**Evaluation of a pooled targeted amplicon deep sequencing (TADS) method based on parasitemia levels for molecular surveillance of *P. falciparum* genes *dhfr* and *mdr1* in Haiti**

Camelia Herman 1,3, Dhruviben Patel 2,3, Julia Kelley 1,3, Janani Guru4, Michelle A. Chang 3, Udhayakumar Venkatachalam3, Jean F. Lemoine 5, Eric Rogier 3, Eldin Talundzic 3

1Centers for Disease Control and Prevention Foundation, Atlanta, GA, USA

2 Williams Consulting LLC, Atlanta, GA, USA

3 Malaria Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA

4The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

5 Programme National de Contrôle de la Malaria, MSPP, Port-au-Prince, Haiti

\*Corresponding author

Email: wrj5@cdc.gov (Eldin Talundzic, PhD)

**Abstract**

Sequencing large numbers of individual samples, which is often needed for countrywide surveillance-based studies, remains expensive despite the decreasing costs of next generation sequencing (NGS). Pooling of several individual samples is potentially an alternative cost and time effective method. Using 100 individual *P. falciparum* samples from a nationwide Haiti drug resistance surveillance study, we compared allele frequency (AF) calls for *dhfr* and *mdr1* for individual and pooled deep sequenced samples. Samples with similar real-time PCR cycle threshold (ct) values (+/- 1.0 ct value) were combined into 1:10 samples per pool. The CT value was used to estimate parasite DNA in each individual sample prior to PCR, NGS library preparation and targeted deep amplicon sequencing (TADS) of the *dhfr* and *mdr1* genes. In both tests, six of the pooled groups had the expected allele frequency (AF) ratios (e.g., individual AF added together/total pool = expected pooled AF). The groups with CT values ≥34.0, equivalent to 100 parasites per uL/blood, showed discordant AF results for *dhfr* and *mdr1* between individual and pooled samples. The discordance in AF estimation for these groups may have been due to low amounts of parasite DNA, which can lead to variable PCR amplification efficiencies. Grouping samples with CT values < 34.0, or an estimated more than 100 parasites per uL/blood, prior to pooling and deep sequencing yielded the expected population level AF for *dhfr* and *mdr1* genes. Pooling samples based on CT values and/or parasitemia prior to deep sequencing provides a more accurate estimate of allele frequencies within *P. falciparum* populations and can be used for rapid genotyping of large number of samples for drug resistant markers in population-based studies.

**Introduction**

Hispaniola is the only island in the Caribbean with reported malaria cases, and 21,430 confirmed malaria cases were reported in Haiti in 2016. {Druetz:2018fw}{Vincent:2018fq} The primary vector transmitting *Plasmodium falciparum* on the island is *Anopheles albimanus.* {WHO:2019ub}and approximately 98.5% of malaria in Haiti is *Plasmodium falciparum*, with *Plasmodium vivax* and *Plasmodium malariae* reported sporadically. {WHO:2019ub}With Haiti’s national goal of interruption of local transmission of malaria by 2025, scale up field diagnostics, and intervention efforts has recently occurred. {Lemoine:2017cp, Rogier:2020kr}

Currently, chloroquine (CQ) with a single dose of primaquine is the primary treatment for uncomplicated *P. falciparum* malaria in the country with second line treatment as sulfadoxine-pyramethamine (SP). {Druetz:2018fw} CQ remaining efficacious against malaria in Haiti and offers many advantages as the country moves towards malaria elimination. {WHO:2019ub, Rogier:2020kr} Routine monitoring of the efficacy of the primary and secondary antimalarial treatment is a public health priority for national malaria control programs. {WHO:2019ub} Additionally, SP mass-drug administration (MDA) may prove useful for Haiti to clear the parasite reservoir from the population.

Health facility sentinel sites currently collect samples for antimalarial drug-resistance genotyping using Sanger sequencing in Haiti, and to date, no molecular markers for CQ resistance have been detected in the country. {Lemoine:2017cp, Rogier:2020kr} Along with the wild-type CQ genotype, only one triple mutant was found for the pyrimethamine resistance gene *dhfr,* though 46% (257/548) had the S108N point mutation {Rogier:2020kr} Additionally, only a single isolate was found with a mutation A437G for the sulfadoxine resistance gene *dhps*. {Rogier:2020kr}

While Sanger sequencing was used previously in drug resistance monitoring in Haiti, future work is exploring the utility of using pooled Targeted Deep Amplicon Sequencing (TADS) for routine population level surveillance. Pooling of individual samples prior to next generation sequencing can provide a more effective strategy for estimation of allele frequencies (AFs). {Taylor:2013fa} However, SNP discovery and detection using pooled targeted deep amplicon sequencing (TADS) can be accompanied by higher probabilities of sequencing errors, unless read coverage depth used to call SNPs is quite high {Cutler:2010ha} In comparison, TADS of individual samples has lower probabilities of false AF calls{Cutler:2010ha}, but is limited to a set number of samples (to obtain sufficient read coverage) and can be prohibitively expensive for large scale population based studies. Thus, pooling can be less efficient when coverage is low and individual sequencing is less efficient with higher expected coverage {Cutler:2010ha}.

To determine whether *P. falciparum* parasitemia levels also have an effect on final AF estimation, 100 samples from a nationwide Haiti drug resistance surveillance study were sequenced using two TADS methods: individual sequenced samples vs. pooled sequenced samples. Specifically, the previously published individual patient sample TADS method {Talundzic:2018jl} was modified to a pooling strategy based on parasitemia levels of ten samples per pool (10:1). Our results suggest that pooling samples based on estimated parasitemia levels prior to TADS can improve final AFs estimations in parasite populations. However, at lower parasitemia levels (less than 40 parasites per uL) pooled TADS may over/underestimate the true AFs of drug resistant mutations. In conclusion, pooling samples based on parasitemia levels, as determined by real-time PCR, prior to sequencing should be considered as an additional strategy for pre-screening or genotyping of large number of samples for molecular markers of resistance.

**Materials and Methods**

**Human subjects.** Project was approved by Haitian IRB and not considered human subjects research by CDC Human Subjects Office (2015-155). Dried blood on filter paper was collected from March 2016- December 2017 from five health facilities in Haiti from persons testing positive to malaria by rapid diagnostic test (RDT, Carestart Pf product info). Patients all ages were enrolled after consenting for testing for molecular markers of malaria. Whole blood was collected by fingerprick on a Whatman 903 Protein Saver cards (GE Healthcare). Each filter paper was air-dried overnight, stored in a desiccant containing bag at 4 degrees until shipment to CDC in Atlanta, GA.

**DNA extraction.** Two 6mm punches from dried blood spots (DBS) were used to extract genomic DNA using the Qiagen DNA extraction kit following the manufacturer’s instructions (QIAGEN, Valencia, CA). The DNA was eluted in 150uL of elution buffer aliquoted and stored at -20°C until use.

**Parasitemia estimation using PET-PCR.** Following DNA extraction, PET-PCR was performed as described previously on 100 samples. {Talundzic:2014cj, Lucchi:2013gl} Cycle threshold (Ct) values below 40 considered to be parasitemia positive and above 40 negative. The 3D7 *P. falciparum* control sample was used in a one- and two-fold dilutions (100,000 p/uL to 0.1 p/uL and 50 p/uL to 1.56 p/uL, respectively) to estimate parasitemia levels based on PET-PCR CT values for reach individual sample.

**PCR enrichment of *dhfr* and *mdr1.*** PCRs were performed to amplify the full-length *P. falciparum dhfr* and *mdr1* genes for 100 individual DBS patient samples, including the *P. falciparum* 7G7 and HB3 *P. falciparum* strains. The New England BioLabs (NEB) High Fidelity PCR kit (New England BioLabs, USA; catalog no. 51104) was used to amplify the genes according to the manufacturer's instructions with a 50.0-μl master mix preparation using the 5× GC buffer.

**Individual patient and pooled patient targeted amplicon deep sequencing (TADS).** The full-length *dhfr* and *mdr1* genes for the 100 Haitan samples were sequenced using the previously described Malaria Resistance Surveillance (MaRS) TADS protocol {Talundzic:2018jl}. For individual patient level sequencing, unique sequence indices were added to the pooled PCR amplicons, prior to pooling all the samples based on patient IDs, and sequenced on a single run. For the pooled level sequencing, similar parasitemia values (e.g, +/- 1.0 Ct values, Table 1) were used to pool individual patients in pools of 10:1 prior to adding unique sequence indices, and then re-sequenced twice, in two separate runs. Illumina-supported sequencing adaptors and unique sequence indices were added to individual patients or pooled patient lots using the Illumina Nextera XT kit (Illumina; catalog numbers FC-131-1096 and FC-131-1002). This step generates a unique sequence barcode identifier (ID) for each individual patient sample or pooled sample lot, including the pooled *dhfr* and *mdr1* genes. The Nano MiSeq Reagant Kit V2 250 by 250 base pair kit (Illumina; catalog numbers FC-102-2003) was used for all three sequencing runs.

**Data analysis and visualization.** The next-generation sequencing analysis toolkit (NeST) (<https://github.com/CDCgov/MaRS>) was used to call non-synonymous single nucleotide polymorphisms (SNPs) in dhfr and mdr1 genes using the same quality threshold cut-offs for the individual and pooled sequenced samples. All SNPs were also visually confirmed using the Geneious Prime software ([www.geneious.com](http://www.geneious.com)). Data visualization (figure generation) was performed using the python seaborn 0.10.1 package (https://seaborn.pydata.org).

**Individual vs. pooled comparison.** All the individual samples were sequenced using exactly the same procedure used to process and sequence the pooled samples. Individual sequenced samples allele frequency (AF) calls for *dhfr* and *mdr1* were used to determine the expected AF of the 10:1 pooled samples as follows: individual AF per sample were added together and then divided by 10 to estimate the expected pooled AF. Importantly, all samples were mono infections and only major AF (>99%) in *dhfr (*N51I, C59R and S108N)and *mdr1* (Y184F, S1034C, N1042D, and D1246Y) were used for this analysis. Samples were combined based on estimated parasitemia levels prior to pooling and sequencing to limit the possibility of a single sample AF calls biasing the expected AF of the pooled samples.

**Results**

From the individual sample sequenced method, high quality reads for 88% (88/100) *dhfr* and 70% (70/100) *mdr1* samples were obtained, Figure 1. In comparison, for the pooled sequenced samples high quality reads were obtained for 9/10 (90%) *dhfr* and 7/10 (70%) *mdr1* samples, Figure 1. Both sequencing approaches had comparable sequencing success rates at greater than one parasites per uL (*dhfr:* greater than 1; *mdr1*: greater than 3), Figure 1. The median read depth coverage for *dhfr* drug resistant associated mutation N51I, C59R and S108N had a 3-fold decrease (median: 280 to 88) from individual to pooled sequenced samples, Figure 2. In comparison, the *mdr1* Y184F, S1034C, N1042D, and D1246Y had 0.2-fold decrease in coverage (median: 79 to 62), Figure 2. The overall coverage distribution for individual sequenced samples was wider (*dhfr:* 10 – 720; *mdr1*: 10 – 460) than the pooled samples (*dhfr:* 10 – 290; *mdr1*: 10 – 110). Comparison of AF estimates between the individual and pooled sequenced samples as a measure of parasitemia indicates that at lower parasitemia levels (less than 40 parasites per uL) the AF of pooled sequenced samples are not accurately estimated when compared to the AFs calls of individual sequenced samples (e.g., expected AF calls), Figure 3.

**DISCUSSION**

Pooling of malaria samples prior to TADS has been shown to be a cost-effective approach for estimating drug resistance associated mutations in *P. falciparum*.{Taylor:2013fa, Brazeau:2019jy, Ngondi:2017bm, Taylor:2015iw} One strategy is to combine dried blood spots (DBS) from individuals prior to nested PCRs and TADS {Ngondi:2017bm}, while another strategy involves combining of genomic DNA (gDNA) in varying pool ratios (i.e. 35 to 273) prior to TADS. {Taylor:2015iw} However, both these strategies can result in biased AFs due to varied parasite densities within individual samples without properly achieving equimolar concentrations of individual DNA in the pools. This is especially the case with alleles present at low frequency in pools that can lead to false positive variants. {Anand:2016jz}

Here, in an effort to minimize some of these potential biases due to parasite densities, we evaluate yet another pooling strategy by using parasite densities to guide the pooling of DBS derived gDNA in 10:1 ratio’s prior to TADS. The 10:1 ratio was chosen to limit further biases introduced to AF due to library preparation chemistries {Rhodes:2014dz} and reduction in sequencing coverage, with the latter having the most influence on accurate AF estimation.{Kofler:2016kn} By grouping ten samples per pool, the total experiment remains cost efficient while also providing further flexibility in overall study design (e.g., such as further grouping by treatment arm or regional site). {Zilinskas:2014gj} By using this strategy, sufficient sequence coverage can be obtained consistently needed to have reliable AF estimation in pools across different full-length drug resistance malaria genes – allowing for identification of novel SNPs in a population which subsequently can be further subject for individual patient level sequencing {Talundzic:2018j} Interestingly, our study also provides evidence that AF estimation in pools can be under or over represented at low parasite levels (less than 40 parasites per uL), Figure 3.

While the added step of having to use a real-time PCR assay to estimate parasitemia levels in samples may seem like a limitation of this strategy, it can serve as an important quality control step to assess DNA quality and confirm or identify accurately the *Plamsmodium* spp. in samples {Talundzic:2014cj}{Akerele:2017bc}. Using a reliable and accurate DNA quantification method prior to sequencing will also improve overall sequencing capacity and further reduce costs – by avoiding either under or over usage of NGS capacity.{Robin:2016ks} Loading of unbalanced indexed DNA libraries, commonly the case with pooling based strategies, may result in unnecessary data loss and further SNP estimation based errors {Robin:2016ks} Thus, the added step of implementing DNA quantification step prior to pooling is an important part of an effective quality control system (QMS) for NGS based surveillance tests. {Hutchins:2019ia}

In order for population level data collected over different periods of time to be useful for meat-analysis, ideally the same laboratory pipeline (e.g., pooling strategy, amplification, library preparation and sequencing chemistries) and analysis should be used. This is especially important in malaria pre-elimination settings or routine therapeutic efficacy studies being conducted on a regular basis across different geographical regions in support of National Malaria Control Programs. Thus, the trade-off between cost savings and quality of data – especially reproducible data over time and space – should be considered when planning an NGS based pooling strategy for molecular surveillance of drug resistance genes.

One of the main limitations of most pooling strategies, including the work described herein, is the inability of identifying accurately reads that belong to individual patients. However, using first a pooling strategy to identify possible SNP signals of interest, especially novel SNPs that might be associated with drug resistance, followed by individual patient level sequencing can be a good strategy for confirming drug resistance in particular samples while keeping overall costs efficient.

**[@ ERIC – can you add here now part about how this could be useful for Haiti moving forward.]**