Supporting Information

Juliano et al. 10.1073/pnas.1007068107

SI Materials and Methods

Nested PCR Genotyping of *msp1* and *msp2*. For the Malawian samples, genomic DNA was extracted from the blood spots using a QIAamp DNA mini kit (Qiagen). Nested PCR of *msp1* and *msp2* was carried out following the recommended genotyping protocols described by the WHO (http://apps.who.int/malaria/docs/drugresistance/RGPText_STI.pdf). nPCR results of the Cambodia samples have previously been reported (1, 2).

Amplicon Preparation for Pyrosequencing. PCR products for 454 sequencing were prepared by using fusion primers specific for three regions of the malaria genome: (*i*) the block 2 region of *msp1*, (*ii*) the central variable region of *msp2*, and (*iii*) the segment of *dhfr* containing aa 51, 59, and 108. Fusion primers consisted of the following structures: forward fusion primer (454 linker A-4 nucleotide barcode-forward PCR primer) and reverse fusion primer (454 linker B- reverse PCR primer; Table S2). By including a 4-nt bar code in the forward fusion primer, we were able to sequence multiple samples in the same reaction (3). The PCR primers for *msp1* were based on previously reported primers (4), whereas the *msp2* and *dhfr* primers were designed using Primer3 (5). Analysis of *glurp* was not done because the length of the variable region exceeded the expected amplicon length capable by 454.

The ability of the primers to amplify all *msp1* (K1, MAD20, and RO33) families and both *msp2* families (FC27 and 3d7) was confirmed by amplification of and correct identification of 10 laboratory stocks of genomic DNA representing these families. Genomic DNA from MR-4 of the following genotypes was used: K1 (MRA-159G, contributed by D. E, Kyle), V1/S (MRA-176G, contributed by D. E. Kyle), FCR3-Gam (MRA-731G, contributed by W. Trager), Dd2 (MRA-150G, contributed by D. Walliker), FCR-C5 (MRA-699G), ITG-2G2 (MRA-326G, contributed by L. H. Miller), 7g8 (MRA-152G, contributed by D. Walliker), D10 (MRA-201G, contributed by Y. Wu), HB3 (MRA-155G, contributed by T. E. Wellems, National Institutes of Health, Bethesda, MD), 3d7 (MRA-102G, contributed by D. J. Carucci), RO33 (MRA-200G, contributed by D. Walliker), and HB3-B2 (MRA-149G, contributed by D. Walliker).

PCR reactions were carried out in 50-μL volumes containing 5 μL of FastStart High Fidelity reaction buffer with 18 mM Mg (Roche), 1 μL 10 mM dNTP mix (Promega), 0.5 μL FastStart High Fidelity enzyme blend (5 U/μL; Roche), 400 nM forward primer (MWG Operon), 400 nM reverse primer (MWG Operon), 5 μL of sample DNA, and ddH_2O. All amplifications were performed by using a Mastercycler EP (Eppendorf). Primers and PCR conditions are summarized in Table S1.

The PCR products were purified by using the Purelink PCR purification kit, using the HC buffer to remove long PCR primers (Invitrogen). Purified PCR product was checked for an OD ratio of 1.8 or greater and to determine the concentration of PCR product using a Nanodrop 1000 spectrophotometer (Thermo Scientific). Up to 500 ng of PCR product was dried using a SpeedVac concentrator (Thermo Scientific). The dried PCR product from either two (Cambodia) or four (Malawi) patients was resuspended in ddH_2O and submitted for sequencing on 1/16th of a picotiter plate.

Interpretation of Sequencing Results. The PCR amplicons were sequenced at the University of North Carolina High-Throughput Sequencing Facility on a 454 Life Sciences sequencer by using GS FLX Titanium chemistry (Roche). This analysis occurred before the publication of Titanium chemistry amplicon protocols. We elected to use Titanium chemistry to achieve higher read lengths than were available in the FLX chemistry at the time. Because of the difference in protocols, a linker DNA fragment complementary to the linker sequence on the amplicon and the Titanium linker on the bead was used to bind the PCR product to the bead. The use of titanium chemistry also precluded the use of the GS Amplicon Variant Analyzer software for analysis of the data. The individual strands of PCR product isolated on the beads were sequenced with only the forward primers for all genes studied. We elected to do this as the *msp2* sequences would likely end in the hypervariable region, and therefore quantitative assembly of forward and reverse sequences would be difficult. In addition, by not reverse sequencing the samples, we also increased the absolute number of forward reads.

The output sequences were sent to the University of North Carolina Center for Bioinformatics and the data separated by barcode into the individual patient samples at each site sequenced. To do this, all sequence reads were initially mapped to the 3d7 genome to make sure only malaria reads were included. In addition, to be included in the analysis, sequence reads needed to meet the following criteria: (*i*) length greater than 50 bp and (*ii*) the complete barcode and forward primer sequence intact. The mapped sequences were then sorted by the target gene and the bar code specific to the patient sample using the following Perl code.

```
my $inputfilename = shift;
my $output_AGAG = $inputfilename . ".AGAG.txt";
my $output_ACAC = $inputfilename . ".ACAC.txt";
my $output_TCTC = $inputfilename . ".TCTC.txt";
my $output_TGTG = $inputfilename . ".TGTG.txt";
print "input $inputfilename\n";
print "output_AGAG $output_AGAG\n";
print "output_ACAC $output_ACAC\n";
print "output_TCTC $output_TCTC\n";
print "output_TGTG $output_TGTG\n";
my $primer_seq = "GCCTCCCTCGCGCCATCAG";
my $primer_len = 19;
my $total = 0;
my $total_AGAG = 0;
my $total_ACAC = 0;
my $total_TCTC = 0;
my $total_TGTG = 0;
my $in = Bio::SeqIO->new(-file => "$inputfilename" , '-format'
=> 'Fasta');
open (OUT_AGAG, ">$output_AGAG") or die "cannot open
$output_AGAG\n";
open (OUT_ACAC, ">$output_ACAC") or die "cannot open
$output_ACAC\n";
open (OUT_TCTC, ">$output_TCTC") or die "cannot open
$output_TCTC\n";
open (OUT_TGTG, ">$output_TGTG") or die "cannot open
$output_TGTG\n";
while (my $seq = $in->next_seq()) {
  # find AGAG tagged sequences
  my $sequence = $seq->seq;
  my $primer_tag = $primer_seq . "AGAG";
  my $ind = index($sequence, "AGAG");
  if ($ind >= 0) {
     my $frag = substr($sequence, 0, $ind + 4);
     my $ind2 = index($primer_tag, $frag);
     if ($ind2 >= 0) { # the frag matched the primer + tag
           print OUT_AGAG ">" . $seq->id . "\n";
           print OUT_AGAG "$sequence\n";
           $total_AGAG++;
```

```
        }
    }
    # find ACAC tagged sequences
    my $sequence = $seq->seq;
    my $primer_tag = $primer_seq . "ACAC";
    my $ind = index($sequence, "ACAC");
    if ($ind >= 0) {
        my $frag = substr($sequence, 0, $ind + 4);
        my $ind2 = index($primer_tag, $frag);
        if ($ind2 >= 0) { # the frag matched the primer + tag
                print OUT_ACAC ">" . $seq->id . "\n";
                print OUT_ACAC "$sequence\n";
                $total_ACAC++;
    }
}
    # find TCTC tagged sequences
    my $sequence = $seq->seq;
    my $primer_tag = $primer_seq . "TCTC";
    my $ind = index($sequence, "TCTC");
    if ($ind >= 0) {
        my $frag = substr($sequence, 0, $ind + 4);
        my $ind2 = index($primer_tag, $frag);
        if ($ind2 >= 0) { # the frag matched the primer + tag
                print OUT_TCTC ">" . $seq->id . "\n";
                print OUT_TCTC "$sequence\n";
                $total_TCTC++;
      }
}
    # find TGTG tagged sequences
    my $sequence = $seq->seq;
    my $primer_tag = $primer_seq . "TGTG";
    my $ind = index($sequence, "TGTG");
    if ($ind >= 0) {
        my $frag = substr($sequence, 0, $ind + 4);
        my $ind2 = index($primer_tag, $frag);
        if ($ind2 >= 0) { # the frag matched the primer + tag
                print OUT_TGTG ">" . $seq->id . "\n";
                print OUT_TGTG "$sequence\n";
                $total_TGTG++;
      }
}
     $total++;
}
    print "total = $total\n";
    print "total_AGAG = $total_AGAG\n";
    print "total_ACAC = $total_ACAC\n";
    print "total_TCTC = $total_TCTC\n";
    print "total_TGTG = $total_TGTG\n";
    close OUT;
```

The partial sequences for each gene for each patient were segregated following the workflow outlined in *Materials and Methods* and represented in Fig. S1. Fig. S2*A* shows an representative initial partial *msp2* alignment in BioEdit of one patient's sequences (6). In this panel we can see several single nucleotide indels in areas of high AT richness. The sequences were then divided into groups, in this case four, as seen in Fig. S2*B*. Of note, a group may not completely colocalize on a single screen as a result of insertions or deletions elsewhere in the sequence. These groups were realigned to evaluate if they contained multiple distinct variants. If not, they would proceed to evaluation of sequence homology and divided again based on criteria previous discussed. This was followed by calling of ambiguous bases by the investigator to create a final consensus sequence for each variant in the population. By processing the sequences in the manner described, we lose potential SNPs (Fig. S2*B*) in the variants. In this case, the vast majority of sequences contain a G at that position and only two have a C. This will lead to a potential underestimation of parasite diversity, but will limit the effect of sequencing errors on the results.

SI Results

Determination of Sequencing Error Rate. To determine an error-free rate for 454 sequencing of the malaria genome, we conducted individual ClustalW alignments of 400 randomly selected *dhfr* sequences to the 3d7 genome. The specific types of errors detected by comparing 400 *dhfr* sequence amplicons to the consensus 3d7 genome are shown in Table S1. Among these 400 amplicons, the distribution of the total number of errors in each sequence resembles a Poisson distribution (Fig. S3). A sensitivity analysis among the samples from Malawi was done to compare the number of unique variants detected with the use of 5, 6, 7, 8, and 9 bp as the required difference between variants. The sequences from each sample were aligned using ClustalW. A variant was considered unique if there was an indel of more than 3 bp as described previously (the smallest difference noted between two variants was two 3 bp deletions between the variants). As shown in Table S1, only 0.3% of the errors in sequencing were indels of this size or larger. The sequences that aligned of the same size were evaluated for uniqueness by the differing definitions. The initial cutoff of 5 bp was selected so that more than 90% of the sequences would contain fewer errors (Fig. S3). We compared the differences between the consensus sequences within each sample to determine how many unique consensus sequences would be lost based on the more stringent cutoffs. The data are summarized in Table S3.

Detection of Novel *dhfr* Polymorphisms. We sequenced the region of *dhfr* that spanned the known drug resistance alleles at aa 51, 59, and 108 in the patient samples from Malawi. More than 85% of the reads achieved lengths that allowed for assessment of all three amino acids, and therefore a drug-resistant haplotype and allele frequency could be determined. The triple mutation in *dhfr* has reached fixation in Malawi; therefore, the vast majority of sequences contained the classic drug resistance mutations at aa 51, 59, and 108 (7). However, one patient did contain a variant with a unique nonsynonymous mutation at aa 51 (with classic drug-resistant mutations at aa 59 and 108; GenBank accession no. HM153165). Fifty-one sequencing reads contained an A-to-C mutation instead of the classic A-to-T mutation at this site.

Correlation of Genotyping Methods. It has been previously shown that nested PCR often fails to detect minority populations (8, 9). For this reason, we evaluated the correlation between nPCR results and sequencing results using only variants detected with more than 100 reads used to make the consensus (or approximately 7% of the population on average). For *msp1* among the Malawian patient samples, the R^2 value was 0.521. For *msp2* among the same patients, the R^2 value was 0.3465. The samples from Malawi had also previously been genotyped using a *msp1* heteroduplex tracking assay (10). Using these previously reported results, we evaluated the correlation of the number of variants detected between HTA and the total number of variants detected by MPP. This showed an R^2 value of 0.5399 (Fig. S4).

1. Juliano JJ, et al. (2009) Misclassification of drug failure in Plasmodium falciparum clinical trials in southeast Asia. *J Infect Dis* 200:624–628.
2. Rogers WO, et al. (2009) Failure of artesunate-mefloquine combination therapy for uncomplicated Plasmodium falciparum malaria in southern Cambodia. *Malar J* 8: 10.
3. Hoffmann C, et al. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 35:e91.
4. Ferreira MU, et al. (1998) Allelic diversity at the merozoite surface protein-1 locus of Plasmodium falciparum in clinical isolates from the southwestern Brazilian Amazon. *Am J Trop Med Hyg* 59:474–480.

5. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386.
6. Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Ser* 41: 95–98.
7. Nkhoma S, Molyneux M, Ward S (2007) Molecular surveillance for drug-resistant Plasmodiumfalciparum malaria in Malawi. *Acta Trop* 102:138–142.
8. Juliano JJ, Kwiek JJ, Cappell K, Mwapasa V, Meshnick SR (2007) Minority-variant pfcrt K76T mutations and chloroquine resistance, Malawi. *Emerg Infect Dis* 13:872–877.
9. Liu S, Mu J, Jiang H, Su XZ (2008) Effects of Plasmodium falciparum mixed infections on in vitro antimalarial drug tests and genotyping. *Am J Trop Med Hyg* 79:178–184.
10. Kwiek JJ, et al. (2007) Estimating true antimalarial efficacy by heteroduplex tracking assay in patients with complex Plasmodium falciparum infections. *Antimicrob Agents Chemother* 51:521–527.

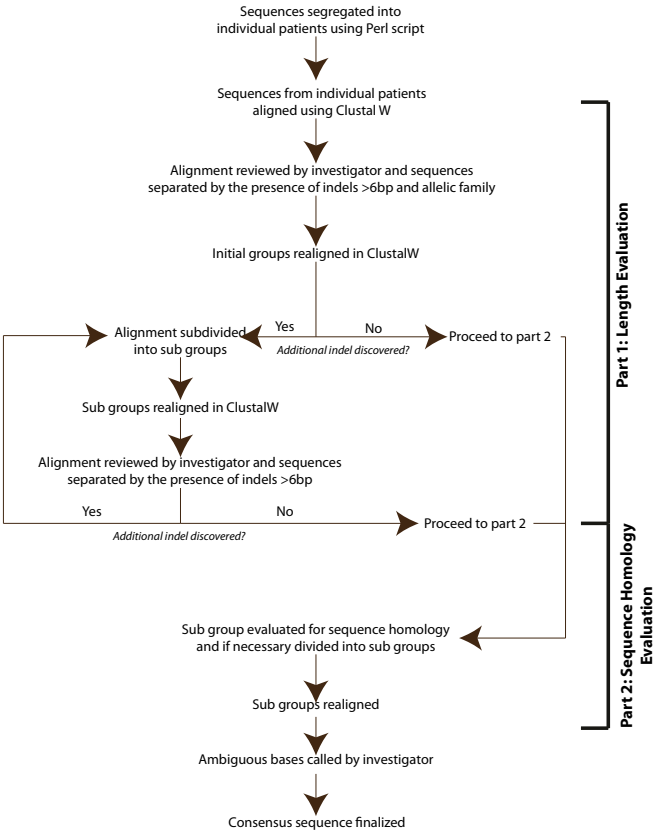

Fig. S1. Workflow of sequence analysis by repetitive alignment.

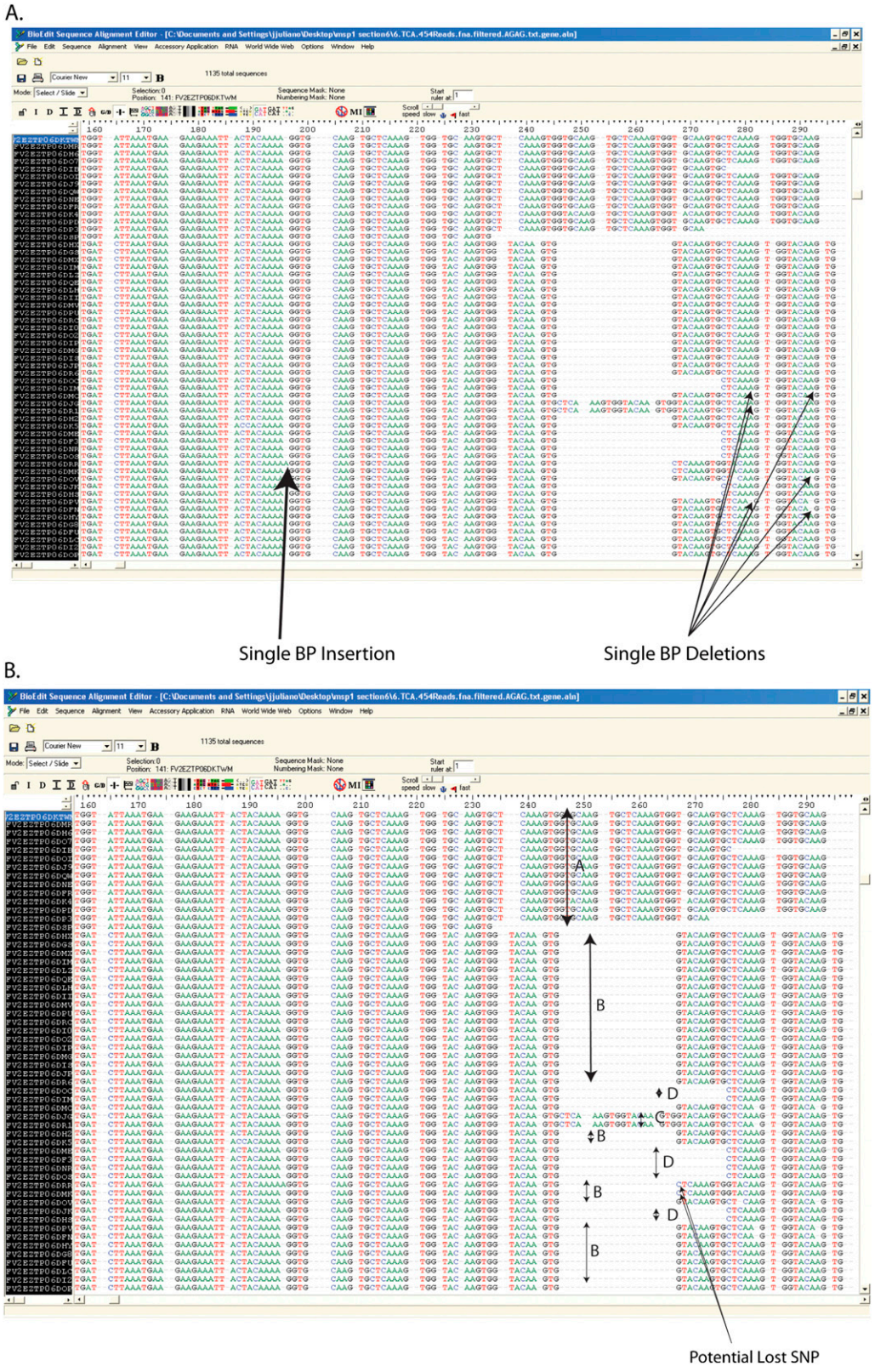

Fig. S2. Interpretation of sequencing results. (*A*) Initial alignment after separation of sequences of an individual patient. Within these sequences, several single nucleotide indels are seen (arrows). (*B*) The method by which sequences would be divided into groups for realignment as described in Fig. S1. After the repetitive alignment was completed, a consensus sequence was made. Ambiguous nucleotides were called by the investigator. In the case of Group B, a single nucleotide polymorphism (marked by the arrows labeled Potential Lost SNP) would possibly be lost. In this case, the majority of sequencing reads contain a G, which would be represented in the final consensus sequence, while the two marked reads contain a C.

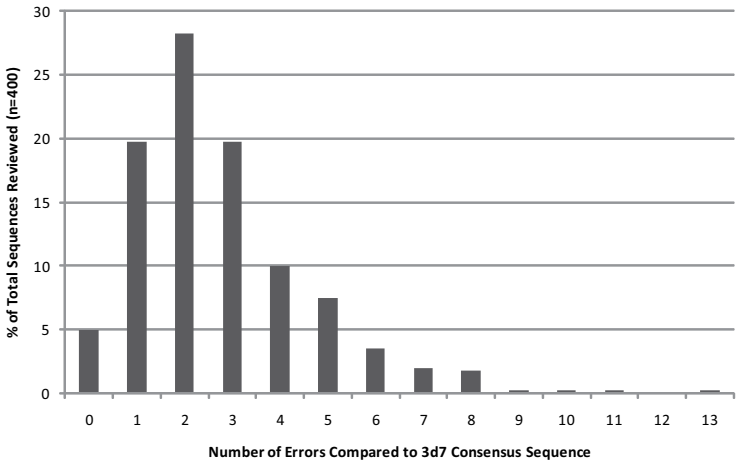

Fig. S3. Distribution of the total number of errors seen in 400 *dhfr* sequence amplicons when compared with the 3d7 consensus sequence.

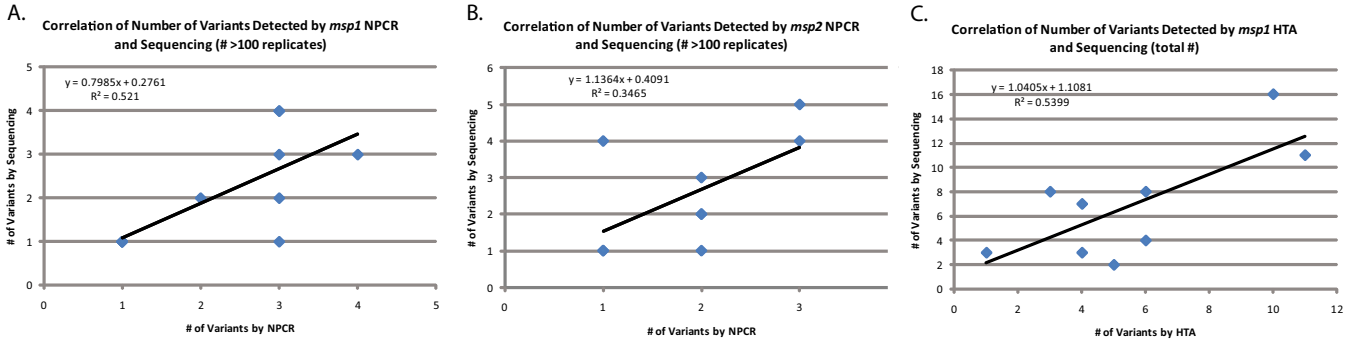

Fig. S4. Correlation of genotyping methods. Correlation between the number of *msp1* (*A*) and *msp2* (*B*) variants detected by nPCR and the number of variants with greater than 100 sequencing replicates by MPP in the same patient. (*C*) Correlation between previously reported *msp1* HTA genotyping results and the total number of variants detected by MPP.

Table S1. Types and numbers of errors in sequences

Error type	bp in errors	Errors (%)
1 bp		
Insertion	459	459 (41.4%)
Deletion	596	596 (53.8%)
2 bp		
Insertion	20	10 (0.9%)
Deletion	20	10 (0.9%)
>2 bp		
Insertion	11	3 (0.3%)
Deletion	0	0
Ambiguous nucleotides	30	30 (2.7%)
Total	1,136	1,108

Table S2. PCR conditions for amplicon preparation

Gene	Primers*	Reaction conditions
msp1[†]	F: GCCTCCCTCGCGCCATCAG-XXXX-AACTAGAAGCTTTAGAAGATGCA R: GCCTTGCCAGCCCGCTCAGACATATGATTGGTTAAATCAAAG	95 °C for 5 min 95 °C 45 s, 50 °C 1 min, 72 °C 1.5 min (35 cycles) 72 °C for 10 min
msp2	F: gcctccctcgcgccatcag-XXXX-gcaacacattcataaacaatgct R: gccttgccagcccgctcagtgcatatgtccatgttgtcc	95 °C for 5 min 95 °C 1 min, 50 °C 45 seconds, 72 °C 1.5 min (35 cycles) 72 °C for 10 min
dhfr	F:GCCTCCCTCGCGCCATCAG-XXXX-CACATTTAGAGGTCTAGGAAATAAAGG R: GCCTTGCCAGCCCGCTCAGTCCCAAGTAAAACTATTAGATCTTCAA	95 °C for 5 min 95 °C 45 s, 45 °C 45 s, 72 °C 1.5 min (35 cycles) 72 °C for 10 min

*XXXX represents a 4 nucleotide bar code unique for each patient.
[†]Primers adapted from Ferreira et al. (4).

Table S3. Sensitivity analysis for cutoff for determining unique variant

	Cutoff				
Gene	5 bp	6 bp	7 bp	8 bp	9 bp
msp2					
Total *msp2* variants	52	52	52	51	51
msp2 variants lost	NA	0	0	1	0
Total *msp2* variants, %*	100	100	100	98	98
msp1					
Total *msp1* variants	63	62	62	61	61
msp1 variants lost	NA	1	0	1	0
Total *msp1* variants, %*	100	98	98	97	97

NA, not applicable.
*Using the total number from the 5-bp cutoff.