Amplicon deep sequencing of low-density *Plasmodium falciparum* infections: an evaluation of analysis approaches

Angela M. Early[1,2]*, Rachel F. Daniels[1,2], Timothy M. Farrell[1,2], Sarah K. Volkman[1,2,3], Dyann F. Wirth[1,2], Bronwyn L. MacInnis[1,2], Daniel E. Neafsey[1,2]

1. Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, 02142 USA
2. Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, 02115 USA
3. Simmons College, School of Nursing and Health Sciences, Boston, MA, 02115 USA

* Corresponding author: early@broadinstitute.org (AME)

Keywords:
Targeted amplicon deep sequencing, Haplotype calling, Multiplicity of infection, Multiclonal infection, Within-host diversity, Molecular epidemiology, Molecular surveillance, Malaria, Plasmodium

Abstract

Background: Deep sequencing of targeted genomic regions is becoming a common tool for understanding the dynamics and complexity of *Plasmodium* infections. Here, Illumina-based amplicon sequencing of two *P. falciparum* genomic regions (*CSP* and *SERA2*) was performed on two types of samples: *in vitro* DNA mixtures mimicking low-density infections (1-200 genomes/µl) and natural patient samples (44-653,080 parasites/µl). The analytical performance of four analysis tools—PASEC, DADA2, HaplotypR, and SeekDeep—was compared on both datasets.

Results: All four analysis tools were able to contend with mock low-density samples, showing reasonable detection accuracy down to a concentration of 5 *Plasmodium* genome copies/µl. Due to increased stochasticity and background noise, however, accuracy was reduced for samples with very low parasitemia (< 5 copies/µl) or very low read count (<100 reads per amplicon). PASEC could distinguish major *vs.* minor haplotypes with an accuracy of 90% in samples with at least 30 *Plasmodium* genome copies/µl, but only 61% at low *Plasmodium* concentrations (< 5 copies/µl) and 46% at low read counts (<25 reads per amplicon). The four tools were additionally compared on a panel of patient samples, and all four provided concordant complexity of infection patterns across four sub-Saharan African countries.

Conclusions: Amplicon deep sequencing successfully determines the complexity and diversity of low-density *Plasmodium* infections, even in the absence of technical PCR/sequencing replicates. Current state-of-the-art tools offer multiple robust approaches for analyzing amplicon data. However, as samples with very low parasitemia and very low read count have higher false positive rates, researchers should consider implementing higher read count thresholds when working with low-density samples.

Background

Amplicon deep sequencing is increasingly displacing other genotyping approaches as it provides a cost-effective approach to profiling the genetic diversity of pathogen infections. Like SNP-based genotyping approaches, both the data-generation and data-analysis steps of amplicon sequencing are highly scalable, allowing for studies of hundreds to thousands of samples. Amplicons, however, can be designed to cover long genetic segments composed of multiple variants. When targeted to a highly polymorphic genomic region, a single amplicon can therefore distinguish among hundreds of unique DNA sequences (haplotypes) [1], which provides higher resolution than either SNP-based or length-based genotyping approaches when estimating the number of lineages within an infection (or complexity of infection; COI) [2–4]. Increasing its ease of use, amplicon analysis in *Plasmodium* has been adapted to multiple sequencing platforms depending on the desired cost, sample size, and sequence length [3, 5–7]. Because of this high resolution and flexibility, amplicon-based methods have been utilized in a range of applications, including studies of allele-specific vaccine efficacy [1], disease severity [6], clearance rates [8], within-host competition [9], relapse rates [5], drug resistance [10–12], host selection [13], and population structure [13, 14]. Amplicon sequencing has high sensitivity for the detection of minority clones and is of particular interest in longitudinal studies that track intra-host dynamics [3, 4].

When used to detect known single variant markers, amplicon sequence data can be analyzed with relatively straight-forward approaches. Often, however, amplicon sequencing is used to target larger, more complex haplotypes. In addition to providing higher clonal resolution, longer haplotypes permit the discovery of unknown alleles, for instance when monitoring variability in resistance-associated genes [10–12], and provide increased information for haplotype-based analyses of epistasis and linkage disequilibrium [13].

Being more information-rich, these longer haplotypes require more sophisticated analysis methods than single-variant detection. Amplicon sequencing data is known to be subject to PCR and sequencing artifacts, particularly for genomic regions with high A/T-content and high rates of homopolymerism [15, 16]. In addition, library preparation method and primer choice can influence the types and extent of errors [17]. Correctly identifying sequence errors is therefore a challenge when applying amplicon sequencing to *P. falciparum*. Fortunately, several new tools for analysis of amplicon data have been developed in recent years [18–21]. Unlike approaches that use reference datasets or cluster sequences with hard percent-identity thresholds, these new methods are more flexible and can distinguish among sequences that differ by only a single nucleotide change [22]. Questions still remain, however, regarding the relative accuracy of these different approaches, their applicability to samples of low parasitemia, and their capacity to recover quantitative information regarding the relative abundance of different haplotypes within patient infections.

In this study, amplicon sequencing of densely polymorphic regions in the *P. falciparum CSP* and *SERA2* genes was applied to two sample collections. The first—a

set of *in vitro* human/parasite DNA mixtures that mimic low density parasite infections—was designed to test the limit of detection for amplicon sequencing. The second sample set consisted of DNA extracted from dried blood spots collected on filter paper in sub-Saharan Africa, capturing the conditions under which samples are typically collected and processed. Initial analysis of the data resulting from these samples was conducted with the Parallel Amplicon Sequencing Error Correction (PASEC) pipeline, a new distance and abundance-based error-correction tool that has been carefully tuned for use with these two amplicons. The performance of PASEC was then compared to that of three previously published tools: DADA2 [18], HaplotypR [19], and SeekDeep [20]. All four tools detected *P. falciparum* haplotypes with high sensitivity, and additionally were able to discriminate between major and minor haplotypes with reasonable accuracy. Overall, the results show that low parasitemia does not impede amplicon analysis of *P. falciparum* samples, although researchers should expect lower sensitivity and lower precision with low read-count samples (<100 reads/amplicon) and at parasite levels under 5 genomes/µl.

Methods

Sample assembly and composition

Mock *Plasmodium*/human DNA mixtures:

Mixtures of cultured *P. falciparum* parasite and human genomic DNA were constructed to mimic human patient infection samples. Up to five culture-adapted parasite strains were combined in various ratios and number (Figure 1; exact sample composition is in Additional File 1, Table S1). Stock mixtures of 200 copies/µl total were prepared by real-time PCR quantification of copies/µl in triplicate relative to a plasmid containing a single copy of the quantification target gene [23]. These stock solutions were then diluted to the indicated concentrations in sequencing-grade water and 10ng commercial human DNA (Promega Corp cat#G3041) was added to all samples. After mixing and dilution, a subset of samples were re-quantified using the same qPCR protocol and reported sample concentrations were adjusted as needed. *Plasmodium*-free negative control samples were also constructed. These contained either 10ng of human DNA or only water.

Patient samples: Previously extracted DNA from 95 patient samples was re-amplified and re-sequenced as part of this study. These samples were acquired from four countries in sub-Saharan Africa as part of the RTS,S malaria vaccine Phase 3 trial and had parasite densities that ranged from 44-653,080 parasites/µl as determined by blood smear (Figure 1; [24]). Full details on sampling and extraction are provided in Neafsey *et al.*, 2015 [1]. In brief, samples were collected as blood spots on Whatman FTA cards, shipped to the Broad Institute, and stored in desiccators until processing. DNA was extracted in batches of 95 samples plus one blank control card using the automated Chemagen Chemagic bead-based extraction platform. Total DNA was stored at -80°C until re-amplification and sequencing.

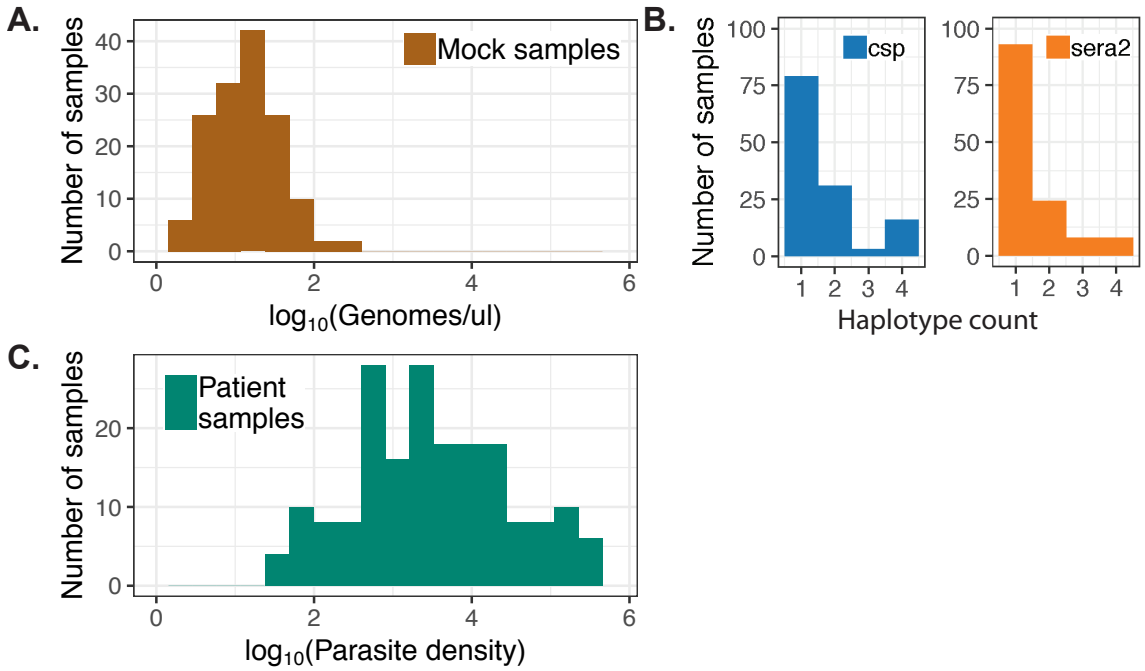

Figure 1. Mock and patient sample composition. (A) Mock samples were constructed from mixtures of *P. falciparum* and human DNA to mimic the parasite DNA concentrations found in extracted low-density infections. (B) DNA from a total of five clonal lab lines were combined in the mock samples, leading to COI values of one to four. (C) Patient samples were previously collected and extracted from a combination of symptomatic patients and asymptomatic carriers. Parasite densities were determined by blood smear.

Positive control plasmid: A plasmid containing synthetic target amplicon sequences for both *CSP* and *SERA2* was obtained from a commercial vendor (Invitrogen/Therm Fisher Scientific) and served as a positive control during the PCR amplification step. Outside the primer regions, the plasmid sequence contains nucleotide variants not observed in natural *P. falciparum* isolates so that any instances of contamination can be readily identified. The plasmid map can be found in Additional File 1, Figure S1.

PCR and sequencing

Two regions from the genes *CSP* (PF3D7_0304600) and *SERA2* (PF3D7_0207900) were PCR amplified as previously described [1]. In brief, targeted regions were amplified then indexed in two separate rounds of PCR. The final *CSP* and *SERA2* amplicons cover 288 and 258 nucleotides respectively (Pf3D7_03_v3:221,352-221,639; Pf3D7_02_v3:320,763-321,020). Both amplicons overlap sequence regions of high nucleotide diversity in sub-Saharan Africa, maximizing the number of distinct haplotypes that can be detected across samples in this geographic area.

All DNA samples and negative controls were amplified and sequenced in duplicate. Paired-end 250-bp reads were generated in one MiSeq run conducted on a pool of 384 PCR products. Unless otherwise noted, each PCR/sequencing replicate

was analyzed as a distinct sample. Before downstream analysis, raw sequencing data were demultiplexed and aligned to amplicon reference sequences to remove all non-*Plasmodium* sequences.

Sample analysis with the PASEC pipeline

For each sample, paired-end reads were merged using FLASH [25] and aligned with BWA-MEM v0.7.12-r1039 [26] to the amplicon regions of the *P. falciparum* reference genome assembly (PlasmoDB v.9.0 3D7). Two short homopolymeric tracts in *CSP* were masked from analysis as such regions are highly error-prone in Illumina sequencing and these tracts were not known to harbor natural polymorphisms. Masked coordinates are given in Additional File 3.

Within each sample, haplotypes were filtered according to a set of pre-specified thresholds developed by Neafsey *et al* [1]. Haplotypes were required to (1) cover the entire amplicon region, (2) have no uncalled bases, (3) be supported by at least two sets of merged read pairs, and (4) have an intra-sample frequency ≥ 0.01. To account for potential PCR and sequencing errors, the filtered haplotypes were clustered based on nucleotide distance and read depth. If two haplotypes within the same sample differed by only one nucleotide and had a read coverage ratio ≥8:1, they were merged, maintaining the identity of the more common haplotype. Previous implementations of this pipeline removed all potential chimeric reads and required samples to contain at least 200 reads for one of the two amplicons [1, 13]. In this analysis, these metrics were analyzed, but hard filters were not applied to the samples before downstream analysis.

Full details on the PASEC pipeline, its customizable parameters, and its implementation in this study are found in Additional Files 2 and 3 and at https://github.com/tmfarrell/pasec.

Sample analysis with DADA2, HaplotypR, and SeekDeep

All samples were also independently analyzed using three additional amplicon analysis tools: DADA2 [18], HaplotypR [19], and SeekDeep [20]. Beyond the changes detailed below, input parameters deviated only modestly from the default settings. Parameters and scripts used for executing each pipeline can be found in Additional File 3. While previous implementations of PASEC applied a 200 reads/sample threshold, no read count filters were applied at the sample level in this analysis comparison.

SeekDeep gives the option of grouping data from technical PCR/sequencing replicates of the same sample and applying clustering and filtering to this grouped data to increase confidence in final calls. We therefore ran the pipeline under two conditions: grouping technical replicates (the recommended, default SeekDeep approach; "SeekDeep2x") and treating each PCR/sequencing replicate independently ("SeekDeep1x"). This permitted a more level comparison with pipelines that do not incorporate replicate information and allowed for a determination of whether a single replicate is sufficient for making accurate haplotype calls.

For HaplotypR, the command-line interface was extended in two ways. First, it was altered to return full haplotype sequences as opposed to only bases at variant

positions. Second, the trimming input command was expanded to allow each amplicon to have different lengths. The version of HaplotypR used in this analysis can be found at https://github.com/tmfarrell/HaplotypR. After running the pipeline, the authors' recommended sample-level filtering was applied to the data. Specifically, each sample was required to have a minimum of 25 reads, and individual haplotypes needed to have a minimum of 3 reads and a within-host frequency of at least 0.1%.

Comparison of analysis tools

All four tools were assessed for their ability to resolve haplotypes at within-sample frequencies down to 1% using the mock low-parasitemia samples. Two performance metrics were computed by comparing expected vs. observed haplotypes in each sample: sensitivity (proportion of all expected haplotypes that were observed) and precision (proportion of all observed haplotypes that were expected). For sensitivity calculations, only haplotypes present at a concentration of at least 1 copy/µl were considered. For each tool, samples were only included in the performance metric calculation if at least one haplotype was identified. Except for the SeekDeep2x implementation, each PCR/sequencing replicate was analyzed as a distinct sample.

Results

Sequencing coverage for low-density mock infections and patient samples from sub-Saharan Africa

In total, 148 DNA mixtures of known haplotypic composition, 190 natural infections from sub-Saharan Africa, 12 positive-control plasmid samples, and 4 negative-control samples without *Plasmodium* DNA were sequenced on a single Illumina MiSeq run.

The 148 mock DNA mixtures were constructed to mimic infections with low parasite density and contained between 1 and 200 *P. falciparum* genomes/µl (Figure 1A). After sequencing, 145 samples had full-length read coverage for at least one of the two amplicons. For each amplicon, initial raw coverage across these samples varied from 0 to 280,876. After implementing the PASEC pipeline, coverage ranged from 0 to 31,787 reads. Coverage was sufficient for both amplicons, although median coverage was higher for *CSP* than for *SERA2* (1872 vs. 909; Figure 2A). All samples with very low coverage (<100 reads) had *Plasmodium* DNA concentrations below 21 genomes/µl. Overall, however, read count and genome copy number were only weakly correlated (Spearman's $\rho = 0.55$, $P = 9.3\text{x}10^{-14}$; Figure 2B).

Sequence coverage was higher for the 190 patient samples from sub-Saharan Africa (Figure 2C). These samples were extracted from dried blood spots and had parasite densities that ranged from 44-653,080 parasites/µl as determined by microscopy of blood smears. As with the *in vitro* DNA mixtures, coverage was generally higher for samples with higher parasite loads, but this correlation was low (Spearman's $\rho = 0.31$, $P = 1.1\text{x}10^{-9}$; Figure 2D). Overall sequencing success was lower for the patient samples than for the mock DNA mixtures (Figure 2C). It is

likely that this partially resulted from difficulties with extracting clean, high quality DNA from the stored filter paper blood spots rather than variability in the sequencing itself. In total, 22 patient samples had at least one PCR/sequencing replicate with 0 or low read counts (<100). Of these, 18 experienced failure with both technical PCR/sequencing replicates.

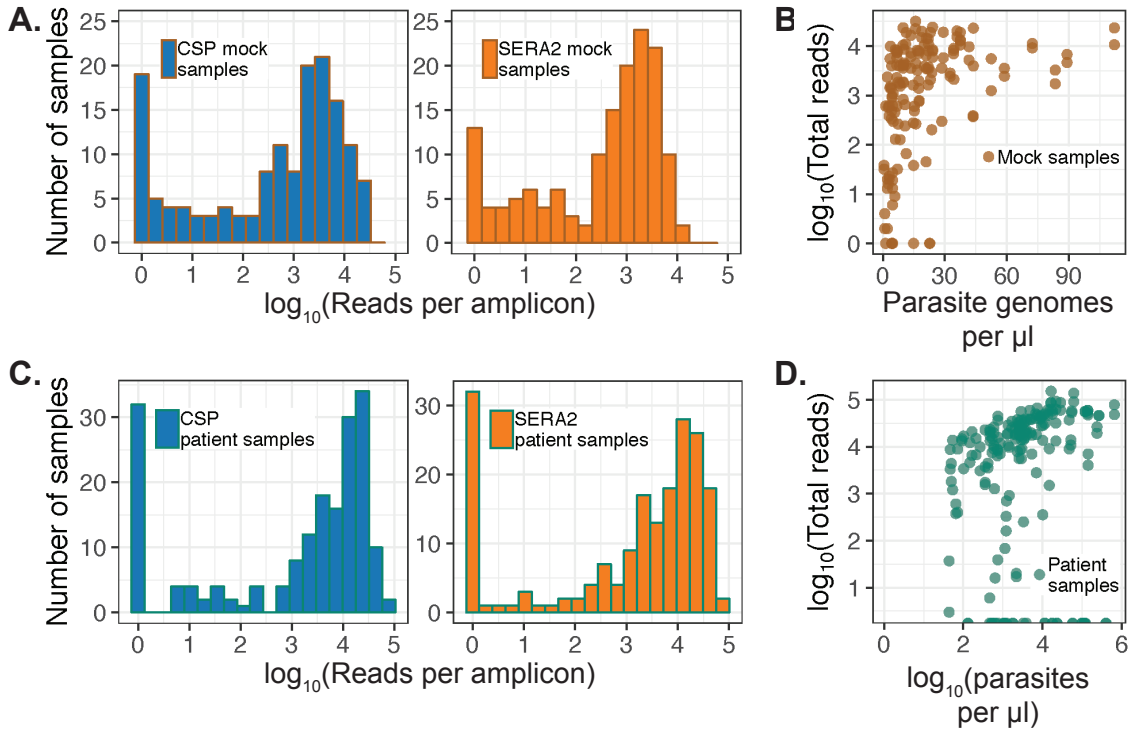

Figure 2. Sequencing coverage of mock and patient samples. Overall sequencing coverage was lower for mock (A) than patient (C) samples although patient samples had a higher failure rate. For both mock (B) and patient (D) samples, total read coverage (reads combined from both amplicons) correlated weakly with parasite genome concentration or parasitemia.

Absolute haplotype concentration affects the probability of sequencing success

Each mock sample contained between one and four unique haplotypes at the *CSP* and *SERA2* amplicons present at concentrations of 1-200 copies/µl (Figure 1B). Overall, there was a high recovery of these expected haplotypes from each of the samples. PASEC correctly identified all haplotypes present at a concentration of 30 copies/µl or higher and 96% of haplotypes with concentrations over 20 copies/µl. Conversely, only 41% of haplotypes with 1-5 copies/µl were recovered (Figure 3A). As discussed in the tool comparison below, this haplotype sensitivity is only slightly

influenced by the post-sequencing analysis method and is instead driven by a failure to initially amplify and/or sequence these low frequency haplotypes.

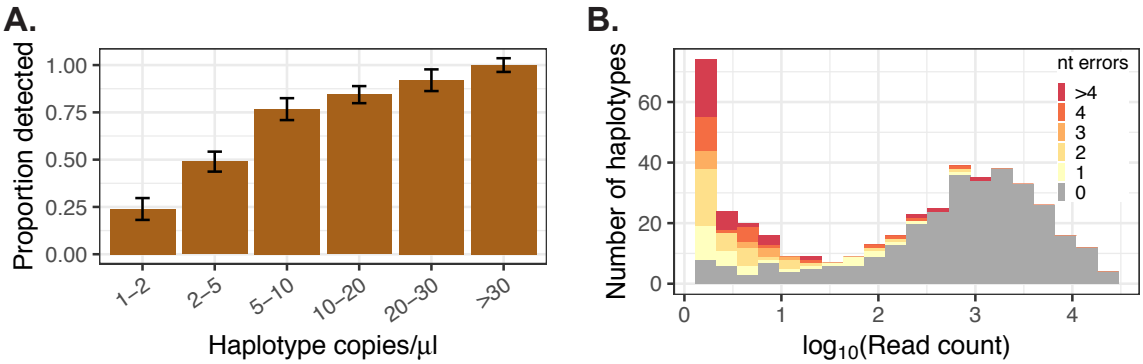

Figure 3. Correct identification of mock sample haplotypes. (A) Detection of known haplotypes within the mock samples was dependent on haplotype concentration within the sample. Error bars represent the binomial-estimated standard deviation. (B) 31% of identified haplotypes were erroneous, but these haplotypes were generally supported by fewer reads than correct haplotypes.

Amplicon sequencing retains some information on within-sample haplotype frequencies, even at low concentrations

When performing direct short-read sequencing, read depth can be used to infer sample features like genotype ratios or genome copy number variations. However, during construction of amplicon libraries, PCR amplification prior to sequencing can introduce stochastic variation in the final read counts. Nevertheless, analysis of the final read ratios in the mock samples shows that some information on the original haplotype ratios can be recovered. For amplicons with at least 100 reads, the correlation between the haplotypic ratio in the template DNA and final read ratio was moderate (Pearson's r = 0.82, $P < 0.001$, Additional File 1, Figure S2). As a result, in 73% of samples with at least a 4% margin between the two most prevalent haplotypes, read ratio correctly identified the most prevalent haplotype in the starting DNA mixture. Again, low sample read count reduced the probability of identifying the correct major haplotype (Figure 4A). Similarly, major haplotype identification was less accurate in samples with very low total *Plasmodium* DNA concentration (<5 genomes/µl; Figure 4B).

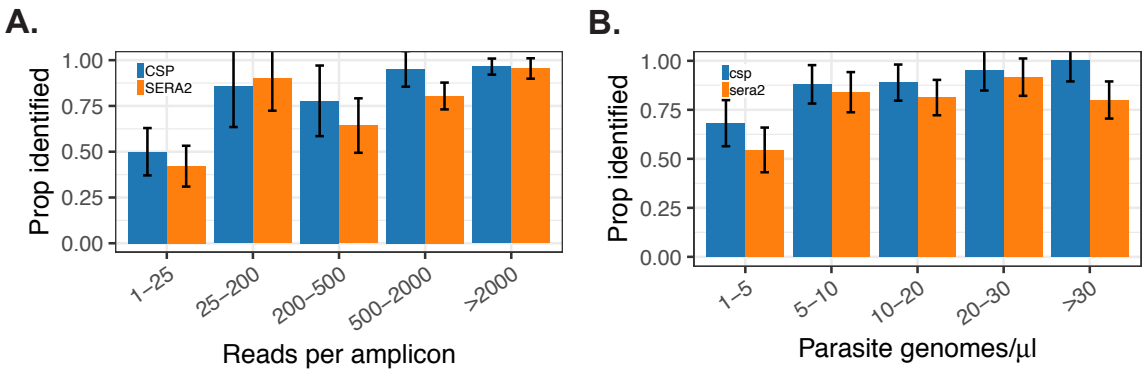

Figure 4. Proportion of mock samples where the major haplotype was correctly identified. Identification of the major haplotype was less reliable at low read counts (A) and low parasite genome concentrations (B). Samples were excluded from the analysis if the difference in prevalence between the top two haplotypes was less than 4%. Error bars represent the binomial-estimated standard deviation.

Erroneous haplotypes in the mock samples have lower read support than correct haplotypes

Results show that read support is a useful indicator of the likelihood that a called haplotype is correct. In keeping with past observations, haplotypes with single-read support were largely sequencing artifacts, with only 0.030% matching a haplotype sequence known to be present in the sample mixtures. The default PASEC pipeline therefore requires haplotypes to have read support ≥2, a filter that eliminated 89.0% of *CSP* and 85.8% of *SERA2* haplotypes from the dataset.

After filtration with the full PASEC pipeline, some erroneous haplotypes remained, but they continued to show lower read support than true haplotypes (Figure 3B). In the final filtered dataset, 31% of the identified haplotypes were erroneous, although combined these haplotypes only accounted for 0.75% of the total reads. Of note, the same percentage of erroneous reads (0.8%) was previously reported by Hathaway *et al* on a different dataset analyzed with their tool SeekDeep [20]. Reads supporting erroneous haplotypes were more prevalent in samples with low read depth and low parasite concentration (Additional File 1, Figure S3). In order to decrease the false positive rate, users could therefore choose to increase the read support threshold per haplotype or the minimum read depth per sample. Striving to completely eliminate false positives, however, would decrease sensitivity, especially for low-frequency haplotypes. For instance, 41% of samples contained at least one erroneous haplotype for one of the two amplicons. In 42% of these cases, the most common erroneous haplotype contained higher read support than the least prevalent true haplotype within the sample.

Frequency and source of haplotype errors in the mock samples

The PASEC pipeline contains customized filtration and error-correction steps to remove erroneous *CSP* and *SERA2* haplotypes. The filtration and error-correction steps in PASEC were designed to address three main sources of erroneous

haplotypes: sequencing errors, chimeric reads, and sample contamination. The frequency of these error types and the efficacy of the various PASEC filters are discussed in more detail below.

Nucleotide sequence errors: The majority of erroneous haplotypes are expected to result from sequence errors (nucleotide substitutions or indels) that occur during Illumina sequencing or the initial rounds of PCR. The PASEC pipeline accounted for these errors with two approaches: (1) hard masking of error-prone sequence regions and (2) clustering of haplotypes that differed by a single nucleotide and had a read coverage ratio ≥8:1. Hard masking was applied to two homopolymeric regions in *CSP* composed of 9 and 6 poly-Ts. In the raw data, erroneous indels within these two regions were detected in 5.7% and 1.2% of full-length reads. While true indels might occur in these sequences in natural populations, this high artifactual indel rate suggests that inference of variants in these regions would be too unreliable. Compared to masking, the clustering of haplotypes had an even greater impact on reducing nucleotide errors: 57.0% of *CSP* haplotypes and 47.9% of *SERA2* haplotypes were eliminated at this step.

In the final filtered dataset, approximately half of the erroneous haplotypes (51%) differed from a true haplotype by one or two nucleotide changes and were likely the result of Illumina sequencing or PCR errors. As discussed above, these haplotypes were supported by fewer reads than true haplotypes (Figure 3B).

Chimeric reads: Chimeric reads are false recombinant haplotypes generated during PCR amplification. While a necessary consideration when performing amplicon sequencing, their overall impact on the mock sample analysis was minimal. Potential chimeras were identified with the isBimera function in DADA2 [18], which identifies all haplotypes that could be constructed from a simple combination of two other haplotypes within the same sample. This analysis flagged 7 *CSP* and 16 *SERA2* samples as containing a total of 36 chimeric haplotypes. Eleven (31%) of the flagged haplotypes were in fact true haplotypes known to be within the given sample. Further analysis showed that 20 of the 25 flagged erroneous haplotypes were only one nucleotide change away from another haplotype in the sample, and the remaining five were related by two nucleotide changes. This suggests that some of these haplotypes may have resulted from PCR or sequencing error instead of chimeric read formation. Eighteen (78%) of the flagged samples had total read counts under 200, the read threshold previously used with the PASEC pipeline [1]. The increased stochasticity associated with low-read samples may explain why these haplotypes were not merged as part of the PASEC sequencing error filter.

Correctly identifying chimeric reads in patient samples presents an additional challenge, especially in regions of high malaria prevalence where recombination among haplotypes is expected to be common. Of the 50 most common *CSP* sequences detected in sub-Saharan Africa [13], 38 (76%) were flagged as chimeric combinations by DADA2. Researchers must therefore consider additional factors like population-level haplotype frequency when identifying chimeric reads in patient samples [19, 20].

Cross-sample or environmental contamination: A large percentage (49%) of erroneous haplotypes had no evidence of chimerism and were unlikely to have resulted from sequencing errors, as they were ≥3 nucleotide changes away from any true haplotype within a given sample. 68% of these haplotypes were present in other samples from the same MiSeq run, suggesting cross-sample or environmental contamination. The remaining haplotypes occurred only once in the whole dataset and may have resulted from environmental contamination. A small amount of cross-sample or environmental contamination was also observed in the negative control samples that contained either water (N=2) or human DNA (N=2). These four *Plasmodium*-free samples contained 5, 7, 16, and 20 reads, respectively. All of these read counts fell well below the 200-read quality threshold previously used with the PASEC pipeline [1].

Comparison of PASEC with three state-of-the-art amplicon analysis tools

The performance of PASEC—a pipeline that has been carefully tuned for use with the *CSP* and *SERA2* amplicons in *P. falciparum*—was compared to that of three analytical tools that were developed to be applied to amplicons from any genomic region: DADA2 [18], HaplotypR [19], and SeekDeep [20]. All four of these tools were designed to detect low-frequency haplotypes and differentiate unique haplotypes with single-nucleotide resolution. There are, however, differences in the analytical approaches. For instance during error filtration, PASEC and HaplotypR rely mainly on variant frequency and read depth, while SeekDeep incorporates k-mer frequencies and base quality scores and DADA2 further models sequencer-specific error likelihoods.

While all these tools have undergone rigorous testing, no previous study has focused on their performance under extremely low parasite densities. Here, each tool was applied to the mock samples and it was evaluated on (1) the proportion of all expected haplotypes that were observed (sensitivity) and (2) the proportion of observed haplotypes that were expected (precision).

Sensitivity and precision: Overall, the four tools performed comparably well on the mock sample panel, although they showed more variability in precision than in sensitivity (Figure 5). This suggests that what differs most between pipelines, is their ability to filter out erroneous haplotypes, not identify correct haplotypes. For instance, while the sensitivity of SeekDeep1x—the SeekDeep implementation using only one technical replicate— was comparable to the other four pipelines, its precision was substantially lower, driven by the identification of a high number of erroneous haplotypes. The use of replicate samples in SeekDeep2x greatly decreased the tool's false positive rate, increasing precision with a small cost in sensitivity.

Each tool's performance varied to a some extent across amplicons. This variation was not consistent across pipelines, however, and as a result, the pipelines' rank order for precision and sensitivity was different for *CSP* and *SERA2* (Additional File 1, Figure S4).

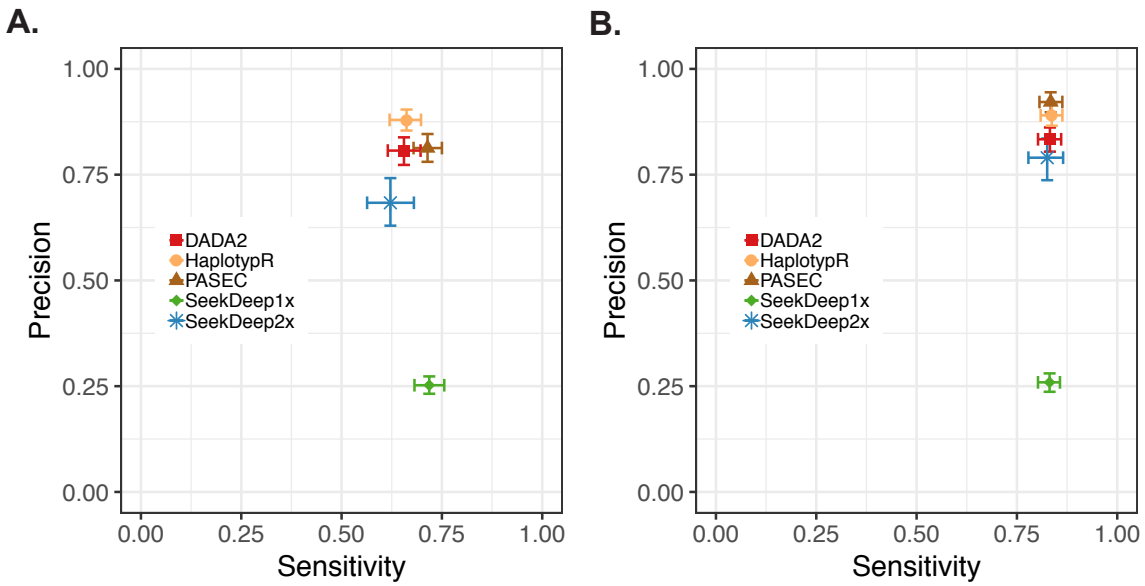

Figure 5. Sensitivity and precision of five analysis pipelines for the detection of haplotypes in mock samples. (A) Analysis approaches vary most in their precision. (B) Performance of all approaches improves when only considering samples that had at least 100 reads for an individual amplicon. Displayed data includes results from both the *CSP* and *SERA2* amplicons. 95% confidence intervals were estimated with 1000 bootstrapped data set replicates.

Effect of sample read depth and genome copy number: All five pipelines showed reduced performance at low read depths (<25 reads) and low parasite concentrations (<5 genomes/µl; Additional File 1, Figure S5). In particular, SeekDeep2x performed most optimally on samples with at least 100 reads (Figure 5B). Parasite genome copy number also affected the tools success at resolving any haplotype within a sample. Overall, the pipelines reported haplotypes for 78% (HaplotypR), 81% (DADA2), 84% (SeekDeep2x), 89% (PASEC), and 96% (SeekDeep1x) of the samples (Additional File 1, Figure S6A). The majority of the samples returning no data contained *Plasmodium* DNA concentrations under 5 genomes/µl Additional File 1, Figure S6B).

Analysis of samples from Sub-Saharan Africa with the four tools

All four tools were also applied to the newly generated amplicon data from 95 previously sampled patient infections from four countries in sub-Saharan Africa (Figure 1C) [1]. These biological samples were PCR amplified and sequenced in duplicate, yielding 190 independently sequenced samples. With the exception of SeekDeep2x, the technical replicates were again treated as separate samples in the analysis step. All tools were run with the same parameters used for the mock samples.

The tools differed in the total number of unique haplotypes identified across the samples, with estimates ranging from 48 to 336 for *CSP* and 38 to 412 for *SERA2* (Additional File 1, Figure S7). For both amplicons, SeekDeep1x and DADA2

identified substantially more haplotypes than the other approaches, although a large percentage of these haplotypes were found at within-sample frequencies under 1%, raising the possibility that they were artifacts.

Consistent with previous observations in sub-Saharan Africa, the majority of the patient samples contained multiple *P. falciparum* clones. COI was estimated for each sample as the maximum number of unique haplotypes identified at either of the two amplicons. With the exception of SeekDeep1x, all four tools produced similar estimates of mean COI per country (Figure 6). This is in keeping with the observation that SeekDeep showed lower precision on the mock samples than the other tools when run with single replicates (Figure 5).

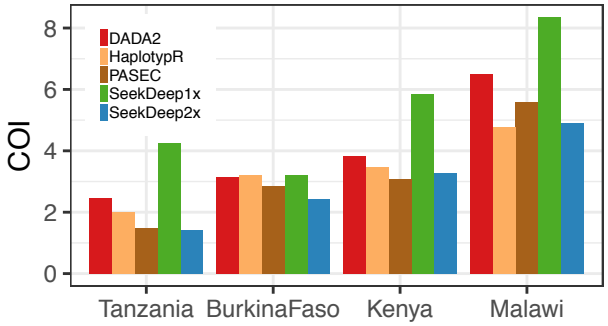

Figure 6. Mean COI estimates for four sub-Saharan African study sites made by the five analysis pipelines. COI was estimated from the maximum number of haplotypes retrieved for the sample from either of the two amplicons.

Discussion

Amplicon sequencing of complex haplotypic regions is being applied to an increasing range of questions in malaria research. This growth of amplicon-based applications has led to the development of a number of new analytical tools and several studies investigating the possibilities and limitations of the approach. Here, the performance of amplicon sequencing was assessed for the first time under a scenario of extremely low parasite densities (1-200 genomes/µl). The results show that amplicon sequencing is more challenging at low parasite densities, however, the approach was still able to detect individual haplotypes present at concentrations of 5-10 genomic copies/µl with 77% accuracy. The study design mimicked samples that could be obtained from asymptomatic carriers, a population segment that is receiving increased attention in regions nearing malaria eradication. The ability of Illumina-based amplicon sequencing to reliably detect *Plasmodium* DNA at these low concentrations shows that it has a limit of detection on par with standard nested PCR [27] and qPCR [28] methods.

The analysis of deep coverage from highly diverse amplicon sequences incurs several technical challenges, including the identification of sequencing errors, chimeric reads and sample contaminants. The four tools compared here—PASEC, DADA2, HaplotypR, and SeekDeep—have different approaches towards error correction yet all performed comparably well on the set of mock samples and

provided similar COI results for a panel of patient samples. PASEC's high performance was the result of extensive pilot work and hand-tuning for use with the amplicons *CSP* and *SERA2*, for instance the hard masking of difficult-to-sequence homopolymer runs in the *CSP* amplicon and the *a priori* identification of indels in *SERA2*. As a result of this customization, it was the only tool to identify a naturally occurring three nucleotide deletion in *SERA2* that is present in Africa. Importantly, however, this study shows that other tools also provide robust results without upfront knowledge of an individual amplicon. In addition, methodological developments are still underway. In particular, work on the identification of sample contamination may lead to lower false positive rates in future studies [29, 30].

As expected, the accuracy of amplicon sequencing is reduced on samples with low parasite densities, regardless of the applied analysis tool. Stochasticity at the level of sample preparation doubtless impacts the approach's ability to quantify haplotypes at low concentrations. In addition, overall error rates are higher in samples with low parasite density and low read counts. Researchers can therefore take steps to lower false positive rates in these challenging classes of samples. Erroneous haplotypes are generally supported by fewer reads (Figure 3B) and samples with lower read counts have a higher proportion of false haplotypes (Additional File 1, Figure S3). It should therefore be standard practice to raise read thresholds when analyzing low parasitemia or low coverage samples.

Conclusion

Amplicon sequencing is a versatile approach for exploring a range of intra-host questions in malaria research. Cost-effective and scalable for use with thousands—or tens of thousands—of samples in high-throughput settings, its use will likely increase in the coming years. As shown here, amplicon sequencing can be applied to samples with both low and high parasite densities, although the consistent detection of parasite clones with very low prevalence (<5 genomes/μl) is challenging. Even at low densities, amplicon sequencing retained some information on haplotype ratio, allowing major and minor clones to be distinguished within 73% of the infections. Currently, several tools exist for the general analysis of long, multi-variant amplicons. Three of these versatile tools (DADA2, HaplotypR, and SeekDeep) showed similar performance compared to PASEC, a method specifically developed for use with the two amplicons sequenced here: *CSP* and *SERA2*. While all tools performed well, final choice of analysis method should take into account aspects of study design (such as the inclusion of technical PCR/sequencing replicates), the read coverage of the samples, and expectations regarding the targeted *Plasmodium* genotypes (for instance, the potential presence of indels). Regardless of the tool used, future studies involving samples with parasitemias <5 parasites/μl will likely benefit from more stringent read-count filtration as accuracy was consistently lower for these samples.

Additional Files

Additional File 1: Supplementary Figures and Tables

Additional File 2: Supplementary PASEC Documentation

Additional File 3: Analysis Pipeline Files for PASEC, DADA2, HaplotypR, and SeekDeep

(zip file)

References

1. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, Griggs A, et al. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. N Engl J Med. 2015;373:2025–37. doi:10.1056/NEJMoa1505819.
2. Zhong D, Koepfli C, Cui L, Yan G. Molecular approaches to determine the multiplicity of Plasmodium infections. Malar J. 2018;17:172. doi:10.1186/s12936-018-2322-5.
3. Juliano JJ, Porter K, Mwapasa V, Sem R, Rogers WO, Ariey F, et al. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. Proc Natl Acad Sci. 2010;107:20138–43. doi:10.1073/pnas.1007068107.
4. Lerch A, Koepfli C, Hofmann NE, Kattenberg JH, Rosanas-Urgell A, Betuela I, et al. Longitudinal tracking of Plasmodium falciparum clones in complex infections by amplicon deep sequencing. bioRxiv. 2018;:306860. doi:10.1101/306860.
5. Lin JT, Hathaway NJ, Saunders DL, Lon C, Balasubramanian S, Kharabora O, et al. Using Amplicon Deep Sequencing to Detect Genetic Signatures of *Plasmodium vivax* Relapse. J Infect Dis. 2015;212:999–1008. doi:10.1093/infdis/jiv142.
6. Patel JC, Hathaway NJ, Parobek CM, Thwai KL, Madanitsa M, Khairallah C, et al. Increased risk of low birth weight in women with placental malaria associated with P. falciparum VAR2CSA clade. Sci Rep. 2017;7:7768. doi:10.1038/s41598-017-04737-y.
7. Dara A, Travassos MA, Adams M, Schaffer DeRoo S, Drábek EF, Agrawal S, et al. A new method for sequencing the hypervariable Plasmodium falciparum gene var2csa from clinical samples. Malar J. 2017;16:343. doi:10.1186/s12936-017-1976-8.
8. Mideo N, Bailey JA, Hathaway NJ, Ngasala B, Saunders DL, Lon C, et al. A deep sequencing tool for partitioning clearance rates following antimalarial treatment in polyclonal infections. Evol Med Public Heal. 2016;2016:21–36. doi:10.1093/emph/eov036.
9. Nair S, Li X, Arya GA, McDew-White M, Ferrari M, Nosten F, et al. Do fitness costs explain the rapid spread of *kelch13* -C580Y substitutions conferring artemisinin resistance? Antimicrob Agents Chemother. 2018;:AAC.00605-18. doi:10.1128/AAC.00605-18.
10. Ngondi JM, Ishengoma DS, Doctor SM, Thwai KL, Keeler C, Mkude S, et al. Surveillance for sulfadoxine-pyrimethamine resistant malaria parasites in the Lake and Southern Zones, Tanzania, using pooling and next-generation sequencing. Malar J. 2017;16:236. doi:10.1186/s12936-017-1886-9.
11. Talundzic E, Ndiaye YD, Deme AB, Olsen C, Patel DS, Biliya S, et al. Molecular Epidemiology of Plasmodium falciparum kelch13 Mutations in Senegal Determined by Using Targeted Amplicon Deep Sequencing. Antimicrob Agents Chemother. 2017;61:AAC.02116-16. doi:10.1128/AAC.02116-16.
12. Rao PN, Uplekar S, Kayal S, Mallick PK, Bandyopadhyay N, Kale S, et al. A Method for Amplicon Deep Sequencing of Drug Resistance Genes in Plasmodium falciparum Clinical Isolates from India. J Clin Microbiol. 2016;54:1500–11.

doi:10.1128/JCM.00235-16.
13. Early AM, Lievens M, MacInnis BL, Ockenhouse CF, Volkman SK, Adjei S, et al. Host-mediated selection impacts the diversity of Plasmodium falciparum antigens within infections. Nat Commun. 2018;9:1381. doi:10.1038/s41467-018-03807-7.
14. Miller RH, Hathaway NJ, Kharabora O, Mwandagalirwa K, Tshefu A, Meshnick SR, et al. A deep sequencing approach to estimate Plasmodium falciparum complexity of infection (COI) and explore apical membrane antigen 1 diversity. Malar J. 2017;16:490. doi:10.1186/s12936-017-2137-9.
15. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14:R51. doi:10.1186/gb-2013-14-5-r51.
16. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17:125. doi:10.1186/s12859-016-0976-y.
17. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;43:e37–e37. doi:10.1093/nar/gku1341.
18. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3. doi:10.1038/nmeth.3869.
19. Lerch A, Koepfli C, Hofmann NE, Messerli C, Wilcox S, Kattenberg JH, et al. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. BMC Genomics. 2017;18:864. doi:10.1186/s12864-017-4260-y.
20. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. Nucleic Acids Res. 2018;46:e21–e21. doi:10.1093/nar/gkx1201.
21. Rask TS, Petersen B, Chen DS, Day KP, Pedersen AG. Using expected sequence features to improve basecalling accuracy of amplicon pyrosequencing data. BMC Bioinformatics. 2016;17:176. doi:10.1186/s12859-016-1032-7.
22. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11:2639–43. doi:10.1038/ismej.2017.119.
23. Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, Park DJ, et al. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. Malar J. 2008;7:223. doi:10.1186/1475-2875-7-223.
24. RTS,S Clinical Trials Partnership, Agnandji ST, Lell B, Fernandes JF, Abossolo BP, Methogo BGNO, et al. A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants. N Engl J Med. 2012;367:2284–95. doi:10.1056/NEJMoa1208394.
25. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27:2957–63. doi:10.1093/bioinformatics/btr507.
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.

27. Snounou G, Viriyakosol S, Xin Ping Zhu, Jarra W, Pinheiro L, do Rosario VE, et al. High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. Mol Biochem Parasitol. 1993;61:315–20. doi:10.1016/0166-6851(93)90077-B.
28. Taylor SM, Juliano JJ, Trottman PA, Griffin JB, Landis SH, Kitsa P, et al. High-Throughput Pooling and Real-Time PCR-Based Strategy for Malaria Detection. J Clin Microbiol. 2010;48:512–9. doi:10.1128/JCM.01800-09.
29. Larsson AJM, Stanley G, Sinha R, Weissman IL, Sandberg R. Computational correction of index switching in multiplexed sequencing libraries. Nat Methods. 2018;15:305–7. doi:10.1038/nmeth.4666.
30. Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. bioRxiv. 2018;:221499. doi:10.1101/221499.