic countries. Am. J. Trop. Med. Hyg. 78, 217-221.

Rogers, M.J., McConkey, G.A., Li, J., McCutchan, T.F., 1995. The ribosomal DNA loci in *Plasmodium falciparum* accumulate mutations independently. J. Mol. Biol. 254, 881-891.

Saravu, K., Rishikesh, K., Kamath, A., Shastry, A.B., 2014. Severity in *Plasmodium vivax* malaria claiming global vigilance and exploration-a tertiary care centre-based cohort study. Malar. J. 13, 10.1186.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537-7541.

Scholzen, A., Cooke, B.M., Plebanski, M., 2014. *Plasmodium falciparum* induces Foxp3hi CD4 T cells independent of surface PfEMP1 expression via small soluble parasite components. Front. Microbiol. 5, 200.

Shaukat, A., Ali, Q., Connelley, T., Khan, M.A.U., Saleem, M.A., Evans, M., Rashid, I., Sargison, N.D., Chaudhry, U., 2019. Selective sweep and phylogenetic models for the emergence and spread of pyrimethamine resistance mutations in *Plasmodium vivax*. Infect. Genet. Evol. 68, 221-230.

Sinka, M.E., Bangs, M.J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., Mbogo, C.M., Hemingway, J., Patil, A.P., Temperley, W.H., 2012. A global map of dominant malaria vectors. Parasit. Vectors. 5, 69.

Steenkeste, N., Incardona, S., Chy, S., Duval, L., Ekala, M.T., Lim, P., Hewitt, S., Sochantha, T., Socheat, D., Rogier, C., Mercereau-Puijalon, O., Fandeur, T., Ariey, F., 2009. Towards high-throughput molecular detection of *Plasmodium*: new approaches and molecular markers. Malar. J. 8, 1475-2875.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725-2729.

William, T., Menon, J., Rajahram, G., Chan, L., Ma, G., Donaldson, S., Khoo, S., Frederick, C., Jelip, J., Anstey, N.M., 2011. Severe *Plasmodium knowlesi* malaria in a tertiary care hospital, Sabah, Malaysia. Emerg. Infect. Dis. 17, 1248-1255.

Wongsrichanalai, C., Barcus, M.J., Muth, S., Sutamihardja, A., Wernsdorfer, W.H., 2007. A review of malaria diagnostic tools: microscopy and rapid diagnostic test (RDT). Am. J. Trop. Med. Hyg. 77, 119-127.

Figure Legends

Fig.1. (A) Giemsa-stained blood smears were examined by 1000 x microscopy, showing rings in the *Plasmodium* positive samples. (B) Immunochromatographic assay for the detection of *P. falciparum* specific histidine-rich protein 2 (Pf-HRP2) and *P. vivax* specific lactate dehydrogenase (Pv-LDH). Adequate volumes of the blood samples were dispensed into the sample well 'S' of the test cassette. If Pf-HRP2 was bind to the HRP2 gold conjugates

forming a burgundy colored pf band, indicating *P. falciparum* positive test. If Pv-LDH was bind to the LDH gold conjugates forming a burgundy colored pv band, indicating *P. vivax* positive test. The absence of any band suggests a negative result and C is positive control.

Fig. 2. Schematic representation of the sample preparation (A) and the bioinformatics data handling (B) of the metabarcoded sequencing library. (A) In the first-round PCR amplification, overhanging forward and reverse primers were used to amplify the rDNA 18S. The adapter base pairs provide the target sites for the primers used for sequencing and the random nucleotides (0-3Ns) were inserted between the primers and the adapter to offset the reading frame, therefore amplicons prevent the oversaturation of the MiSeq sequencing channels. The second-round PCR amplification was then performed using overhanging barcoded primers bound to the adapter tags to add indices, as well as the P7 and P5 regions required to bind to the MiSeq flow cell. (B) Text files containing rDNA 18S sequence data (FASTQ files) were generated from the Illumina MiSeq binary raw data outputs, and data analyses were performed using a bespoke modified pipeline in Mothur v1.39.5 software (Schloss et al., 2009) and Illumina MiSeq standard procedures (Kozich et al., 2013) as described in materials and methods section 2.4.

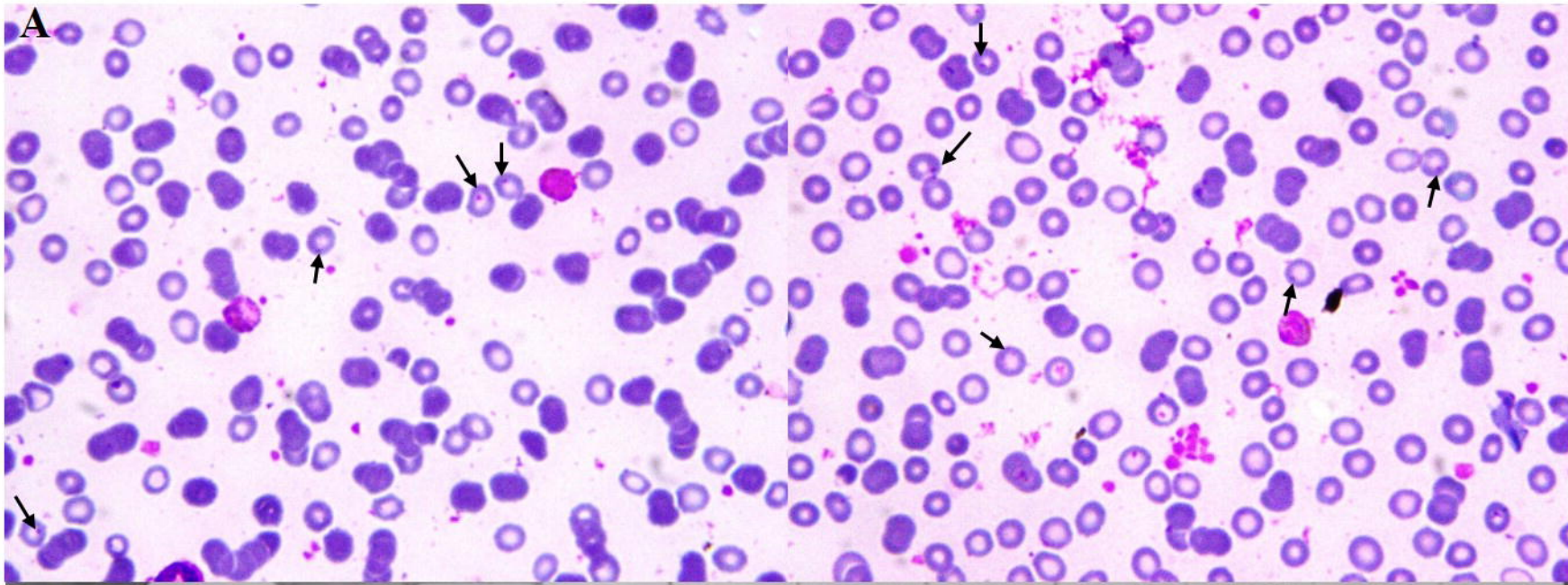
Fig. 3. The maximum-likelihood tree was obtained from the *P. falciparum* and *P. vivax* rDNA 18S region. The sequences were first calculated the number of reference sequences generated from both species (Supplementary Data S1). A total of 34 reference sequences of the rDNA 18S locus were identified in *P. falciparum* and 74 reference sequences were identified in *P. vivax*. The reference sequences were aligned on the MUSCLE tool of the Geneious v9.0.1 software. The neighbor-joining algorithm (HKY parameter model) was computed with 1000 bootstrap replicates using MEGA 7 software. Both species were identified with different color shades (*P. falciparum* in blue and *P. vivax* in brown).

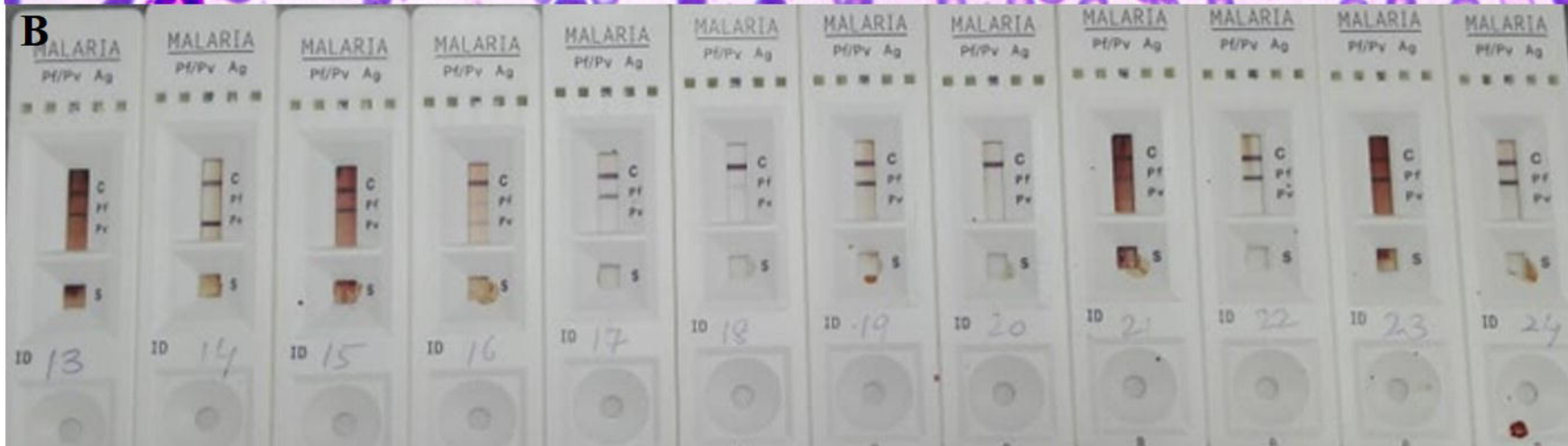
Fig. 4. Validation of the metabarcoding sequencing using three separate mock pools [Mix 1 (*P. falciparum*), Mix 2 (*P. vivax*), Mix 3 (*P. falciparum* and *P. vivax*)] of unknown numbers of parasites from each species. Panel 2A shows that metabarcoded sequencing was used on four replicates of each mock pool to amplify both species as denoted on the X-axis. The Y-axis shows the percentage proportions of each species. Panel 2B shows how the replicates were grouped and averaged based on the amplification. The species are identified with different colours (*P. falciparum* in blue and *P. vivax* in brown).

Fig. 5. Immunochromatographic assay (RDT) and the metabarcoding sequencing (Illumina MiSeq) was performed for the detection of *P. falciparum* and *P. vivax*. A total of 365 malaria-positive samples were collected in EDTA tube from basic health units in the tribal area of the Pakistan-Afghanistan border and Chughtai Diagnostic Laboratory in the Punjab province of Pakistan. The immunochromatographic assay and the metabarcoded 18S rDNA sequencing methods were applied to each sample; the X-axis shows the proportion of each species being estimated and the Y-axis shows the percentage proportions of each species. The species are identified with different colours (*P. falciparum* in pink and *P. vivax* in light blue).

Fig. 6. Split tree was made for the *P. falciparum* and *P. vivax* 18S rDNA sequence data. 134 genotypes were identified in *P. falciparum* and 48 genotypes were identified in *P. vivax* (Supplementary Data S2). The genotypes were aligned on the MUSCLE tool of the Geneious v9.0.1 and the tree was constructed with the UPGMA method in the Jukes-Cantor model of substitution in the SplitsTrees4 software. The appropriate model of nucleotide substitutions was selected by using the jModeltest 13.1.0 program. The pie chart circles in the tree

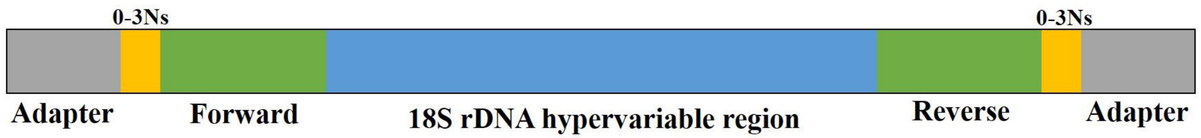
represent the different 18S genotypes containing different colours as follows: (A) *P. falciparum* from the field samples are coloured pink (type A group 1 and 2) and NCBI database sequences are coloured blue (type A group 1 and type S). (B) *P. vivax* from the field samples are coloured light blue (type A group 1, 2 and 3) and NCBI database sequences are coloured brown (type A group 3 and type S).


A

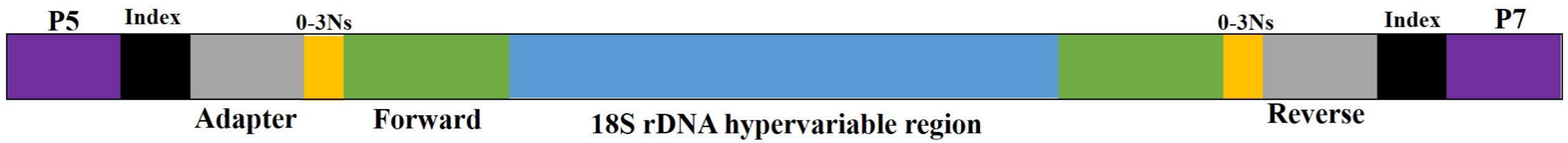

B
MALARIA
Pf/Pv Ag
C
Pf
Pv
S
ID 13
ID 14
ID 15
ID 16
ID 17
ID 18
ID 19
ID 20
ID 21
ID 22
ID 23
ID 24

(A) Sample preparation

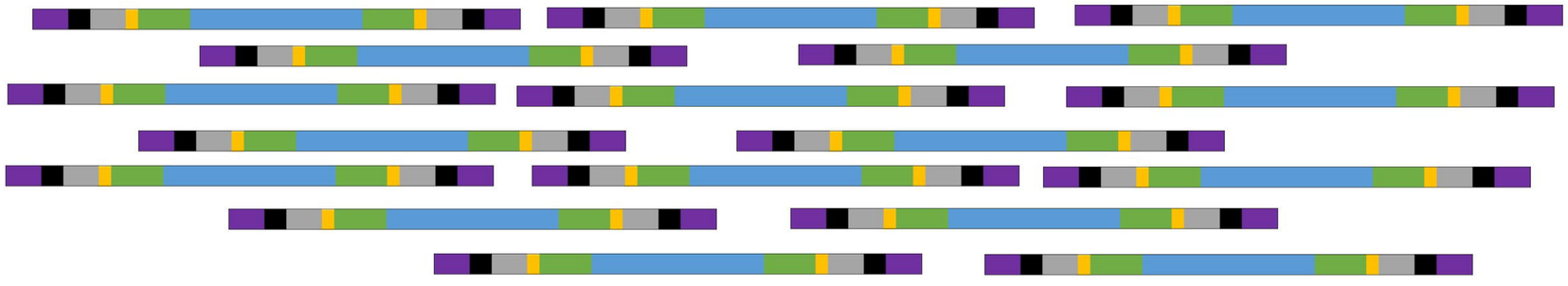
1. Adapter PCR



2. Barcoded PCR



3. Pooling of all samples



4. Sequencing each sample on Illumina MiSeq platform

(B) Bioinformatics data handling

Step 1

Raw paired-ends reads were analysed (make.contigs' command)

Remove any ambiguous bases (screen.seqs command)

Align dataset with a rDNA 18S reference sequence library (align.seqs command)

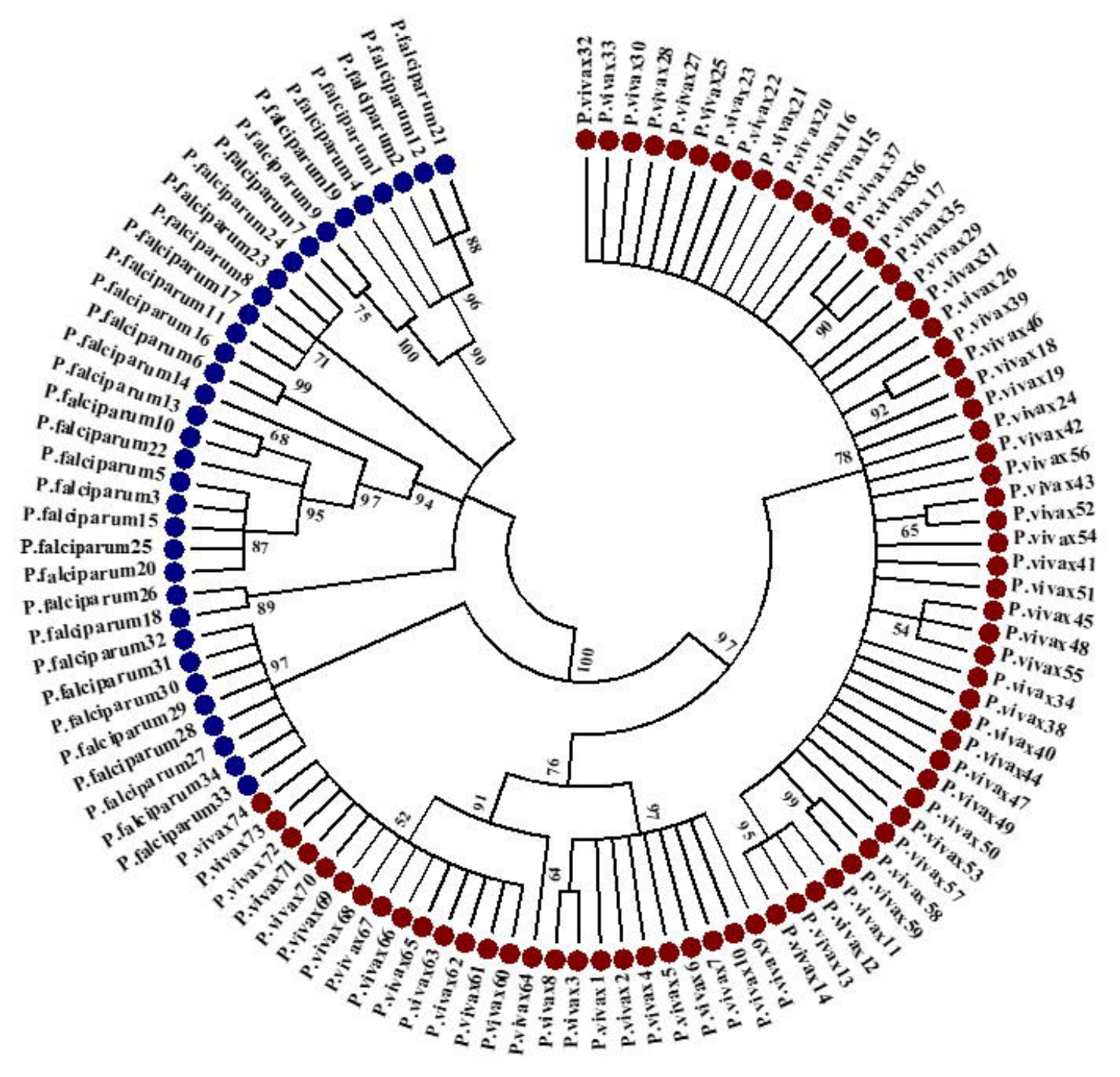
Confirm the filtered sequences overlap the same region of the rDNA 18S reference sequence library (screen.seqs command)

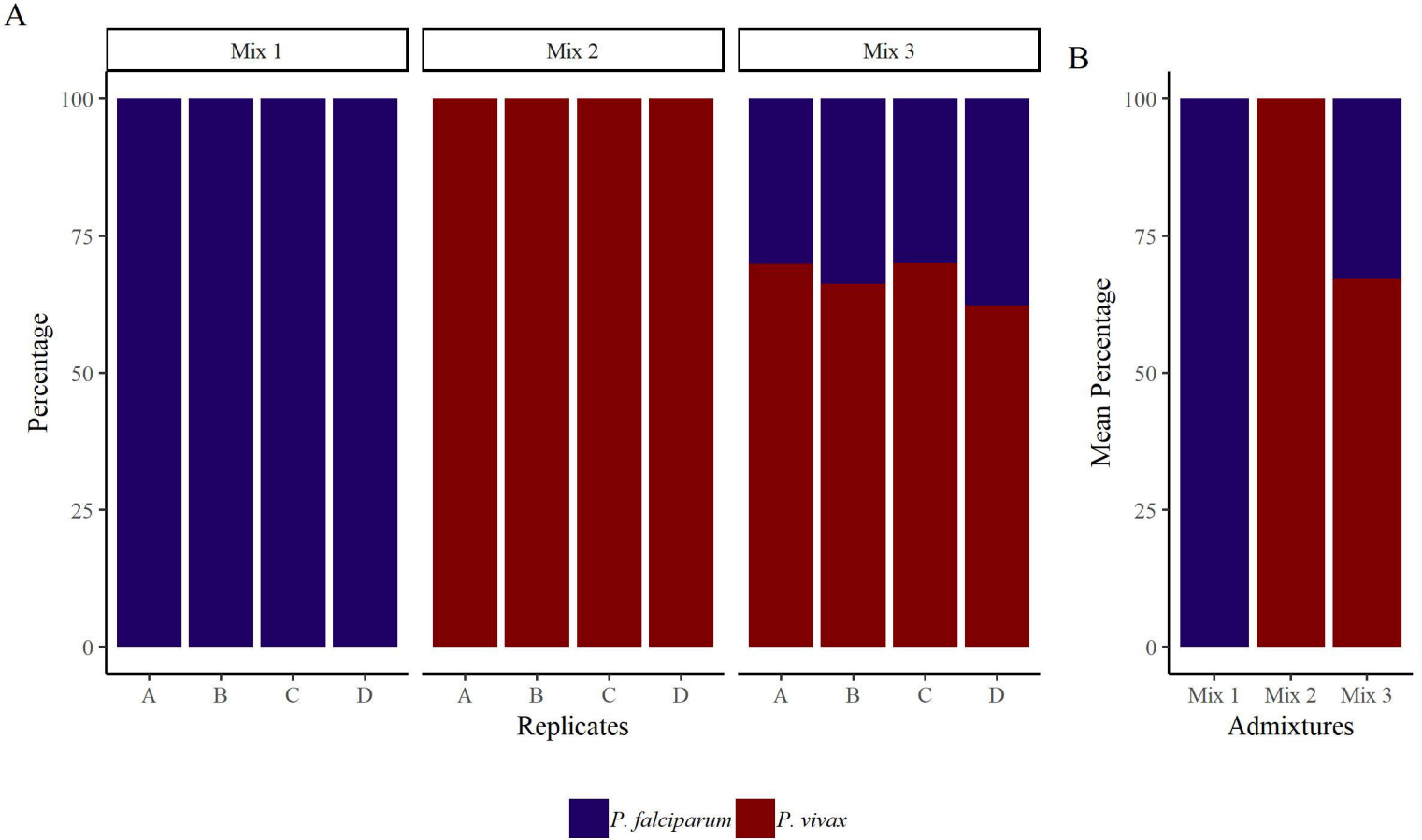
Classify the sequences into either of the two species (*P. falciparum* and *P. vivax*) (classify.seqs command)

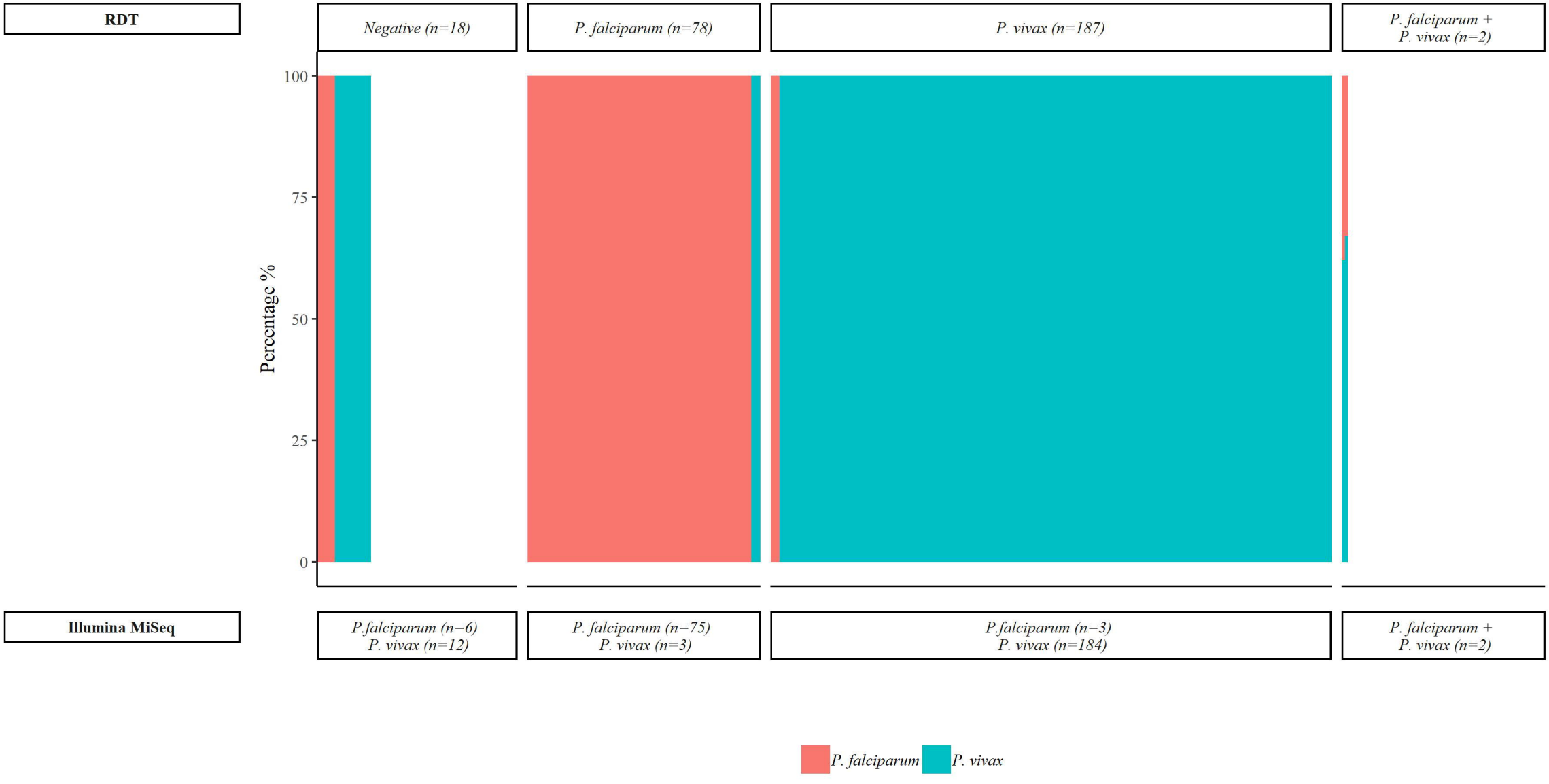
Step 2

Count list of the consensus sequences for each parasite samples based on the total numbers of filtered sequences (unique.seqs command)

Count list was used to create the FASTQ files of the consensus sequences of each sample (split.groups command)


P.falciparum21
P.falciparum12
P.falciparum2
P.falciparum1
P.falciparum4
P.falciparum19
P.falciparum9
P.falciparum7
P.falciparum24
P.falciparum23
P.falciparum8
P.falciparum17
P.falciparum11
P.falciparum16
P.falciparum6
P.falciparum14
P.falciparum13
P.falciparum10
P.falciparum22
P.falciparum5
P.falciparum3
P.falciparum15
P.falciparum25
P.falciparum20
P.falciparum26
P.falciparum18
P.falciparum32
P.falciparum31
P.falciparum30
P.falciparum29
P.falciparum28
P.falciparum27
P.falciparum34
P.falciparum33
P.vivax74
P.vivax73
P.vivax72
P.vivax71
P.vivax70
P.vivax69
P.vivax68
P.vivax67
P.vivax66
P.vivax65
P.vivax63
P.vivax62
P.vivax61
P.vivax60
P.vivax64
P.vivax8
P.vivax3
P.vivax1
P.vivax2
P.vivax4
P.vivax5
P.vivax6
P.vivax7
P.vivax10
P.vivax9
P.vivax14
P.vivax13
P.vivax12
P.vivax11
P.vivax59
P.vivax58
P.vivax57
P.vivax53
P.vivax50
P.vivax49
P.vivax47
P.vivax44
P.vivax40
P.vivax38
P.vivax34
P.vivax55
P.vivax48
P.vivax45
P.vivax51
P.vivax41
P.vivax54
P.vivax52
P.vivax43
P.vivax56
P.vivax42
P.vivax24
P.vivax19
P.vivax18
P.vivax46
P.vivax39
P.vivax26
P.vivax31
P.vivax29
P.vivax35
P.vivax17
P.vivax36
P.vivax37
P.vivax15
P.vivax16
P.vivax20
P.vivax21
P.vivax22
P.vivax23
P.vivax25
P.vivax27
P.vivax28
P.vivax30
P.vivax33
P.vivax32
88
96
90
75
100
71
99
68
94
97
95
87
89
97
100
97
76
91
52
64
97
95
99
78
92
90
65
54

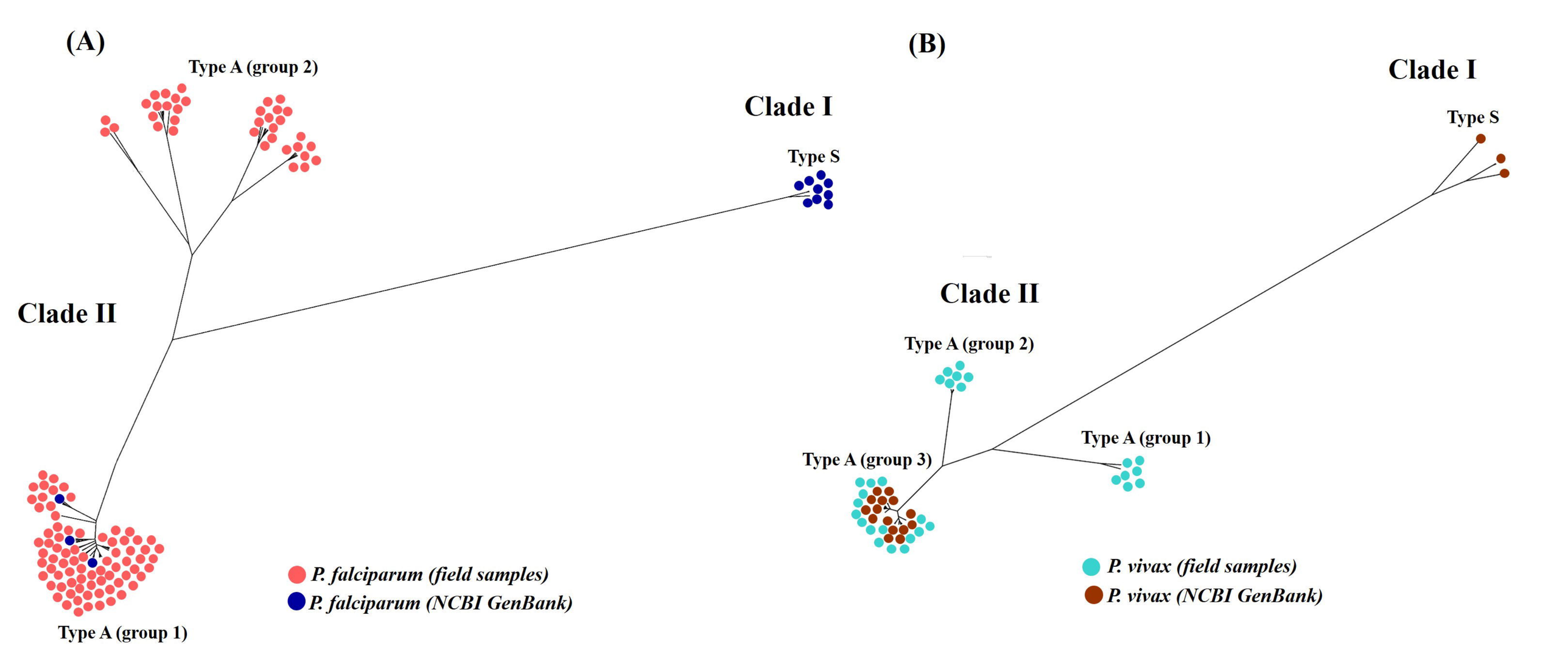

A
Mix 1
Mix 2
Mix 3
100
75
50
25
0
Percentage
A
B
C
D
Replicates
B
Mean Percentage
Mix 1
Mix 2
Mix 3
Admixtures
P. falciparum
P. vivax



RDT performance, using Illumina Miseq as a gold standard:

.	Predictive Value (95% CI)	Category-Specific Classification Probability (95% CI)
RDT Negative	0.816 (0.724 - 0.883)	1*
RDT P. falciparum	0.962 (0.860 - 0.990)	0.893 (0.807 - 0.943)
RDT P. vivax	0.984(0.938 - 0.996)	0.924 (0.881 - 0.953)
RDT P. falciparum + P. vivax	1*	1*

* It was not possible to calculate confidence intervals for these values due to the absence of false classifications in this dataset.


(A)
Type A (group 2)
Clade I
Type S
Clade II
Type A (group 1)
P. falciparum (field samples)
P. falciparum (NCBI GenBank)
(B)
Clade I
Type S
Clade II
Type A (group 2)
Type A (group 1)
Type A (group 3)
P. vivax (field samples)
P. vivax (NCBI GenBank)