Optimization of parasite DNA enrichment approaches to generate whole genome sequencing data for Plasmodium falciparum from low-parasitaemia samples

Zalak Shah
University of Maryland School of Medicine
Matthew Adams
University of Maryland School of Medicine
Kara A Moser
University of Maryland School of Medicine
Biraj Shrestha
University of Maryland School of Medicine
Emily M Stucke
University of Maryland School of Medicine
Miriam K Laufer
University of Maryland School of Medicine
David Serre
University of Maryland School of Medicine
Joana C Silva
University of Maryland School of Medicine
Shannon Takala-Harrison (stakala@medicine.umaryland.edu)
University of Maryland School of Medicine https://orcid.org/0000-0003-4674-8500

Methodology

Keywords: Plasmodium falciparum, malaria, whole genome sequencing, selective whole genome amplification, vacuum filtration

Posted Date: March 18th, 2020

DOI: https://doi.org/10.21203/rs.2.17581/v3

Version of Record: A version of this preprint was published at Malaria Journal on March 30th, 2020. See the published version at https://doi.org/10.1186/s12936-020-03195-8.

Abstract

Background Owing to the large amount of host DNA in clinical samples, generation of high-quality Plasmodium falciparum whole genome sequencing (WGS) data requires enrichment for parasite DNA. Enrichment is often achieved by leukocyte depletion of infected blood prior to storage. However, leukocyte depletion is difficult in low-resource settings and limits analysis to prospectively-collected samples. As a result, approaches such as selective whole genome amplification (sWGA) are being used to enrich for parasite DNA. However, sWGA has had limited success in generating reliable sequencing data from low parasitaemia samples. In this study, enzymatic digestion with MspJI prior to sWGA and whole genome sequencing was evaluated to determine whether this approach improved genome coverage compared to sWGA alone. The potential of sWGA to cause amplification bias in polyclonal infections was also examined. Methods DNA extracted from laboratory-created dried blood spots was treated with a modification-dependent restriction endonuclease, MspJI, and filtered via vacuum filtration. Samples were then selectively amplified using a previously reported sWGA protocol and subjected to WGS. Genome coverage statistics were compared between the optimized sWGA approach and the previously reported sWGA approach performed in parallel. Differential amplification by sWGA was assessed by comparing WGS data generated from lab-created mixtures of parasite isolates, from the same geographical region, generated with or without sWGA. Results MspJI digestion did not enrich for parasite DNA. Samples that underwent vacuum filtration (without MspJI digestion) prior to sWGA had the highest parasite DNA concentration and displayed greater genome coverage compared to MspJI+sWGA and sWGA alone, particularly for low parasitaemia samples. The optimized sWGA (filtration + sWGA) approach was successfully used to generate WGS data from 218 non-leukocyte depleted field samples from Malawi. Sequences from lab-created mixtures of parasites did not show evidence of differential amplification of parasite strains compared to directly sequenced samples. Conclusion This optimized sWGA approach is a reliable method to obtain WGS data from non-leukocyte depleted, low parasitaemia samples. The absence of amplification bias in data generated from mixtures of isolates from the same geographic region suggests that this approach can be appropriately used for molecular epidemiological studies.

Keywords Plasmodium falciparum , malaria, whole genome sequencing, selective whole genome amplification, vacuum filtration

Background

Next-generation sequencing has greatly advanced research on malaria parasite genomics. Several molecular epidemiological studies have used genomic approaches in an effort to better understand *Plasmodium falciparum* genetic diversity in relation to malaria transmission, drug resistance and vaccine design [1–5]. However, a majority of such population genomics studies rely on whole genome sequencing of samples collected in malaria endemic areas. Since patient blood samples contain mostly human DNA, enrichment for parasite DNA is required in order to obtain parasite sequence data with adequate genome coverage. Leukocyte depletion is an effective method for reducing the amount of host DNA for parasite sequencing and hence increase the proportion of parasite DNA prior to sequencing; however, depletion

must be performed within hours of sample collection and can be logistically challenging in some resource-limited settings [6,7] . In addition, once the sample is frozen and cells are lysed, leukocyte depletion is no longer effective, thus limiting the application of this approach to prospectively-collected samples. To reduce the need for extensive sample processing in the field and to enable examination of the wealth of historical samples collected as dried blood spots or whole venous blood, malaria researchers have explored alternative parasite DNA enrichment approaches including enzymatic digestion of human DNA [8], selective whole genome amplification (sWGA) [9], and hybrid selection (capture-based method) [10].

MspJI is a restriction endonuclease that cleaves specific motifs containing methylated cytosines that has been used to selectively digest human DNA prior to parasite DNA sequencing [11]. The success of this approach is based on the assumption of different methylation patterns in the human and parasite genomes; however, methylation patterns in *P. falciparum* are not fully understood [12]. sWGA of the parasite genome over the human genome has also shown promising results as a method for enrichment of parasite DNA prior to whole genome sequencing. This approach uses multiple displacement amplification with phi29 DNA polymerase using primers binding at greater density in the parasite genome compared to the human genome. Phi29 results in amplification of long DNA fragments and is known to have a low error rate. An existing sWGA protocol has been shown to work best with samples that have parasitaemia greater than ~1200 parasites/μL [9]. While this parasitaemia threshold may allow sequencing of most clinical infections, it limits studies of lower parasitaemia infections, including submicroscopic infections that may significantly contribute to the malaria burden in some areas [13]. In addition, because sWGA primers were designed against the 3D7 reference sequence without consideration of *P. falciparum* genetic diversity, the potential for differential amplification of particular parasite clones within a polyclonal infection (e.g. those most genetically similar to the reference) is a concern. Since a large proportion of infections from high-transmission areas are polyclonal, the possibility of amplification bias introduced by sWGA warrants further investigation, as such bias could lead to inaccurate inferences in downstream analyses.

In this study, enzymatic digestion with MspJI prior to sWGA and whole genome sequencing was evaluated to determine whether this approach improved genome coverage compared to sWGA alone when applied to samples representing a range of parasitaemias. In addition, this study also evaluated whether the optimized sWGA protocol results in biased estimates of multiplicity of infection or allele frequencies by comparing whole genome sequence data generated from laboratory-created mixtures of isolates from Malawi that underwent sWGA prior to sequencing or were directly sequenced.

Methods

Laboratory-created samples to evaluate enrichment approaches

Dried blood spots. To test the different parasite DNA enrichment approaches, dried blood spots were created by mixing cultured NF54-infected red blood cells with uninfected whole human blood. The

laboratory-adapted isolate, NF54, was maintained in culture following the method of Trager and Jensen [14]. Parasite concentrations were microscopically enumerated from sorbitol-synchronized ring-stage NF54 culture and mixed with uninfected whole human blood (Interstate Blood Bank, Nashville, TN) resulting in parasite concentrations ranging from 10,000 to 500 parasites/μL and subsequently spotting 12.5 mL of blood onto Whatman 3 MM filter paper. DNA was extracted from dried blood spots using the protocol described by Zainabadi *et al.* [15]. The DNA was treated under three conditions, including sWGA only and MspJI or MspJI ⁻control (no enzyme)+ followed by sWGA, as illustrated in Supplementary Fig. S1. The MpsJI⁻ condition followed the same protocol as MspJI, without the enzyme.

Mixtures of DNA to assess amplification bias. DNA from four previously cultured and sequenced field isolates from Malawi [16] were mixed in equal proportions to assess potential amplification bias introduced by sWGA. The DNA concentration for each sample was measured using a picogreen assay (Thermo Fisher Scientific, Waltham, MA) and the samples were mixed in equal proportions. None of the cultured isolates represented polyclonal infections, based on analysis of prior sequencing data using estMOI [17]. The sample was further split into six tubes, three of which underwent direct whole genome sequencing and the other three which underwent sWGA, followed by whole genome sequencing, as shown in Fig. 4.

MspJI digestion. MspJI digestion was performed in a 0.2 mL 96-well PCR plate. The reaction mixture contained 1x CutSmart Buffer, 10 g of bovine serum albumin and 6 units of MspJI (New England Biolabs, Ipswich, MA). The MspJI digestion control (i.e., MspJI⁻) contained the same buffers but excluded the enzyme. 25 μL of sample DNA was added to both reaction mixtures (total volume, 30 μL), and reactions were incubated in a thermocycler. Two different incubation protocols were tested, one with a 16 hrs incubation at 37°C, followed heating at 65°C for 20 min to inactivate the enzyme and cooling at 4°C, and a second protocol where the 37°C incubation lasted only 4 hrs.

Vacuum filtration of DNA. Following enzymatic digestion, the entire reaction mixture was transferred to a MultiScreen® PCR Filter Plate (Millipore) and filtered to remove digested DNA fragments using a MultiScreen® Vacuum Manifold with a pressure of -7 inches Hg until the wells were emptied and the filters appeared dry. Filtered samples were reconstituted with 30 μL of water, and the plate was gently agitated for 15 mins. Samples were then transferred to a new plate.

sWGA. Amplification was performed in a 0.2mL 96-well PCR plate. The reaction mixture contained 1x BSA, 1mM dNTPs, 2.5μM of each amplification primer, 1x Phi29 reaction buffer and 30 units of Phi29 polymerase. 17μL of template DNA was added to the reaction mixture (total volume, 50 mL) which was then placed in a thermocycler programmed for a stepdown protocol (35° C for 5 min, 34°C for 10 min, 33°C for 15 min, 32°C for 20 min, 31°C for 30 min, 30°C for 16 hrs), followed by heating at 65°C to inactivate the enzyme and cooling at 4°C. Primers used for sWGA were the same as those published by Oyola *et al.* [9] (see Supplementary Material).

qPCR. Quantitative PCR of the human actin gene and *P. falciparum* 18S rRNA gene was used to estimate the amount of human and parasite DNA, respectively, before and after sWGA. The reaction mixture contained QuantiTech 2x QT Multiplex Master Mix, 10 μM of each of the primers and 1.5 mL of template DNA (total volume, 10 mL). A two-tailed Mann-Whitney U test was used to estimate differences between different experimental conditions (e.g. MspJI-sWGA, MspJI^{-}-sWGA, sWGA).

Whole genome sequencing. Genomic DNA libraries were constructed for sequencing using the KAPA Library Preparation Kit (Kapa Biosystems, Woburn, MA). DNA (500 hg) was fragmented with the Covaris E210 to ~200 bp. Libraries were prepared using a modified version of the manufacturer's protocol. The DNA was purified between enzymatic reactions and library size selection was performed with AMPure XT beads. Libraries were assessed for concentration and fragment size using the DNA High Sensitivity Assay on the LabChip GX (Perkin Elmer, Waltham, MA). Library concentrations were also assessed by qPCR using the KAPA Library Quantification Kit. Libraries were pooled and sequenced on a 150 bp paired-end Illumina HiSeq 4000 run (Illumina, San Diego, CA).

Data analysis

Read mapping and coverage. Each dataset was analysed by mapping raw fastq files to the 3D7 reference genome using Bowtie2 [18]. Bam files were processed according to GATK's Best Practices workflow to obtain analysis-ready reads [19,20]. Bedtools [21] was used to generate coverage and depth estimates from the processed reads. Differences in the proportion of the genome covered in samples from untreated *vs*. filtered sWGA were tested using the z-score test for difference in proportions.

Variant calling. GATK's Best Practices workflow was followed for variant calling [19,20]. Haplotype Caller was used in reference confidence mode to create genomic variant call format (GVCF) files for each sample and joint SNP Calling (GATK v3.7). Variants were removed if they met the following filtering criteria: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, QUAL < 50. Variant sites with >20% missing genotypes were additionally removed using vcftools.

Assessment of amplification bias. Whole genome sequence data generated from laboratory-created isolate mixtures that either underwent sWGA and sequencing or direct sequencing were compared to evaluate the potential for sWGA to introduce amplification bias. The experimental design is illustrated in Fig. 4. Reference allele frequencies for each sample were estimated using samtools mpileup and were compared to examine the variability between and within samples from sWGA and non-sWGA groups. Sites with coverage depth lower than 20x were excluded from this analysis, as were sites not called in all samples. After applying these filters, 1,786,088 sites remained. The distribution and Spearman correlation coefficient (rho) were estimated in R.

F_{WS}, a measure of within-sample diversity, was used to estimate infection complexity for each isolate mixture for comparison between the sWGA and non-sWGA conditions. F_{WS} was estimated using the R

package, *moimix* (https://github.com/bahlolab/moimix). Significance was determined using the Mann-Whitney U test. Only the core *P. falciparum* genome was used to estimate F_{WS} [22].

The composition of each laboratory-created mixture was also compared to determine if there were significant differences between the sWGA and non-sWGA groups, implying differential amplification of certain strains over others. To assess potential amplification bias, the proportion of isolate-specific SNPs called from WGS data generated from isolate mixtures that did or did not undergo sWGA prior to sequencing was compared for each isolate. Isolate-specific SNPs were identified by comparing WGS data from each of the four isolates used to create the experimental mixtures. All variant sites with missing alleles in any of the isolates were removed from the analysis. For each isolate mixture, the predominant allele (defined as an allele comprising >70% of reads) at each position was called; if no allele was predominant based on this threshold, the SNP was called missing. The positions of unique SNPs for each isolate were extracted from the sequence data generated from the isolate mixtures. The proportion of unique SNPs from each isolate in the mixture was estimated and compared between the sWGA and non-sWGA groups. Significance was estimated using a z-score test for difference in proportions.

Results

Vacuum filtration, but not enzyme digestion, prior to sWGA increases parasite DNA concentration and improves the quality of whole genome sequence data

Laboratory-created dried blood spots representing a range of parasitaemias were created to test the different DNA enrichment approaches. DNA extracted from the dried blood spots underwent one of three conditions prior to sWGA: 1) MspJI digestion, 2) MspJI$^-$control (same conditions as MspJI but without enzyme), or 3) untreated, as illustrated in Fig. S1. Samples that underwent MspJI digestion had significantly less human (p=0.028, Mann-Whitney U test) and parasite DNA (p=0.028, Mann-Whitney U test) compared to the untreated samples. Similarly, MspJI-digested samples also had significantly less human (p=0.028, Mann-Whitney U test) and parasite DNA (p=0.028, Mann-Whitney U test) compared to the MspJI$^-$ control (Fig. 1A). This pattern was consistent following sWGA, with MspJI-sWGA samples having significantly less human DNA (p-value=0.028, Mann-Whitney U test) as well as parasite DNA (p-value=0.028, Mann-Whitney U test) compared to the untreated-sWGA samples. Surprisingly, the MspJI$^-$ control samples that underwent sWGA had a significantly higher parasite DNA concentration compared to samples digested with MspJI (p-value=0.028, Mann-Whitney U test) and untreated samples (p-value=0.028, Mann-Whitney U test) that underwent sWGA. Further experiments suggested that the vacuum filtration step in the MspJI$^-$ condition (Fig. S1) was likely responsible for the improved parasite DNA concentration (Fig. 1B, p-value=0.028, Mann-Whitney U test), possibly due to removal of small DNA fragments that may lead to non-specific binding. This result was also consistent across samples with lower parasitaemias, ranging from 5000 parasites/μL to 500 parasites/μL (Fig. S2, p-value=0.028, Mann-Whitney U test).

The untreated-sWGA (no treatment, only sWGA) and filtered-sWGA samples (vacuum filtration, followed by sWGA), with different parasitaemias (10,000, 1000, 500 parasites/μL), underwent whole genome sequencing, and the sequence reads were trimmed and mapped to the *P. falciparum* 3D7 reference genome. As seen in Fig. 2A, the filtered-sWGA samples had a higher percentage of reads map to 3D7 ($p < 0.0001$, z score test for difference in proportions). This pattern was more notable for samples with lower parasitaemia (Fig. 2A). Sequence data from filtered-sWGA samples displayed a slightly higher percent of the genome with 5x coverage per million reads sequenced compared to sequence data from untreated-sWGA samples (Fig. 2B). Further in-depth coverage analysis showed that this pattern was consistent across all chromosomes (Fig. S3A). Visual inspection of coverage along chromosome 1 showed uneven coverage in both filtered and unfiltered samples that underwent sWGA but filtered-sWGA samples had higher coverage in most regions, particularly in the lowest parasitaemia sample. Indeed, in the lowest parasitaemia sample, higher coverage was observed in several regions where the untreated-sWGA sample had very little or no coverage (Fig. S3B).

Next, the optimized sWGA protocol was tested on 218 red blood cell pellets from Malawi that were PCR-positive for *P. falciparum* and had parasitaemias, ranging from 0 to >200,000 parasites/μL (by microscopy). Only 23/218 (10.55%) samples had less than 75% of the genome with >= 5x coverage. Samples with >75% of the genome with >=5x coverage had a median average read depth of ~137x. Out of 50 samples with parasitaemia <500 parasites/mL, only 9 (18%) samples had less than 75% of the genome with >=5x coverage, while the remaining samples with >75% of the genome with >= 5x coverage had a median average read depth of ~129x. The correlation between percent genome coverage and parasitaemia was also estimated

(Fig. 3) and found to be weak, but statistically significant ($p = 0.00012$, Spearman's correlation $rho = 0.25$).

Optimized sWGA does not show evidence of amplification bias when applied to mixtures of parasite isolates from the same geographic region.

To examine the potential for amplification bias introduced by sWGA, equal mixtures of four *P. falciparum* isolates from Malawi were created. Three aliquots of these mixtures underwent sWGA followed by whole genome sequencing and three underwent whole genome sequencing without prior sWGA (Fig. 4). Mixtures that underwent sWGA had a larger proportion of sequenced reads that mapped to the 3D7 reference genome compared to directly-sequenced mixtures, while directly-sequenced mixtures had a higher percentage of the genome with at least 5x coverage (Table 1).

Three approaches were used to evaluate the potential for amplification bias by sWGA (Fig. 4). First, the correlation between reference allele frequencies was estimated for each variant site in sequence data generated from mixtures that did or did not undergo sWGA prior to sequencing. Based on visual examination, the distribution of reference allele frequencies was similar in all six samples, but directly-sequenced samples had more clearly distinct peaks representing each of the four isolates compared to samples that underwent sWGA prior to sequencing (Fig. S4). The correlation between reference allele

frequencies was high (>94%) and similar within and between sWGA or directly-sequenced samples (Fig. S5).

Second, F_{ws}, a measure of within-sample diversity [2,23,24], was estimated to examine differences in infection complexity between samples that underwent sWGA prior to sequencing and those that did not. F_{ws} values range from 0 to 1, with 0 indicating a mixture of highly unrelated clones and 1 indicating a single clone. Directly-sequenced samples had lower average F_{ws} estimates than samples that underwent sWGA (Table 1). However, when F_{ws} was estimated based on the subset of SNPs called in both groups, there was no significant difference in F_{ws} between groups (p-value=0.1, Mann Whitney U Test).

Finally, the proportion of each isolate in our isolate mixtures was compared based on the frequency of isolate-specific variants in sequence data generated from the mixtures that did or did not undergo sWGA. The proportion of isolate-specific SNPs in each mixture did not differ significantly between mixtures that underwent sWGA and those that did not (Fig. 5).

Discussion

While advances in next-generation sequencing technologies have greatly expanded research in the field of malaria genomics, the difficulties of enriching for malaria parasite DNA in clinical samples, particularly those collected from submicroscopic infections, has limited population genomics analyses to include mostly high-parasitaemia, symptomatic infections. In this study, two published parasite DNA enrichment approaches were combined, namely enzyme digestion of human DNA [8] and sWGA [9], to determine whether the combined approaches improved the ability to generate high-quality whole genome sequence data from non-leukocyte-depleted clinical samples with low parasitaemia. Oyola *et al.* reported up to ~9 fold *P. falciparum* DNA enrichment resulting in >98% of the parasite genome with at least 5x coverage when using MspJI digestion only (no sWGA) prior to sequencing [8]. However, in this study, enzyme digestion with MspJI resulted in a significant decrease in both human and parasite DNA concentrations, before and after sWGA. This reduced DNA concentration may be due to digestion of both human and parasite DNA, consistent with the presence of a “smear” of small DNA fragments observed when digested samples were subjected to gel electrophoresis (data not shown). This finding contrasts with that of Oyola *et al.* who observed a band representing intact parasite DNA along with digested human DNA following MspJI digestion, although it is notable that the amount of starting DNA (1ug total) used in the Oyola study was much larger than that obtained from the dried blood spots in this study, and likely larger than what would be obtained from dried blood spots collected in the field. In addition, Cowell *et al.* [24] observed no significant difference in genome coverage with digestion using MspJI or FspEI prior to sWGA and sequencing of *Plasmodium vivax*. However, since MspJI targets specific motifs containing methylated cytosines, and *P. vivax* has higher GC content than *P. falciparum,* it is possible that the targeted motifs may be more common in *P. vivax,* leading to the failure of MspJI in this context. Though MspJI digestion was not successful, further investigation is warranted to identify alternative enzymes

that target the human genome over the parasite genome that could potentially perform better and be useful in combination with sWGA.

While digestion with MpsJI did not improve parasite DNA concentration, surprisingly,the parasite DNA concentration was significantly greater in the MspJI^{-}control compared to MspJI and no treatment. Subsequent experiments suggested that this increased DNA concentration resulted from the vacuum filtration step. Indeed, it was discovered that the filtration of extracted DNA prior to sWGA resulted in a greater parasite DNA concentration compared to unfiltered DNA that underwent sWGA. This result may be due to removal of small fragments of parasite DNA that may bind sWGA primers but not lead to effective amplification because of their short length. The increased DNA concentration obtained from the optimized sWGA protocol also resulted in increased genome coverage of whole genome sequencing data, although this increase was most pronounced in lower parasitaemia samples.

Using the optimized sWGA approach, high coverage whole genome sequencing data was obtained from DNA extracted from laboratory-created dried blood spots with parasitaemias as low as 500 parasites/μL, as well as from non-leukocyte depleted red blood cell pellets from the field, including 41 samples from low parasitaemia, sub-microscopic infections. Although statistically significant, the correlation between parasitaemia and percent of the genome with 5x coverage was small, suggesting that this optimized sWGA approach can be used to obtain high-quality whole genome sequence data from low parasitaemia samples. Additional testing of this approach on dried blood spots stored under different environmental conditions and storage times will be required to further characterize which samples are likely to yield successful results using this optimized approach. If the parasite DNA is highly degraded, this optimized approach may not be successful, both because small DNA fragments may be removed during filtration and lead to less effective whole genome amplification and because of a general lack of longer fragments needed for successful sWGA. For more highly degraded samples, capture-based methods of enrichment may yield better results.

Amplification bias following sWGA or capture-based parasite DNA enrichment has been largely understudied but is essential to understand given the high prevalence of polyclonal infections in high transmission areas. The potential for amplification bias following sWGA was further explored by comparing whole genome sequence data generated from experimental mixtures of parasite DNA from four culture-adapted parasite isolates from Malawi that were either directly sequenced or underwent sWGA prior to sequencing. Consistent with other studies, genome coverage of sequence data generated from samples that underwent sWGA was more uneven compared to coverage of sequence data generated through direct sequencing, most likely due to the sparse sWGA primer coverage in diverse and AT-rich subtelomeric regions [9,25]. However, pairwise per base reference allele frequencies within the core genome were highly correlated both within and between the sWGA and directly-sequenced groups, suggesting that sWGA is not substantially biasing allele frequencies. Experimental mixtures that underwent sWGA had significantly higher estimates of F_{ws}, implying some loss of within-mixture diversity. This result is in agreement with the results of Cowell *et al.* who found reduced estimates of infection complexity following sWGA of *P. vivax* DNA, based on analysis of both sequence data and

microsatellites [24], and may be explained by differences in genome coverage between groups. Indeed, directly sequenced samples had higher genome coverage than samples that underwent sWGA, even in the core genome. To test this hypothesis, F_{WS} was estimated based only on SNPs called in both the sWGA samples and the directly sequenced samples and no significant difference in F_{WS} was observed between groups, suggesting the lower diversity in samples that underwent sWGA was not the result of amplification bias favouring some isolates over others. This conclusion is also supported by the lack of significant differences in the proportion of isolate-specific variants between the sWGA and directly sequenced isolates. More in-depth analysis will be required to evaluate which genomic regions have reduced coverage in sWGA samples and the implications for downstream analyses.

While preferential amplification was not observed based on equal mixtures of isolates from Malawi based on variants called against the 3D7 reference (believed to be of African origin [16,26]), it would be informative to evaluate this phenomenon in mixtures of parasites from other geographic regions or in varying proportions to determine whether one set of sWGA primers will provide unbiased results and high genome coverage in all settings or whether sWGA primers designed based on regional reference genomes is necessary. Because parasites from different continents are more genetically differentiated than parasites from the same geographic region [2], amplification bias may be more of a concern for comparisons involving parasites sampled from different continents than for comparisons of parasites from the same geographic region.

Conclusions

The optimized sWGA approach is a reliable method to obtain WGS data from non-leukocyte depleted, low parasitaemia samples. The absence of amplification bias in data generated from mixtures of isolates from the same geographic region suggests that this approach can be appropriately used for molecular epidemiological studies.

List Of Abbreviations

DBS: dried blood spots

MOI: multiplicity of infection

sWGA: selective whole genome amplification

WGS: whole genome sequencing

Declarations

Ethics approval and consent to participate

Red blood cell pellets used in this study were collected as part of the Mfera Cohort Study conducted in Malawi from 2014-2017 [27]. Written informed consent was obtained from parents or guardians of all study participants according to protocols approved by institutional review boards at the University of Maryland School of Medicine and the National Health Sciences Research Committee of Malawi.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by funding from the following awards granted by the National Institutes of Health: R01AI101713, R01AI125579, U19AI110820, K24AI114996, and the Malawi International Center of Excellence for Malaria Research U19AI089683.

Authors' contributions

ZS, MA, DS, JCS and ST-H designed the experiments. MKL provided samples. ZS and BS carried out the experiments. ZS, KM and EMS performed data analysis. MA, MKL, DS, JCS and ST-H provided guidance on data analysis. ZS and ST-H drafted the manuscript, and MA, KM, BS, EMS, MKL, DS and JCS approved the manuscript for publication.

Acknowledgements

We thank the participants in the Mfera Cohort Study. We would also like to thank Sudhaunshu Joshi and Gillian Mbambo for their assistance in culturing parasites and other lab experiments. We would also like to acknowledge Karl Seydel, Amed Ouattara, and Andrea Buchwald for their valuable input and suggestions, and Don Mathanga and Terrie Taylor for their role in leading the Malawi International Center of Excellence for Malaria Research.

References

1. Auburn S, Barry AE. Dissecting malaria biology and epidemiology using population genetics and genomics. Int J Parasitol. 2017;47:77–85.

2. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. Nature. 2012;487:375–9.
3. Agrawal S, Moser KA, Morton L, Cummings MP, Parihar A, Dwivedi A, et al. Association of a novel mutation in the *Plasmodium falciparum* chloroquine resistance transporter with decreased piperaquine sensitivity. J Infect Dis. 2017;216:468–76.
4. Takala SL, Coulibaly D, Thera MA, Batchelor AH, Cummings MP, Escalante AA, et al. Extreme polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development. Sci Transl Med. 2009;1:2ra5-2ra5.
5. Dwivedi A, Reynes C, Kuehn A, Roche DB, Khim N, Hebrard M, et al. Functional analysis of *Plasmodium falciparum* subpopulations associated with artemisinin resistance in Cambodia. Malar J. 2017;16:493.
6. Venkatesan M, Amaratunga C, Campino S, Auburn S, Koch O, Lim P, et al. Using CF11 cellulose columns to inexpensively and effectively remove human DNA from *Plasmodium falciparum*-infected whole blood samples. Malar J. 2012;11:41.
7. Auburn S, Campino S, Clark TG, Djimde AA, Zongo I, Pinches R, et al. An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. PLoS One. 2011;6:e22213.
8. Oyola SO, Gu Y, Manske M, Otto TD, O'Brien J, Alcock D, et al. Efficient depletion of host DNA contamination in malaria clinical sequencing. J Clin Microbiol. 2013;51:745–51.
9. Oyola SO, Ariani CV, Hamilton WL, Kekre M, Amenga-Etego LN, Ghansah A, et al. Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification. Malar J. 2016;15:797.
10. Melnikov A, Galinsky K, Rogov P, Fennell T, Van Tyne D, Russ C, et al. Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biol. 2011;12:R73.
11. Cohen-Karni D, Xu D, Apone L, Fomenkov A, Sun Z, Davis PJ, et al. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. Proc Natl Acad Sci USA. 2011;108:11040–5.
12. Baumgarten S, Bryant JM, Sinha A, Reyser T, Preiser PR, Dedon PC, et al. Transcriptome-wide dynamics of extensive m6A mRNA methylation during *Plasmodium falciparum* blood-stage development. Nat Microbiol. 2019;4:2246–59.
13. Bousema T, Okell L, Felger I, Drakeley C. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. Nat Rev Microbiol. 2014;12:833–40.
14. Trager W, Jensen JB. Human malaria parasites in continuous culture. 1976. J Parasitol. 2005;91:484–6.
15. Zainabadi K, Adams M, Han ZY, Lwin HW, Han KT, Ouattara A, et al. A novel method for extracting nucleic acids from dried blood spots for ultrasensitive detection of low-density *Plasmodium falciparum* and *Plasmodium vivax* infections. Malar J. 2017;16:377.

16. Moser KA, Drábek EF, Dwivedi A, Stucke EM, Crabtree J, Dara A, et al. Strains used in whole organism *Plasmodium falciparum* vaccine trials differ in genome structure, sequence, and immunogenic potential. Genome Med. 2020;12:6.
17. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics. 2014;30:1292–4.
18. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
19. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
20. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1-33.
21. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
22. Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. Genome Res. 2016;26:1288–99.
23. Chan ER, Menard D, David PH, Ratsimbasoa A, Kim S, Chim P, et al. Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium vivax*. PLoS Negl Trop Dis. 2012;6:e1811.
24. Cowell AN, Loy DE, Sundararaman SA, Valdivia H, Fisch K, Lescano AG, et al. Selective whole-genome amplification is a robust method that enables scalable whole-genome sequencing of *Plasmodium vivax* from unprocessed clinical samples. mBio. 2017;8:e02257-16 .
25. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, et al. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. Nat Commun. 2016;7:11078.
26. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains. Nat Commun. 2014;5:4052.
27. Buchwald AG, Sixpence A, Chimenya M, Damson M, Sorkin JD, Wilson ML, et al. Clinical implications of asymptomatic *Plasmodium falciparum* infections in Malawi. Clin Infect Dis. 2019;68:106–12.

Table

Additional File

Additional file 1:

File name: AdditionalFile1.pdf

Title: AdditionalFile1

Table 1. Sequencing statistics in samples that underwent direct sequencing versus sWGA.

	Direct Sequencing (n = 3)	sWGA (n = 3)
Total reads sequenced	30,878,334	33,586,527
Reads mapped to *P. falciparum* (%)	92.21	95.52
Genome with 5x coverage (%)	98.68	95.53
Mean coverage depth	196x	213x
Total SNPs	32,775	22,355
F_{WS}	0.068 ± 0.009	0.116 ± 0.015

Description: Supplementary Material.

Figures

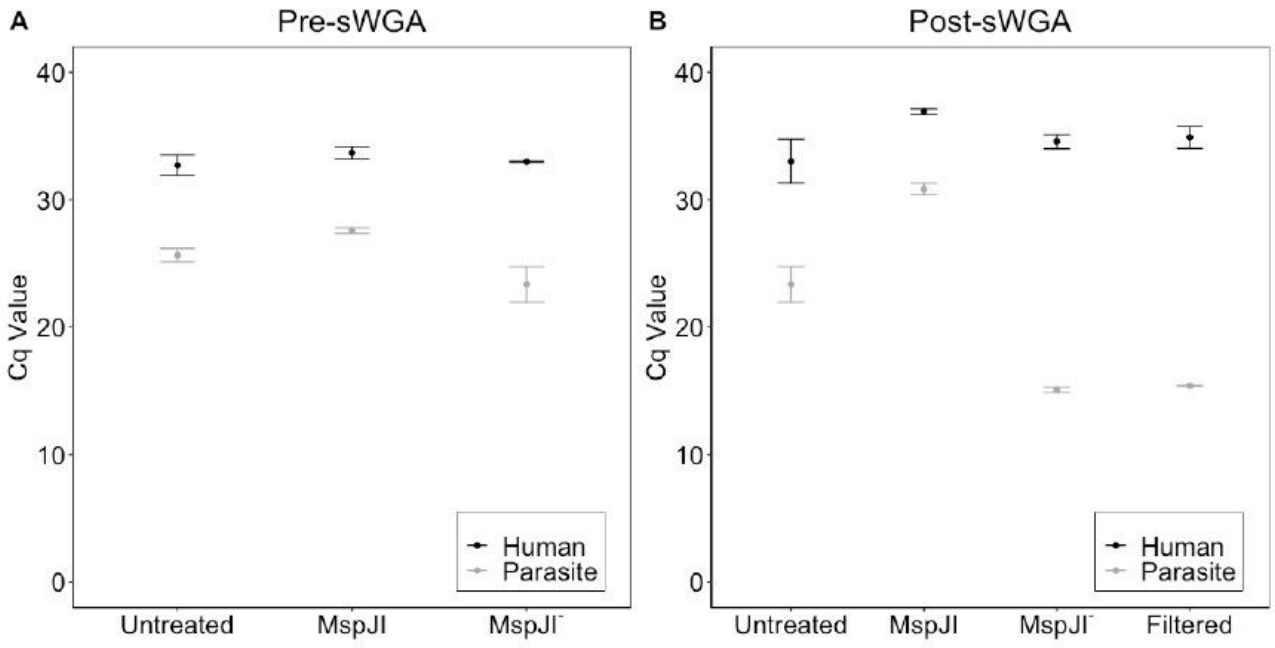

Figure 1

Figure 1

Effect of MspJI and sWGA treatments on parasite DNA concentration. (A) Human and P. falciparum Cq values prior to sWGA on samples with 10,000 parasites/μL (n = 3). Cq value indicates the number of cycles required to detect a signal, where higher Cq values indicate lower DNA concentrations. (B) Human and P. falciparum Cq values after sWGA on samples with 10,000 parasites/μL (n = 3).

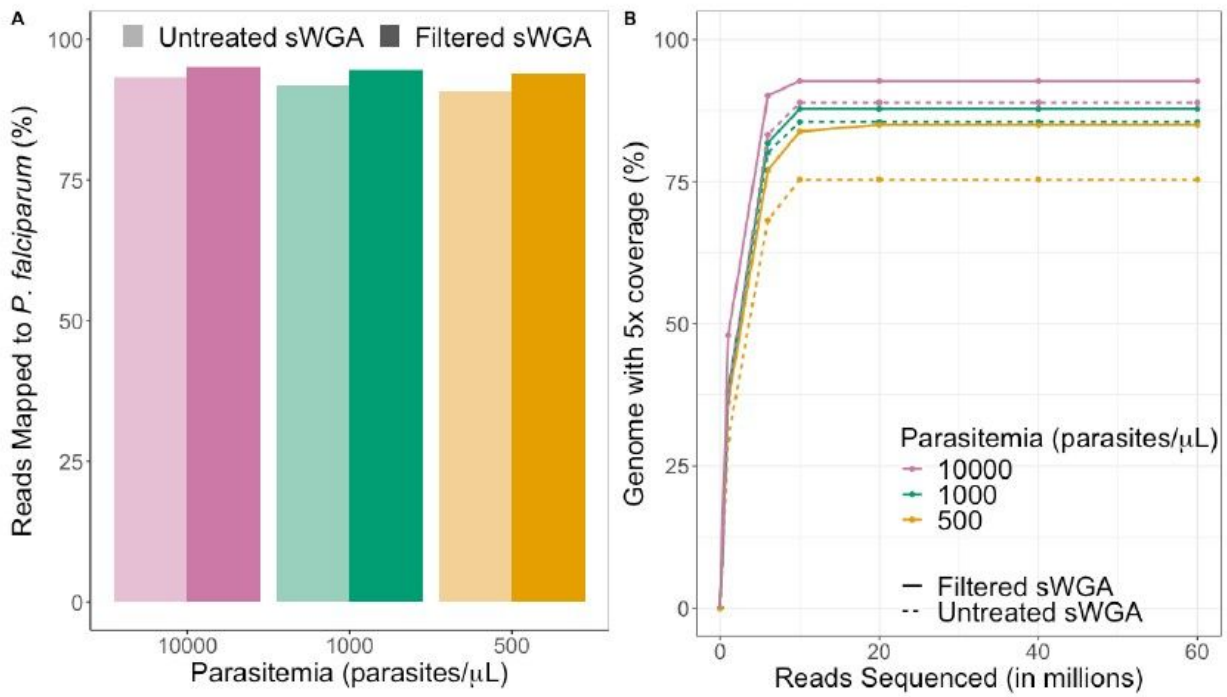

Figure 2

Figure 2

P. falciparum genome coverage in filtered and untreated samples that underwent 424 sWGA. (A) The percentage of reads that mapped to the P. falciparum 3D7 reference are shown 425 for filtered and untreated samples with different parasitemias that underwent sWGA prior to 426 sequencing. (B) Percentage of the P. falciparum 3D7 genome with at least 5x coverage is shown 427 relative to the number of reads sequenced (in millions) in filtered and untreated samples of 428 different parasitemias that underwent sWGA prior to sequencing.

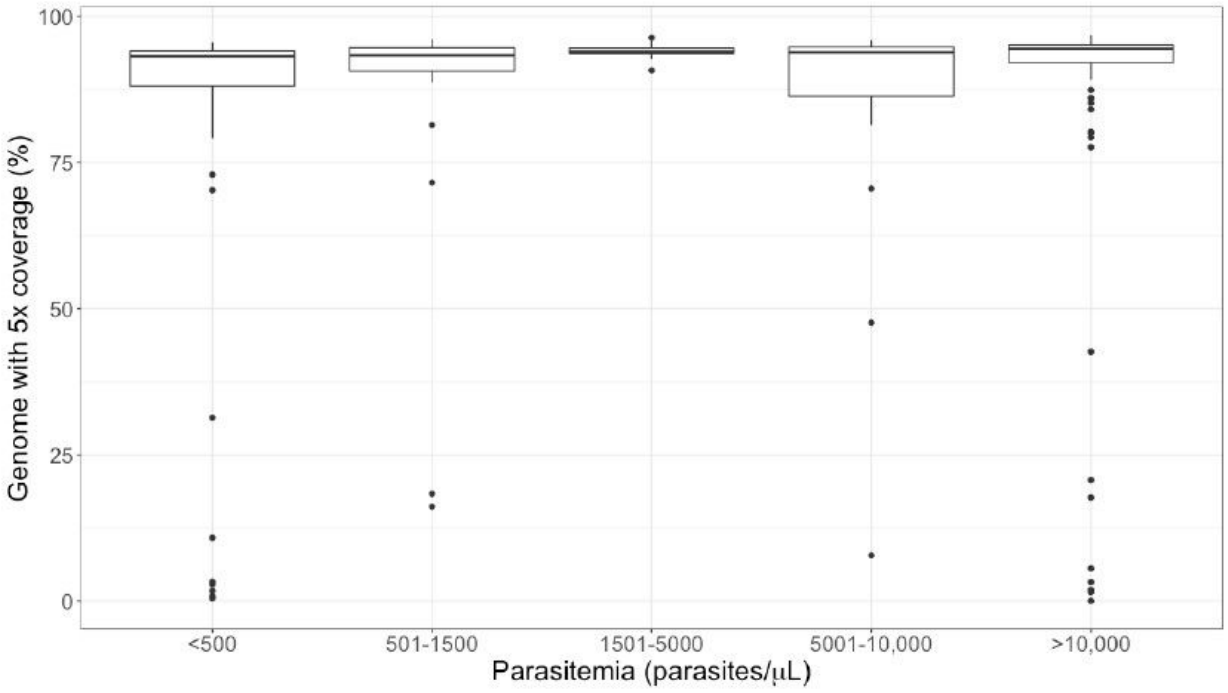

Figure 3

Figure 3

Percent of genome with 5x coverage in whole genome sequences of field isolates 431 with different parasitemias.

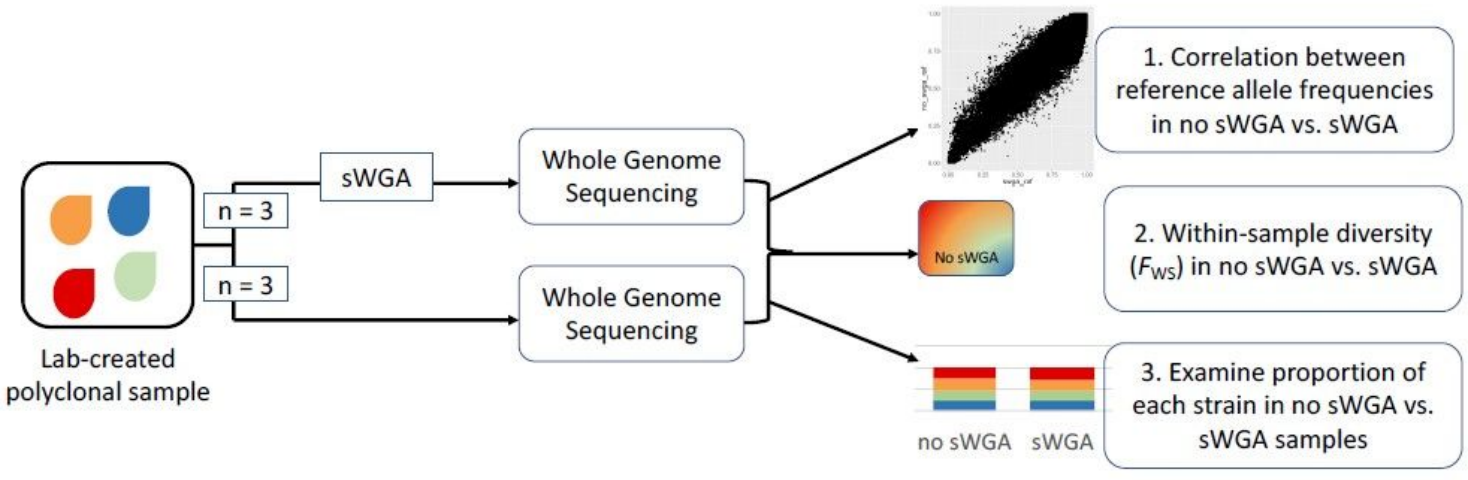

Figure 4

Figure 4

Schematic of experimental design to evaluate potential amplification bias in samples 434 undergoing sWGA.

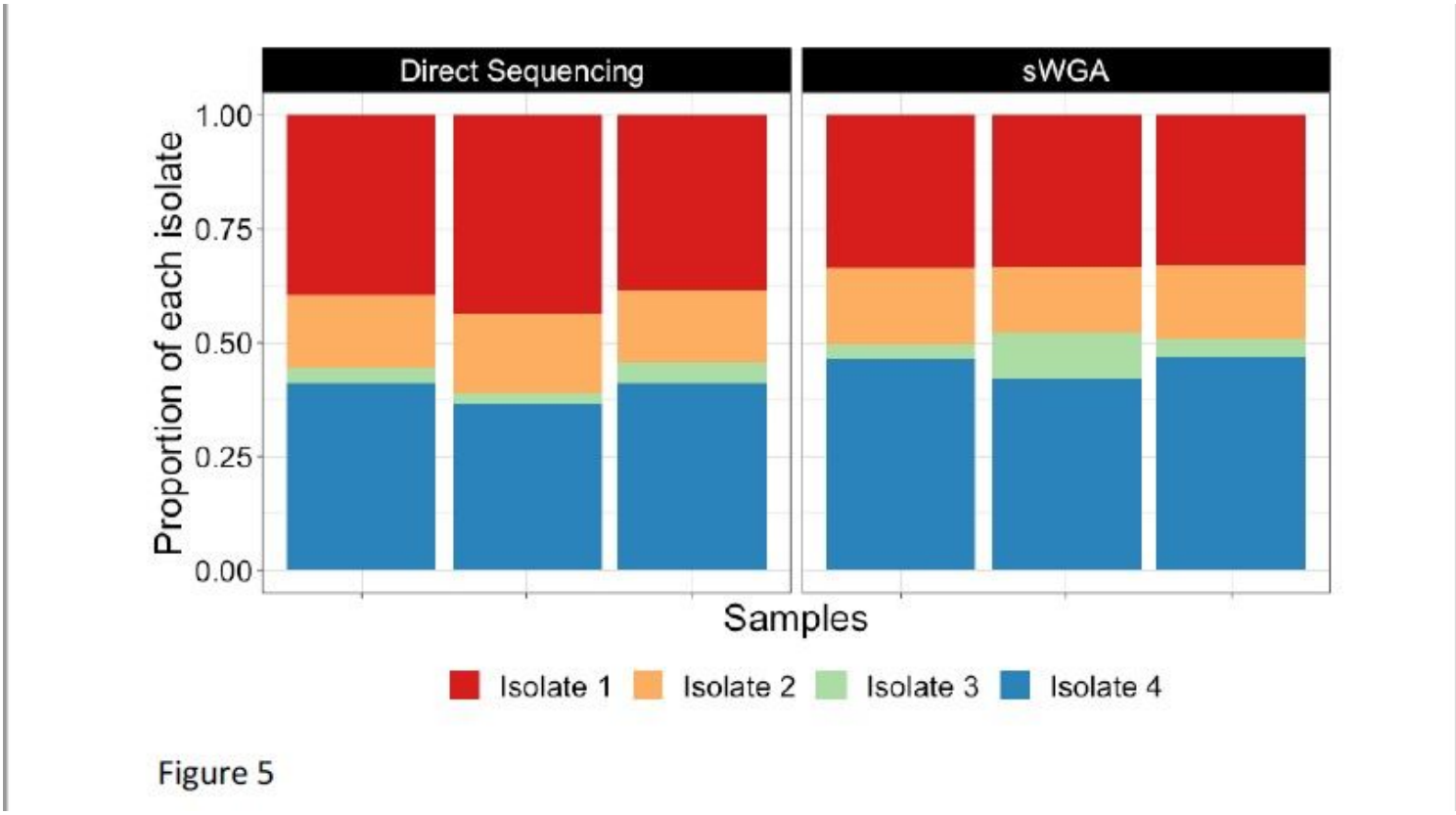

Figure 5

Proportion of isolate-specific SNPs in mixtures that underwent sWGA prior to 437 sequencing or were directly sequenced.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFilerevision1final.pdf