Single cell sequencing of the small and AT-skewed genome of malaria parasites

Shiwei Liu[1], Adam C. Huckaby[1], Audrey C. Brown[1], Christopher C. Moore[2], Ian Burbulis[3, 5], Michael J. McConnell[3, 4, 6], Jennifer L. Güler[1, 2]*

[1]Department of Biology, University of Virginia, Charlottesville, VA.

[2]Division of Infectious Diseases and International Health, University of Virginia, Charlottesville, VA

[3]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA

[4]Department of Neuroscience, University of Virginia School of Medicine, Charlottesville, VA

[5]Escuela de Medicina, Universidad San Sebastian, Puerto Montt, Chile

[6]Current address:

Lieber Institute for Brain Development, Baltimore, MD

*Corresponding author contact: jlg5fw@virginia.edu

Abstract

Single cell genomics is a rapidly advancing field; however, most techniques are designed for mammalian cells. Here, we present a single cell sequencing pipeline for the intracellular parasite, *Plasmodium falciparum*, which harbors a relatively small genome with an extremely skewed base content. Through optimization of a quasi-linear genome amplification method, we achieve more even genome coverage and better targeting of the parasite genome over contaminants. These improvements are particularly important for expanding the accessibility of single cell approaches to new organisms and cell types and for improving the study of genetic mechanisms of adaptation.

Keywords: whole-genome amplification, AT-skewed genome, malaria, single cell sequencing, MALBAC

Background

Malaria is a life-threatening disease caused by protozoan *Plasmodium* parasites. *P. falciparum* causes the greatest number of human malaria deaths [1]. The clinical symptoms of malaria occur when parasites invade human erythrocytes and undergo rounds of asexual reproduction by maturing from early forms into late stage parasites and bursting from erythrocytes to begin the cycle again [2]. In this asexual cycle, parasites possess a single haploid genome during the early stages; rapid genome replication in the later stages leads to an average of 16 genome copies [2].

Due to a lack of an effective vaccine, antimalarial drugs are required to treat malaria. However, drug efficacy is threatened by the frequent emergence of resistant populations [3]. Copy number

variations (CNVs), or the amplification and deletion of a genomic element, is one of the major sources of genomic variation in *P. falciparum* that contribute to antimalarial resistance [4–15]. Similar to bacteria and viruses [16–18], a high rate of CNVs may initiate genomic changes that contribute to the rapid adaptation of this organism [7, 19]. Despite the importance of CNVs, their dynamics in evolving populations are not well understood.

The majority of CNVs in *P. falciparum* have been identified by analyzing bulk DNA in which the CNVs are present in a substantial fraction of individual parasites in the population due to positive selection [8, 10, 11, 20, 21]. However, many CNVs likely remain undetected because they are presumably either deleterious or offer no advantages for parasite growth or transmission [22] and are therefore present in low frequency [20, 22]. Currently, most CNVs are identified using read-depth analysis of short read sequencing data, which derives an average signal across the population. For this reason, genetic variants must be present in a high frequency (i.e. ~50%) in the population to be detected [23–25]. Sequencing at very high depth improves the detection of low frequency CNVs, but the sensitivity is limited to large-scale CNVs present in > 5% cells [26–28]. Analysis methods that rely on the detection of reads that span CNV junctions have improved the sensitivity and specificity of CNV detection [29], but continue to struggle with minor allele detection.

Recent investigations have analyzed single cells to detect low frequency CNVs within heterogeneous populations [25, 30–34]. This approach provides a significant advantage for detecting rare genetic variants by no longer deriving an average signal from large quantities of cells. However, short read sequencing requires nanogram to microgram quantities of genomic

material for library construction, which is many orders of magnitude greater than the genomic content of individual *Plasmodium* cells. Therefore, whole genome amplification (WGA) is required to generate sufficient DNA quantities. Several WGA approaches have been reported and each has advantages and disadvantages for different applications [35–38] but most were optimized for mammalian cell analysis [28, 36, 38–49].

To date, reliable detection of CNV in single *P. falciparum* parasites using whole genome sequencing has not been achieved. The application of existing WGA methods is complicated by this parasite's small genome size and extremely imbalanced base composition (23Mb haploid genome with 19.4% GC-content [50]). Each parasite haploid genome contains 25 femtograms of DNA, which is ~ 278 times less than the ~6400Mb diploid human genome. Therefore, an effective *P. falciparum* WGA method must be both highly sensitive and able to handle its imbalanced base composition. One WGA method, multiple displacement amplification (MDA), has been used to amplify single *P. falciparum* genomes with near complete genome coverage [51, 52]. These studies successfully detected single nucleotide polymorphisms between single parasites; however, MDA is less useful for CNV detection because analysis is disrupted by low genome coverage uniformity [37] and the generation of chimeric reads [53].

Multiple annealing and looping-based amplification cycling (MALBAC) is another WGA method that exhibits improved uniformity over MDA, which is advantageous for detecting CNVs in single cells [27]. MALBAC has the unique feature of quasi-linear pre-amplification, which reduces the bias associated with exponential amplification [27]. However, standard MALBAC is less tolerant to AT-biased genomes, unreliable with low DNA input, and prone to

contamination [54–56]. Thus, optimization of this WGA method is necessary for *P. falciparum* genome analysis.

In this study, we developed a single cell sequencing pipeline for *P. falciparum* parasites, which included efficient isolation of single infected erythrocytes, an optimized WGA method inspired by MALBAC, and a sensitive method of assessing sample quality prior to sequencing. We tested our pipeline on erythrocytes infected with laboratory-reared parasites as well as patient-isolated parasites with heavy human genome contamination. Genome amplification using our optimized protocol showed increased genome coverage, better coverage uniformity, and strong amplification reproducibility. These improvements will enable the detection of parasite-to-parasite heterogeneity to clarify the role of genetic variations, such as CNVs, in the adaptation of *P. falciparum*. This study also provides a framework for the optimization of single cell amplification in other organisms with challenging genomes.

Methods

Parasite Culture

We freshly thawed erythrocytic stages of *P. falciparum* (*Dd2*, MRA-150, Malaria Research and Reference Reagent Resource Center, BEI Resources) from frozen stocks and maintained them as previously described [57]. Briefly, parasites were grown in *vitro* at 37°C in solutions of 3% hematocrit (serotype A positive human erythrocytes, Valley Biomedical, Winchester, VA) in RPMI 1640 (Invitrogen, USA) medium containing 24 mM $NaHCO_3$ and 25 mM HEPES, and supplemented with 20% human type A positive heat inactivated plasma (Valley Biomedical, Winchester, VA) in sterile, plug-sealed flasks, flushed with 5% O_2, 5% CO_2, and 90% N_2 [7].

We maintained the cultures with media changes every other day and sub-cultured them as necessary to maintain parasitemia below 5%. All parasitemia measurements were determined by SYBR green based flow cytometry [58]. Cultures were routinely tested using the LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich, USA) to confirm negative infection status.

Clinical Sample Collection

We obtained parasites from an infected patient admitted to the University of Virginia Medical Center with clinical malaria. The patient had a recent history of travel to Sierra Leone, a malaria-endemic country, and *P. falciparum* infection was clinically determined by a positive rapid diagnostic test and peripheral blood smear analysis. We obtained the sample of 1.4% early stage parasites within 24h of phlebotomy, incubated in the conditions described in Parasite Culture for 48 hours and washed the sample 3 times with RPMI 1640 HEPES to decrease levels of white blood cells. In order to fully evaluate our amplification method in the presence of heavy human genome contamination, we did not perform further leukodepletion. We set aside some of the sample for bulk DNA preparation (see *Bulk DNA Extraction*). Using another portion of the sample, we enriched for parasite-infected erythrocytes using SLOPE (Streptolysin-O Percoll) method [59], which increased the parasitemia from 1.4% to 48.5% (**Figure S1**). We then isolated the single *P. falciparum*-infected erythrocytes using the CellRaft AIR^{TM}System (Cell Microsystems, Research Triangle Park, NC) as detailed in *Parasite Staining and Isolation*.

Bulk DNA Extraction

We lysed asynchronous *P. falciparum*-infected erythrocytes with 0.15% saponin (Sigma-Aldrich, USA) for 5min and washed them with 1x PBS (diluted from 10x PBS Liquid Concentrate,

139 Gibco, USA). We then lysed parasites with 0.1% Sarkosyl Solution (Bioworld, bioPLUS, USA)
140 in the presence of 1mg/ml proteinase K (from *Tritirachium album*, Sigma-Aldrich, USA)
141 overnight at 37°C. We extracted nucleic acids with phenol/chloroform/isoamyl alcohol (25:24:1)
142 pH 8.0 (Sigma-Aldrich, USA) using 2ml light Phase lock Gels (5Prime, USA). Lastly, we
143 precipitated the DNA with ethanol using the standard Maniatis method [60].
144

145 **Parasite Staining and Isolation**

146 For late stage parasite samples, we obtained laboratory *Dd2* parasite culture with a starting
147 parasitemia of 1.7% (60% early stage parasites). We separated late stage *P. falciparum*-infected
148 erythrocytes from non-paramagnetic early stages using a LS column containing MACS®
149 microbeads (Miltenyi Biotec, USA, [61]). After elution of bound late stage parasite, the sample
150 exhibited a parasitemia of 80.8% (74.0% late stage parasites, **Figure S1**). For early stage
151 parasites, we obtained laboratory *Dd2* parasites culture with a starting parasitemia of 3% (46%
152 early stage parasites). We harvested the non-paramagnetic early stages parasites which were
153 present in the flow-through of the LS column containing MACS® microbeads. Next, we enriched
154 the infected erythrocytes using the SLOPE method, which preferentially lysed uninfected
155 erythrocytes [59]. The final parasitemia of enriched early stage parasites was 22.8% (97.0% early
156 stage parasites, **Figure S1**). To differentiate *P. falciparum*-infected erythrocytes from remaining
157 uninfected erythrocytes or cell debris, we stained the stage specific *P. falciparum*-infected
158 erythrocytes with both SYBR green and MitoTracker Red CMXRos (Invitrogen, USA). We then
159 isolated single *P. falciparum*-infected erythrocytes using the CellRaft AIR™ System (Cell
160 Microsystems, Research Triangle Park, NC). We coated a 100-micron single reservoir array
161 (CytoSort Array and CellRaft AIR user manual, CELL Microsystems) with Cell-Tak Cell and

162 Tissue Adhesive (Corning, USA) following the manufacture's recommendations. Then, we
163 adhered erythrocytes on to the CytoSort array from a cell suspension of ~20,000 cells in 3.5 mL
164 RPMI 1640 (Invitrogen, USA) with AlbuMAX II Lipid-Rich BSA (Thermo Fisher Scientific,
165 USA) and Hypoxanthine (Sigma-Aldrich, USA). Lastly, we set up the AIR™ System to
166 automatically transfer the manually selected single infected erythrocytes (SYBR+, Mitotracker+)
167 into individual PCR tubes.
168

169 **Steps to Limit Contamination**

170 We suspended individual parasite-infected erythrocytes in freshly prepared lysis buffer, overlaid
171 them with one drop (approx. 25 µl) of mineral oil (light mineral oil, BioReagent grade
172 for molecular biology, Sigma Aldrich, USA), and stored them at -80°C until WGA. We
173 amplified DNA in a clean positive pressure hood located in a dedicated room, using dedicated
174 reagents and pipettes, and stored them in a dedicated box at -20°C. We wore a new disposable
175 lab coat, gloves and a face mask during reagent preparation, cell lysis, and WGA steps. We
176 decontaminated all surfaces of the clean hood, pipettes, and tube racks with DNAZap (PCR
177 DNA Degradation Solutions, Thermo Fisher Scientific, USA), followed by Cavicide (Metrex
178 Research, Orange, CA), and an 80% ethanol rinse prior to each use. We autoclaved all tubes,
179 tube racks and the waste bin on a dry vacuum cycle for 45 min. Finally, we used sealed sterile
180 filter tips, new nuclease-free water (Qiagen, USA) for each experiment, and filtered all salt
181 solutions through a 30mm syringe filter with 0.22 µm pore size (Argos Technologies, USA)
182 before use in each experiment.
183

184 **Whole Genome Amplification**

Standard MALBAC: The MALBAC assay was originally designed for human cells [27, 48]. This approach involved making double stranded DNA copies of genomic material using random priming and then exponentially amplifying these copies via suppression PCR. Here, we transferred individual cells into sterile thin-wall PCR tubes containing 2.5 µl of lysis buffer that yielded a final concentration of 25 mM Tris pH 8.8 (Sigma-Aldrich, USA), 10 mM NaCl (BAKER ANALYZED A.C.S. Reagent, J.T.Baker, USA), 10mM KCl (ACS reagent, Sigma-Aldrich, USA), 1mM EDTA (molecular biology grade, Promega, USA), 0.1% Triton X-100 (Acros Organics, USA), 1mg/ml Proteinase K (*Tritirachium album,* Sigma-Aldrich, USA). After overlaying one drop of mineral oil, we lysed cells at 50°C for 3h and inactivated the proteinase at 75°C for 20min, then 80°C for 5min before maintaining at 4°C. We added 2.5µl of amplification buffer to each sample to yield a final concentration of 25mM Tris pH 8.8 (Sigma-Aldrich, USA), 10mM $(NH_4)_2SO_4$ (Molecular biology grade, Sigma-Aldrich, USA), 8mM $MgSO_4$ (Fisher BioReagents, Fisher Scientific, Product of India), 10mM KCl (ACS reagent, Sigma-Aldrich, USA), 0.1% Triton X-100 (Acros Organics, USA), 2.5mM dNTP's (PCR grade, Thermo Fisher Scientific, USA), 1M betaine (PCR Reagent grade, Sigma-Aldrich, USA) and 0.667µM of each random primer (5'GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNTTT 3', and 5'GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNGGG 3') ordered from Integrated DNA Technologies, USA. To denature DNA, we heated samples to 95°C for 3min and snap-cooled on an ice slush before gently adding 0.5µl of enzyme solution (8,000 U/ml *Bst* DNA Polymerase Large Fragment, New England Biolabs, USA, in 1X amplification buffer) into the aqueous droplet.

207 We placed the samples into a thermo-cycler (Bio-Rad, USA) holding at 4°C and heated
208 according to the following cycles: 10°C – 45s, 15°C – 45s, 20°C – 45s, 30°C – 45s, 40°C – 45s,
209 50°C – 45s, 64°C – 10min, 95°C – 20s. The samples were immediately snap-cooled on an ice
210 slush and held for at least 3min to maintain the DNA in a denatured state for the next round of
211 random priming. We added another 0.5µl of enzyme solution and mixed thoroughly with a
212 pipette on ice as above. We placed the samples back into the 4°C thermo-cycler and heated
213 according to the cycles listed above with an additional 58°C step for 1min before once again
214 cooling on an ice slush for 3min. We repeated the addition of enzyme mix (as above) and
215 performed an additional round of amplification cycles (as above, including the 58°C step) four
216 more times (for a total of 6 cycles). Once completed, we placed the samples on ice and
217 supplemented with cold PCR master mix to yield 50µl with the following concentrations: 0.5µM
218 Common Primer (5'GTGAGTGATGGTTGAGGTAGTGTGGAG3', Integrated DNA
219 Technologies, USA), 1.0mM dNTPs (PCR grade, Thermo Fisher Scientific, USA), 6.0mM
220 $MgCl_2$ (Molecular biology, Sigma-Aldrich, USA), 1X Herculase II Polymerase buffer and
221 1X Herculase II polymerase (Agilent Technologies, USA). We immediately thermo-cycled
222 samples with the following temperature-time profile: 94°C – 40s, 94°C – 20s, 59°C – 20s, 68°C –
223 5min, go to step two 14 times, and an additional extended at 68°C – 5min, and finally, a hold at
224 4 °C. Overall, for single parasite genomes amplified by standard MALBAC samples, we used
225 18/19 linear cycles and 17 exponential cycles.
226
227 **Optimized MALBAC**: We made the following modifications to standard MALBAC to produce
228 our improved method. **1)** We froze cells at -80°C until usage because freeze-thaw enhanced cell
229 lysis as previously reported [52]; **2)** We removed betaine from the amplification buffer because it

230 improved amplification of GC-rich sequences [62], which are infrequent in *P. falciparum*
231 genomes **(Table S1)**; **3)** We used a single random primer where the GC-content of the
232 degenerate bases were 20% instead of 50%
233 (5'GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNTTT 3') at final concentration of
234 1.2µM; **4)** We reduced the volume of the random priming reaction by added only 0.29µl of
235 2X amplification buffer to the lysed samples and 0.13µl of enzyme solution to the aqueous
236 droplet each amplification cycle; **5)** We added additional random priming cycles over prior
237 MALBAC studies for a total of 18 (for late stage parasites) or 19 (for early stage parasites)
238 cycles; **6)** We reduced the total volume of exponential amplification from 50µl to 20µl and
239 increased the number of exponential amplification cycles from 15 to 17; **7)** We verified the
240 presence of high molecular weight DNA products in the samples before purifying nucleic acids
241 by Zymo DNA Clean & Concentrator-5 (ZYMO Research).
242

243 **Pre-Sequencing Quality Assessment**

244 We assayed 6 distinct genomic loci across different chromosomes to determine variations in
245 copy number following the whole genome amplification step. We included this step, which
246 employs highly sensitive droplet digital PCR (ddPCR, QX200 Droplet Digital PCR
247 system, Bio-Rad, USA), to identify samples that exhibited more even genome coverage prior to
248 short read sequencing. The sequence of primers and probes are described in **Table S2** [7, 63, 64].
249 Each ddPCR reaction contained 5µl of DNA (0.3ng/ul for single cell samples), 10µl ddPCR
250 Supermix for Probes (without dUTP), primers and probes with the final concentration in **Table**
251 **S2**, and sterile H_2O to bring the per-reaction volume to 22µl. We prepared droplets with the PCR
252 mixture following the manufacture's protocol: 95°C – 10 min; 40 cycles of 95°C – 30s, 60°C –

60s, and an infinite hold at 4°C. After thermal cycling, we counted positive droplets using the Bio-Rad QX200 Droplet Reader (Bio-Rad, USA). We analyzed data through QuantaSoft (Bio-Rad, USA). For each gene, a no template control (sterile water, NTC) and a positive control (0.025ng *Dd2* genomic DNA) are included in each ddPCR run. Following ddPCR, we calculated the "uniformity score" using the locus representation of the 6 genes (displayed as gene-1, gene-2, gene-3, gene-4, gene-5, gene-6 in the equation) in the amplified DNA sample relative to non-amplified DNA using the following equation:

$$Uniformity\ score = \frac{gene1}{gene2} + \frac{gene1}{gene3} + \frac{gene1}{gene4} + \frac{gene1}{gene5} + \frac{gene1}{gene6} + \frac{gene2}{gene3} + \frac{gene2}{gene1} + \frac{gene2}{gene4} + \frac{gene2}{gene5} + \frac{gene2}{gene6} + \frac{gene3}{gene4} + \frac{gene3}{gene1} + \frac{gene3}{gene2} + \frac{gene3}{gene5} + \frac{gene3}{gene6} + \frac{gene4}{gene5} + \frac{gene4}{gene1} + \frac{gene4}{gene2} + \frac{gene4}{gene3} + \frac{gene4}{gene6} + \frac{gene5}{gene1} + \frac{gene5}{gene2} + \frac{gene5}{gene3} + \frac{gene5}{gene4} + \frac{gene5}{gene6} + \frac{gene6}{gene1} + \frac{gene6}{gene2} + \frac{gene6}{gene3} + \frac{gene6}{gene4} + \frac{gene6}{gene5}$$

When certain loci were over- or under-represented in the amplified sample, this score increased above the perfect representation of the genome; a uniformity score of 30 indicates that all genes are equally represented. We calculated the locus representation from the absolute copies of a gene measured by ddPCR from 1ng of amplified DNA divided by the absolute copies from 1ng of the bulk DNA control [65]. We only included samples in which all six genes were detected by ddPCR. The relative copy number of the *PfMDR1 (P. falciparum* multi-drug resistance gene 1), which was amplified in the *Dd2* parasite line [6], was also used to track the accuracy of amplification. We calculated this value by dividing the ddPCR-derived absolute copies of *PfMDR1* by the average absolute copies of the 6 assayed loci (including *PfMDR1*, normalized to a single copy gene*)*.

Short Read Sequencing

We fragmented MALBAC amplified DNA (> 1ng/µL, 50µL) using Covaris M220 Focused Ultrasonicator in microTUBE-50 AFA Fiber Screw-Cap (Covaris, USA) to a target size of 350bp using a treatment time of 150s. We determined the fragment size range of all sheared DNA samples (291bp-476bp) with a Bioanalyzer on HS DNA chips (Agilent Technologies, USA). We used the NEBNext Ultra DNA Library Prep Kit (New England Biolabs, USA) to generate Illumina sequencing libraries from sheared DNA samples. Following adaptor ligation, we applied 3 cycles of PCR enrichment to ensure representation of sequences with both adapters and the size of the final libraries range from 480bp to 655bp. We quantified the proportion of adaptor-ligated DNA using real-time PCR and combined equimolar quantities of each library for sequencing on 4 lanes of an Illumina Nextseq 550 using 150bp paired end cycles. We prepared the sequencing library of clinical bulk DNA as above but sequenced it on an Illumina Miseq using 150bp paired end sequencing.

Sequencing Analysis

We performed read quality control and sequence alignments essentially as previously described [66] (**Figure S2A**). Briefly, we removed Illumina adapters and PhiX reads, and trimmed WGA common primers from reads in each fastq file with BBDuk tool in BBMap [67]. To identify the source of DNA reads, we randomly subsetted 10,000 reads from each sample busing the reformat tool in BBMap [67] and blasted reads in nucleotide database using BLAST+ remote service. We aligned each fastq file to the hg19 human reference genome and kept the unmapped reads (presumably from *P. falciparum*) for analysis. Then, we aligned each fastq file to the 3D7 *P. falciparum* reference genome with Speedseq [68]. We discarded the reads with low-mapping quality score (below 10) and duplicated reads using Samtools [69]. To compare the coverage

296 breadth between single cell samples at the same sequencing depth, we downsampled BAM files
297 using Samtools [69]. We calculated the coverage statistics using Qualimap 2.0 [70] and subsetted
298 genic, intergenic, and non-telomere regions using Bedtools [71]. We conducted single cell
299 sequencing analysis following the steps in **Figure S2B.** We compared the variation of
300 normalized read abundance (log10 ratio) at different bin sizes using boxplot analysis (R version
301 3.6.1) and determined the bin size of 20 kb using the plateau of decreasing variation of
302 normalized read abundance (log10 ratio) when increasing bin sizes. We then divided the *P.*
303 *falciparum* genome into non-overlapping 20 kb bins using Bedtools [71]. The normalized read
304 abundance was the mapped reads of each bin divided by the total average reads in each
305 sample. To show the distribution of normalized read abundance along the genome, we drew the
306 circular coverage plot using Circos software and ClicO FS [72, 73]. To assess uniformity of
307 amplification, we calculated the coefficient of variation of normalized read abundance by
308 dividing the standard deviation by the mean and multiplying by 100 [37, 74] and analyzed the
309 equality of coefficients of variation using the R package "cvequality" version 0.2.0 [75]. We
310 used Pearson correlation coefficients calculated using the R package "Hmisc" version 4.3-0 [76]
311 to assess amplification reproducibility as it has been applied in a previous study to compare
312 different amplification methods on different sequencing platforms [77]. We drew the scatter plot
313 of normalized read abundance using the "plot" function in R and drew heatmap and hierarchical
314 clustering of Pearson correlation coefficient with the "gplots" R package version 3.0.1.1 [78].
315 Additionally, to estimate the chance of random primer annealing during MALBAC pre-
316 amplification cycles, we generated all possible 5-base sliding windows with 1 base step-size in
317 the *P. falciparum* genome and calculated the GC-content of the 5-bases windows using Bedtools
318 [71].

Results

Plasmodium falciparum genomes from single-infected erythrocytes are amplified by MALBAC

Our single cell sequencing pipeline for *P. falciparum* parasites included stage-specific parasite enrichment, isolation of single infected erythrocytes, cell lysis, whole genome amplification, pre-sequencing quality control, whole genome sequencing, and analysis steps (**Figure 1A**). We collected parasites from either an *in vitro*-propagated laboratory line or from a blood sample of an infected patient (referred to as 'laboratory' and 'clinical' parasites, respectively). This allowed us to test the efficiency of our procedures on parasites from different environments with varying amounts of human host DNA contamination. Furthermore, for laboratory samples, we isolated both early (1n) and late (~16n) stage parasite-infected erythrocytes to evaluate the impact of parasite DNA content on the performance of WGA. For single cell isolation, we used the microscopy-based CellRaft Air system (**Figure 1B**), which has the benefit of low capture volume (minimum: 2µl) and visual confirmation of parasite stages. Following isolation, using the standard MALBAC protocol (termed non-optimized MALBAC), we successfully amplified 3 early (ENM) and 4 late stage (LNM) laboratory samples. We also applied a version of MALBAC that we optimized for the small AT-rich *P. falciparum* genome (termed optimized MALBAC) to 42 early (EOM) and 20 late stage (LOM) laboratory samples as well as 4 clinical samples (COM). Compared to standard MALBAC, our optimized protocol had a lower reaction volume, more amplification cycles, and used a modified pre-amplification random primer (see *Methods* for more details). Using this method, we successfully amplified 43% of the early and

90% of the late stage laboratory samples and 100% of the clinical samples (see post-amplification yields in **Table S3**).

A novel pre-sequencing quality control step identifies samples with more even genome amplification.

We assessed the quality of WGA products from single cells using droplet digital PCR (ddPCR) to measure the copy number of single and multi-copy genes dispersed across the *P. falciparum* genome (6 genes in total including *PfMDR1*, which is present at ~3 copies in the *Dd2* laboratory parasite line). Using this sensitive method, along with calculation of a "uniformity score", we were able to select genomes that had been more evenly amplified (i.e. fewer genes were over- or under-amplified); a low uniformity score and accurate copy number values indicated a genome that has been amplified with less bias (see *Methods* for details on score calculation). When we analyzed differences between all successfully amplified samples by optimized MALBAC (17 EOM samples and 14 LOM samples) and non-optimized MALBAC (3 ENM and 4 LNM samples), we found that samples amplified with optimized MALBAC were more evenly covered than those amplified with standard MALBAC (**Table 1**). For example, the average uniformity score of EOMs and ENMs was 391 and 16,104, respectively and the average *PfMDR1* copy number of EOMs and ENMs was 2.9 and 0.1, respectively.

Based on the results of ddPCR detection, we selected a subset of 13 EOM and 10 LOM laboratory parasite-infected erythrocytes for sequencing (**Table 1**). Overall, selected samples had lower average uniformity scores (i.e. 248 and 1012 for selected and unselected EOMs, respectively, see **Table 1**). Three out of 4 COM samples showed a lack of ddPCR detection on at

least one parasite gene; thus, we were not able to calculate a uniformity score for these samples and the amplification of clinical genomes was likely more skewed than laboratory samples (**Table 1**).

Table 1. Pre-sequencing quality control by droplet digital PCR

Single cell	MALBAC type	Sample type (n)	Pre-sequencing ddPCR assessment				Selected for Seq.
			Uniformity score Average (SD)		Mean *PfMDR1* CN Average (SD)		
Yes	Optimized	EOM* (13)	391 (364)	248 (202)	2.9 (1.9)	2.6 (0.8)	Yes
		EOM (4)		1012 (195)		3.7 (3.9)	No
		LOM* (10)	227 (334)	118 (69)	2.4 (1.5)	2.2 (1.3)	Yes
		LOM (4)		775 (683)		2.8 (2.1)	No
		COM*(2)	369 (-)	369 (-)	3.3 (4.1)	1.9 (0.8)	Yes
		COM(2)		^ (-)		4.7 (6.6)	No
	Non-optimized	ENM* (1)	16104 (3416)	18519 (-)	0.1 (0.1)	0.2 (-)	Yes
		ENM (2)		13689 (-)		0 (-)	No
		LNM* (1)	7350 (8162)	13121 (-)	0.1 (0.1)	0.1 (-)	Yes
		LNM (3)		1578 (-)		0.1 (0.1)	No
No	NA	*Dd2* Bulk (1)	30	-	3	-	Yes
		Clinical Bulk (1)	30	-	1	-	Yes

*EOM: Early stage single parasites amplified by optimized MALBAC; LOM: Late stage single parasites amplified by optimized MALBAC; COM: Clinical single parasites amplified by optimized MALBAC; ENM: Early stage single parasites amplified by non-optimized MALBAC; LNM: Late stage single parasites amplified by non-optimized MALBAC. Uniformity scores were calculated when all of the six genes were detected by ddPCR in the sample. ^Due to the lack of ddPCR detection of some genes in COM samples, the uniformity score could not be calculated. (-) Indicates only one sample was included in the calculation. SD, standard deviation.

Optimized MALBAC limits contamination of single cell samples.

Both standard and optimized MALBAC-amplified parasite genomes were short-read sequenced alongside a matched bulk DNA control (**Table 1**). After quality control, we aligned the reads to the *P. falciparum* 3D7 reference genome (see *Methods* and **Figure S2** for details). We first assessed the proportion of contaminating reads in our samples; NCBI Blast results showed that

381 the majority of non-*P. falciparum* reads were of human origin (**Figure 2A**). The proportions of
382 human reads in 6 out of 13 EOM samples (1.1%-6.9%) and 8 out of 10 LOM samples (1.4%-
383 6.1%) were lower than that in the control bulk sample (7.4%, **Figure 2A**). Conversely, the
384 proportion of human reads in ENM and LNM samples were much higher (81.8% and 18.9%,
385 respectively). As shown in other studies [79, 80], our clinical bulk DNA (81.9%) contained a
386 much higher level of human contamination than the laboratory bulk DNA (7.4%). However, we
387 found that the proportion of the human contaminating DNA in the two single cell COM samples
388 was considerably lower (58.8% and 65.5%). The second most common source of contaminating
389 reads in the sequenced samples was from bacteria such as *Staphylococcus* and *Cutibacterium*.
390 The ENM sample exhibited a ~10-fold increase in the proportion of bacterial reads over
391 averaged EOM samples (8.2% versus 0.8%, respectively) whereas the LNM samples showed the
392 same proportion of bacterial reads as the averaged LOM samples (0.2%). These results indicated
393 that the optimized MALBAC protocol reduced the amplification bias towards contaminating
394 human and bacterial genomes.
395
396 **Optimized MALBAC reduces amplification bias of single cell samples.**
397 To further assess the optimized MALBAC protocol, we evaluated GC-bias at several steps of our
398 pipeline (i.e. WGA, library preparation, and the sequencing platform itself). Analysis of the bulk
399 genome control (without WGA) indicated that there was little GC-bias introduced by the library
400 preparation, sequencing, or genome alignment steps; the GC-content of mapped reads from bulk
401 sequencing data is 18.9% (**Table 2**), which was in line with the GC-content (19.4%) of the
402 reference genome [50]. We then compared values from single cell samples to those from the
403 appropriate bulk control to evaluate the GC-bias caused by MALBAC amplification (**Figure**

404 **2B**). The average GC-content of all EOM (21.4%), LOM (22.4%), and COM (20.7%) samples
405 was within 1-3.5% of the bulk controls from laboratory and clinical samples (18.9% and 19.7%,
406 respectively, **Table 2**). However, the average GC-content of ENM and LNM samples was 6.1%
407 and 5.4% greater than that of the bulk control; this results is consistent with the high GC
408 preference of the standard protocol [36, 56]. ENM and LNM samples also showed a greater
409 proportion of mapped reads with high GC-content (> 30%) than EOM, LOM, and bulk DNA
410 samples (**Figure 2B**).

411

412 **Table 2. Average GC-content and coverage breadth of sequenced samples**

MALBAC type	Single cell	Sample type (#)	Average GC content of all mapped reads	Average of mean coverage (X)	Coverage breadth		
					Whole genome	Genic regions	Intergenic regions
Optimized	Yes	EOM (13)	21.4%	37.54	57.9%	78.0%	27.8%
				1.00*	24.1%	37.4%	4.2%
		LOM (10)	22.4%	43.10	57.3%	79.0%	25.0%
				20.00*	52.3%	74.6%	18.9%
		COM (2)	20.7%	9.54	48.0%	67.7%	18.5%
Non-optimized		ENM (1)	25.0%	1.47	23.0%	34.4%	6.1%
				1.00*	19.4%	29.3%	4.7%
		LNM (1)	24.3%	20.43	47.4%	67.9%	16.9%
				20.00*	47.3%	67.7%	16.7%
NA	No	*Dd2*_Bulk (1)	18.9%	75.83	96.1%	97.0%	94.9%
		Clinical Bulk(1)	19.7%	0.03	0.3%	0.3%	0.2%

413 * Indicate samples have been randomly downsampled to the same mean coverage for the pair-wise comparison
414 between EOM and ENM samples, and between LOM and LNM samples.

415

416 Since GC-bias during the amplification step can limit which areas of the genome are sequenced,
417 we assessed whether the optimization of MALBAC improved genome coverage. The coverage
418 breadth of single cell samples increased by 34.9% in early stage samples (**Figure 2C**, orange-
419 ENM to grey-EOM lines) and by 9.9% for late stage samples following optimization (**Figure**
420 **2C**, red-LNM to purple-LOM lines, see values in **Table 2**). Even when we randomly

421 downsampled EOM and ENM samples to the same coverage (1X), EOM samples continued to
422 show higher coverage breadth than the ENM sample (**Table 2**). Similarly, LOM samples showed
423 improvement in coverage breadth over the LNM sample (**Table 2**). Even though optimized
424 MALBAC showed less bias towards GC-rich regions, it was still problematic for highly AT-rich
425 and repetitive intergenic regions (mean of 13.6% GC-content, [50]). The fraction of intergenic
426 regions covered was only 27.8% for EOM samples and 25.0% for LOM samples on average.
427 When we excluded intergenic regions, the fraction of genic regions of the genome covered by at
428 least one read reached an average of 78.0% and 79.0% for EOM and LOM samples (**Table 2**).
429 Conversely, the coverage of intergenic and genic regions was significantly lower for the non-
430 optimized samples (**Table 2**). Coverage of the *P. falciparum* genome in the clinical bulk sample
431 was very low due to heavy contamination with human reads (0.3% of the genome was covered
432 by at least one read). This was much lower than that from the 2 COM samples (an average of
433 48%, **Figure 2C** and **Table 2**).
434
435 **Optimized MALBAC improves uniformity of single cell genomes.**
436 To investigate the uniformity of read abundance distributed over the *P. falciparum* genome, we
437 divided the reference genome into 20kb bins and plotted the read abundance in these bins over
438 the 14 chromosomes (**Figure 3A, Figure S3**, **Figure S4A**). We selected a 20kb bin size based on
439 its relatively low coverage variation compared to smaller bin sizes and similar coverage variation
440 as the larger bin sizes (**Figure S5**). To quantitatively measure this variation, we calculated the
441 normalized read abundance per bin in each sample (by dividing the raw read counts with the
442 mean read counts per 20kb bin, **Figure 3B, Figure S3C**). Here, the bulk control displayed the
443 smallest range of read abundance for outlier bins (blue circles) and lowest interquartile range

(IQR) value of non-outlier bins (black box, **Figure 3B, Figure S3C**), indicating less bin-to-bin variation in read abundance. Both EOM and LOM samples exhibited a smaller range of normalized read abundance in outlier bins than ENM and LNM samples (**Figure 3B, Figure S3C**). In addition, the read abundance variation of COM samples was similar to EOM or LOM samples (**Figure 3B, Figure S4B**). Finally, due to the extremely low coverage of the clinical bulk sample, the read abundance variation was much higher than all other samples (**Figure 3B, Figure S4B**).

We then calculated the mean coefficient of variation (CV) for read abundance in the different sample types to compare the variation of read abundance in all sequenced samples (**Table 3, Table S4, Figure 3C**). The CV from the ENM sample was significantly higher compared to the CV of each EOM sample (147% versus a mean of 89%, respectively, pairwise p value < 0.01, **Table S5**). Similarly, the LNM-CV was significantly higher compared to the CV of each LOM sample (111% versus a mean of 79%, respectively, pairwise p value <0.01, **Table S5**). These data showed improvement in levels of read uniformity across the genome when using optimized MALBAC over the standard protocol. In support of this finding, the CV value of COM samples was similar to EOM and LOM samples (**Table 3, Figure 3C**).

Table 3. Coefficient variation of normalized read abundance in each sample type

Sample name	Mean Coefficient of Variation (CV)	SD
Dd2 Bulk (1)	22	-
ENM (1)	147	-
EOM (13)	89	4
LNM (1)	111	-
LOM (10)	79	2
COM (2)	87	12
Clinical Bulk (1)	472	-

SD, standard deviation.

Optimized MALBAC exhibits reproducible coverage of single cell genomes.

To better assess the amplification patterns across the genomes, we compared the distribution of binned normalized reads from single cell samples to the bulk control using the Pearson correlation test as used in other single cell studies [36, 81]. This analysis revealed that amplification patterns of optimized EOM and LOM samples were positively correlated with the bulk control (Pearson correlation coefficient of 0.32 and 0.30, respectively, **Table S6**), while the non-optimized samples were not (ENM: -0.13 and LNM: -0.05) (**Figure 4A**). Therefore, the parasite genome was better represented by single cell samples amplified by optimized MALBAC than those amplified by the standard MALBAC protocol. To quantify the reproducibility of amplification distributions, we compared the Pearson correlation coefficient between different single cell samples amplified by MALBAC. The read abundance across all single cell samples was highly correlated; two individual EOM or LOM samples had a correlation coefficient of 0.93 (**Figure 4B**). When we expanded our analysis to calculate the correlation of binned normalized reads between all 26 sequenced samples (**Table S6**) and hierarchically clustered the Pearson correlation coefficient matrix between these samples, all 23 optimized single cell samples (EOM and LOM) clustered with a mean Pearson correlation coefficient of 0.87 (**Figure 4C**). In addition, the two COM samples were correlated with each other (Pearson correlation coefficient of 0.81) (**Figure S4**C) as well as with EOM samples (0.75) and LOM samples (0.73) (**Table S6**). This correlation indicated high reproducibility of normalized read distribution across the genomes that were amplified by optimized MALBAC. Within the large cluster, two LOM samples (LOM12 and LOM13) displayed the highest correlation (0.95, **Figure 4C**).

MALBAC amplification maintains improvements over MDA-based amplification of single cell samples.

To test whether optimized MALBAC provided better uniformity and reproducibility than the MDA-based amplification of parasite genomes from single infected erythrocytes, we compared our results with those generated by a previous study ([52], **Figure S6**). This study sorted individual infected erythrocytes with high (H), medium (M) and low (L) DNA content corresponding to late, mid, and early stage parasites applied MDA WGA to single erythrocytes and sequenced the DNA products. In order to enhance the comparison of the two methods, we analyzed data from the twelve MDA samples using our analysis pipeline and parameters (six MDA-H and three of each MDA-M and -L samples). While MALBAC-amplified genomes exhibited a consistent amplification pattern (**Figure S3A, Figure S3B**), the MDA-amplified genomes showed substantially more variation in amplification pattern across cells (**Figure S6A**). We also detected higher variation in normalized read abundance in the MDA-H samples (compared to MDA-M and -L samples, **Figure S6B**), which was consistent with the report that the MDA method amplifies high DNA content better than parasites with lower DNA content [52]. Even though the bulk DNA controls used in both studies showed similar CVs (24% versus 22%), the MDA amplified samples displayed a higher CV than MALBAC amplified samples regardless of the parasite stage (a mean of 186% versus 85%, respectively, **Table 3, Table S7, Figure S6C**). Finally, the correlation between MDA amplified cells (mean correlation coefficient: 0.36; **Table S8, Figure S6D**) was much lower than that between our optimized MALBAC amplified cells (mean correlation coefficient: 0.87; **Table S6**, **Figure 4C**).

Discussion

510 This study is the first to optimize the standard MALBAC procedure for single cell sequencing of
511 a genome with extreme GC-content (*P. falciparum*: 19.4%). We showed that this optimized
512 method can reliably amplify early stage parasite genomes, which contain < 30fg of DNA per
513 sample. Single cell experiments are innately very sensitive to contaminating DNA from other
514 organisms and single cell samples amplified with optimized MALBAC show a lower proportion
515 of contaminating DNA, which improves parasite genome coverage (**Figure 2A, Figure 2C**).
516 Furthermore, we showed that this method reduced the GC-bias in amplification to improve the
517 breadth and uniformity of *P. falciparum* genome amplification (**Figure 2B**, **Figure 3**). Finally,
518 we observed high reproducibility in optimized MALBAC amplified single cell samples (**Figure**
519 **4**); this characteristic may impact CNV analysis to expand capabilities of single-cell approaches
520 for *Plasmodium*, as well as other organisms.
521

522 **MALBAC Volume and Cycles**

523 MALBAC amplification has been used in studies of human cells, where each single genome
524 harbors a picogram level of DNA [27, 48]. In this study, we successfully improved the sensitivity
525 of the MALBAC method to amplify a femtogram level of DNA from single *P. falciparum*
526 parasites. Reducing the total reaction volume (from 50µl to 20µl) and increasing the number of
527 amplification cycles (pre-amplification: from 5 to 19-20; exponential: from 15 to 17) was likely
528 responsible for this improvement in sensitivity. It was essential to combine these two changes;
529 the lower sample volume and decreased starting material reduced the overall DNA yield and
530 therefore, we increased the number of amplification cycles to generate enough material for
531 sequencing. Additional benefits of these modifications included: less contaminating DNA
532 introduced by reagents and reduced costs due to the lower reagent requirement. Importantly,

533 these simple steps can be applied to the MALBAC amplification of small genomes or genomes
534 with skewed GC-content from other organisms such as bacteria [82]. For example, studies of
535 *Mycoplasma capricolum* (GC-poor) *[83]*, *Rickettsia prowasekii* (GC-poor) *[84]*, and *Borrelia*
536 *burgdorferi* (GC-poor) [85], *Entamoeba histolytica* (GC-poor) *[86]*, *Micrococcus luteus* (GC-
537 rich) [87] could be improved using this method.
538

539 **Primers and Coverage Bias**

540 The modification of the primer was essential for the successful amplification of the AT-rich *P.*
541 *falciparum* genome. This change was meant to prevent the preferential amplification of GC-rich
542 regions of the genome observed for human and rat single cell genomes [36, 56]. We increased
543 coverage breadth of *P. falciparum* genic regions (a mean of 21.7% GC-content) from as low as
544 <40% to ~80% (ENM versus EOM and LOM samples, **Table 2)** by specifically altering the base
545 content of the degenerate 5-mer of the MALBAC pre-amplification primer from 50% to 20%
546 GC-content. The initial priming step is crucial for whole genome amplification and controlling
547 this step can limit amplification bias [88]. Indeed, 5-mers with ~20% GC-content across the *P.*
548 *falciparum* genome are > 700,000 times more common than those with ~50% GC-content (**Table**
549 **S1**). This difference indicated that annealing of the optimized MALBAC primer based on the
550 degenerate bases was more specific for the parasite's genome than the standard MALBAC
551 primer. Interestingly, during this study we observed a preferential amplification of genic over
552 intergenic regions (**Table 2**), which may be explained by lower percentage of 5-mers with 20%
553 GC-content in intergenic regions than in genic regions (**Table S1**). Furthermore, when we
554 searched for reads that contained the MALBAC common sequence to identify WGA binding
555 sites across the *P. falciparum* genome, we found that binding sites were predominantly located in

556 the genic regions (**Table S9**); this result indicated that there was an issue with primer annealing
557 in intergenic regions, which may be caused by a high predicted rate of DNA secondary structure
558 formation in intergenic regions of the *P. falciparum* genome [66]. The polymerase used in the
559 MALBAC linear amplification steps (*Bst* large fragment) exhibits strand displacement activity,
560 which allowed resolution of secondary structure [89, 90]. However, a longer extension time may
561 be required for amplification of repetitive DNA sequence, either during linear or exponential
562 steps.
563
564 **Parasite and Contaminating Genomes**
565 The improved sensitivity of optimized MALBAC not only enhanced the amplification of the
566 small parasite genome, but also improved the sensitivity to amplify contaminating non-parasite
567 DNA. Nevertheless, when comparing the two MALBAC protocols, the optimized method
568 yielded a greatly reduced proportion of contaminating DNA (ENM and LNM: 13.6% vs EOM
569 and LOM: 6.9% of total reads, **Figure 2A**). We speculate that this decrease was once again due
570 to our modification of the GC-content of the degenerate bases of the primer (see below); this
571 limited the preferential amplification by standard MALBAC of contamination DNA with higher
572 GC content, improving the representation of parasite DNA in the overall WGA product.
573
574 The major contaminating DNA source that we detected in our samples was from humans (**Figure**
575 **2A**). This was not surprising given that, in our experimental system, the culture and host
576 environments are rich in human DNA [79, 80, 91]. It is also possible that human DNA was
577 introduced during the single cell isolation or WGA steps [55]. The former situation is a larger
578 issue for clinical parasite isolates due to the abundance of white blood cells that contribute to

extracellular DNA when they decay outside of the host [92]. Indeed, we observed more human DNA in clinical bulk and single cell samples (an increase of ~11-fold over laboratory-derived bulk and single cell samples, respectively). The massive level of contamination in the clinical bulk sample and limited coverage of the parasite genome (0.3%) was exacerbated by 1) omission of a leukodepletion step that is routinely employed to limit host cell contamination [93–95] and 2) lower overall sequencing output of that particular run (**Table S10**).

The second most common source of contaminating DNA was bacteria (**Figure 2A**). Since this contaminant was absent in the bulk DNA control and increased in early stage parasite samples (representing an average of 0.8% of EOM reads compared to 0.2% for LOM samples), we predict that bacterial material was introduced during single cell isolation and WGA steps. Although we took precautions to limit this occurrence (see *Methods*), environmental cells and DNA could have been introduced during parasite isolation using the open microscopy chamber of the CellRaft AIR System. In addition, contaminating DNA could be have introduced from the molecular biology grade water [96–98] or WGA reagents [99–102]. Reducing the reaction volume could further reduce this source of contamination.

Early and Late Stage Parasites

Depending on when a novel CNV arises (i.e. early or late in replication), each parasite stage holds advantages for its detection. If the CNV arises in the first round of replication and is copied into most of the genomes of a late stage parasite, having multiple genomes will be advantageous for detection. If the CNV arises later in replication, it will be present in only few of the genomes; therefore, averaging across the genomes, as with bulk analysis, will limit its detection. Since only

602 one haploid genome is present in an early stage parasite, the sensitivity for detecting rare/unique
603 CNVs within parasite populations will be enhanced in this situation.
604 For this reason, staging of parasites in this study was important. We performed stage-specific
605 enrichment before single cell isolation and confirmed that the majority of parasites were the
606 desired stage using flow cytometry (see *Methods*, **Figure S1**, 97% for early stage enrichment and
607 74% for late stage enrichment). Furthermore, during selection of cells by microscopy (before
608 automated collection by the IsoRaft instrument), we confirmed the expected fluorescence
609 intensities for each stage; early stage parasites had a significantly smaller genome and
610 mitochondrion size compared to late state (as in **Figure 1B**). However, differences in preparation
611 of samples may have impacted our parasite stage comparisons. While all late stage samples were
612 isolated, lysed and amplified in the same batch under the same conditions, early stage samples
613 were lysed and amplified in three separate batches (**Table S4**). Despite conserved methods and
614 good concordance in CV between all samples (**Table S4**), minor differences could have
615 contributed to variations in the amplification steps.
616
617 Differences in our genome analysis results from optimized MALBAC samples provided further
618 confidence that the parasites were of the expected stage. Firstly, late stage parasites showed a
619 higher WGA success rate than early stage parasites (90% versus 43%, **Table S3**). This result was
620 explained simply by the presence of extra genomes in the late stage samples (~16n versus 1n)
621 and was consistent with a previous study that used MDA-based amplification methods [52]. Late
622 stage parasites also displayed better uniformity of read abundance (**Table 3**), indicating less
623 amplification bias because fewer regions were missed when more genomes were present.
624 Additionally, there were fewer contaminating reads found in late stage parasites than early stage

625 parasites overall (5.1% versus 8.6%). Once again, this was likely due to a higher ratio of parasite
626 DNA to contaminating DNA in the late stage samples.
627
628 Despite these differences, we observed similar coverage breadth and Pearson correlation
629 coefficients of read abundance for both early and late stage MALBAC-amplified parasites
630 (**Table 2 and Table S6**). This was contrary to the MDA study in single *P. falciparum* parasites
631 that found a higher breadth of genome coverage from the late stage parasites [52]. Our findings
632 confirmed that the pattern of amplification across the genome was determined by the binding of
633 the optimized MALBAC primers.
634
635 **Reproducibility and Future CNVs**
636 Although MDA-based amplification achieved better coverage breadth than optimized MALBAC
637 (~90% versus ~60% of the genome is covered by > 1 read, respectively, [52] and **Table 2**), we
638 observed higher uniformity of coverage with our method. This more even representation of the
639 genome was evident from a lower variance and a high correlation of read abundance between
640 optimized MALBAC samples (MALBAC: **Figures 3 and 4**, MDA: **Figure S6**). The high level
641 of amplification reproducibility (i.e. the same regions are over- and under-amplified across
642 multiple genomes) was especially advantageous for CNV detection because the amplification
643 bias could be normalized across cells, as has been successfully performed for human cells [27,
644 103]. We aim to take advantage of this feature to detect CNVs at a single parasite level.
645 However, before this is possible, we will need to analyze a larger cohort of single cell genomes
646 and gain access to improved CNV analysis tools that are relevant for our system. Fortunately,
647 CNV analysis tools are improving and some have been designed specifically for single cell

648 sequencing data (Scope, Ginkgo, SCCNV) [30, 104, 105]. However, so far, these tools lack the
649 flexibility to be applied to diverse genome types without optimization (i.e. reference genome and
650 ploidy adjustments). By combining improved amplification methods and analysis tools with the
651 higher coverage achievable with microbial genomes, we aim to achieve a resolution in single
652 cells that is well below the current limits of CNV detection of single cell sequencing with
653 common WGA methods (hundreds of kb) [27, 28, 44, 49, 56, 106–108].
654

655 **Limitations and Scope**

656 One limitation in our comparison between standard and optimized MALBAC-amplified samples
657 was the sequencing of only a single standard MALBAC sample from each parasite stage.
658 However, we evaluated an additional 5 samples amplified with the standard protocol and
659 detected a consistently poor ddPCR uniformity score (mean of 16104 across 2 ENM samples and
660 7350 across 2 LNM samples; **Table 1**) and could not calculate the uniformity score for the
661 remaining ENM and LNM samples due to the lack of detection for one or two genes. This result
662 indicated biased coverage and high levels of contaminating DNA coverage in these samples,
663 which made sequencing of these samples futile.
664

665 Since our goal in this study was to evaluate amplification bias, we did not perform SNP analysis
666 on samples to address accuracy of the MALBAC method. Other studies showed that the WGA-
667 induced single nucleotide error rate with MALBAC was higher than that for MDA [27, 55, 109].
668 This was likely due to the use of error-prone large fragment *Bst* polymerase in MALBAC pre-
669 amplification cycles compared to the use of phi29 DNA polymerase with proofreading activity in
670 MDA.

Conclusions

With the modification of reaction volume, cycle number, and GC-content of degenerate primers in MALBAC, we successfully amplified low GC-content DNA from single *P. falciparum* parasites. Our modifications will expand the use of MALBAC amplification approaches to organisms not previously accessible because of small genome size or skewed base content. Furthermore, as we demonstrated with contaminating human and bacterial DNA, primer modification and reduced volumes decreased amplification of undesired genomes in a sample. The reproducible nature of our WGA method, combined with new genome analysis tools, will reduce the effect of amplification bias when conducting large scale single cell analysis and enhance our ability to explore genetic heterogeneity. Thus, we expect this approach to broadly improve study of mechanisms of genetic adaptation in a variety of organisms.

List of abbreviations

MALBAC: Multiple annealing and looping-based amplification cycling

CNVs: Copy number variations

WGA: Whole genome amplification

MDA: Multiple displacement amplification

PfMDR1: Plasmodium falciparum multidrug resistance 1

ddPCR: Droplet digital PCR

NA: Not available

SD: Standard deviation

EOM: Early stage single parasites amplified by optimized MALBAC

LOM: Late stage single parasites amplified by optimized MALBAC

COM: Clinical single parasites amplified by optimized MALBAC

ENM: Early stage single parasites amplified by non-optimized MALBAC

LNM: Late stage single parasites amplified by non-optimized MALBAC

IQR: Interquartile range

CV: Coefficient of variation

SLOPE: Streptolysin-O Percoll

Declarations

Ethical Approval and Wavier for Informed Consent

The University of Virginia Institutional Review Board for Health Sciences Research provided ethical approval for clinical samples used in this study (IRB-HSR protocol #21081). We handled all samples in accordance with approved protocols and in agreement with ethical standards of the Declaration of Helsinki. The University of Virginia Institutional Review Board for Health Sciences Research provided a wavier for informed consent because our study design met the following criteria: the research involved minimal risk to subjects, the waiver does not adversely affect the rights and welfare of subjects, and the research could not practicably be carried out without the waiver.

Availability of data and materials

The full data sets have been submitted to NCBI with bioproject accession number: PRJNA607987. The datasets for the uniformity and reproducibility analysis of MDA-based

amplification on parasite DNA from single infected erythrocytes are available in the NCBI short read archive under the accession PRJNA385321[52].

Competing interests

The authors declare that they have no competing interests.

Funding

Research reported in this publication was supported by the University of Virginia 3Cavaliers Grant (to JLG and CCM), University of Virginia Global Infectious Disease Institute iGrant (to SL), the McDonnell Fellowship (IEB), FONDECYT Regular #1191737 (IEB), and the NIAID of the National Institutes of Health under award number NIH 7R21AI111072 (IEB). These funding agencies were not involved in the experimental design, analysis and interpretation of the data, or writing of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' contributions

MJM and JLG conceived of the project. SL, JLG, and ACH designed the experiments. IB and MJM provided access to essential protocols and equipment (CellRaft AIR System) at the start of the project. ACB and CCM procured and processed clinical samples from the University of Virginia Medical Center. SL conducted all of the experiments. SL analyzed data, with support from ACH and JLG. SL and JLG wrote the manuscript. ACH, ACB, CCM, IB, and MJM edited the manuscript. All authors critically reviewed and approved the manuscript.

739 **Acknowledgements**

740 Our thanks to Michelle Warthan for culturing support, AnhThu Nguyen at Genomics Core
741 (Biology, University of Virginia) for sequencing support, and all members of Dr. Jennifer
742 Guler's laboratory, as well as the laboratory of Dr. William Petri Jr at University of Virginia, for
743 their helpful discussion and insight.

744

References

1. Rich SM, Leendertz FH, Xu G, LeBreton M, Djoko CF, Aminake MN, et al. The origin of malignant malaria. Proc Natl Acad Sci U S A. 2009;106:14902–7.

2. Matthews H, Duffy CW, Merrick CJ. Checks and balances? DNA replication and the cell cycle in *Plasmodium*. Parasites & Vectors. 2018;11:216.

3. Blasco B, Leroy D, Fidock DA. Antimalarial drug resistance: linking *Plasmodium falciparum* parasite biology to the clinic. Nature medicine. 2017;23:917–28.

4. Bopp SER, Manary MJ, Bright AT, Johnston GL, Dharia NV, Luna FL, et al. Mitotic Evolution of *Plasmodium falciparum* Shows a Stable Core Genome but Recombination in Antigen Families. PLOS Genetics. 2013;9:e1003293.

5. Cheeseman IH, Gomez-Escobar N, Carret CK, Ivens A, Stewart LB, Tetteh KKA, et al. Gene copy number variation throughout the *Plasmodium falciparum* genome. BMC Genomics. 2009;10:353–353.

6. Cowman AF, Galatis D, Thompson JK. Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the *pfmdr1* gene and cross-resistance to halofantrine and quinine. Proc Natl Acad Sci U S A. 1994;91:1143–7.

7. Guler JL, Freeman DL, Ahyong V, Patrapuvich R, White J, Gujjar R, et al. Asexual Populations of the Human Malaria Parasite, *Plasmodium falciparum*, Use a Two-Step Genomic Strategy to Acquire Accurate, Beneficial DNA Amplifications. PLOS Pathogens. 2013;9:e1003375.

8. Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, Zhou Y, et al. A systematic map of genetic variation in *Plasmodium falciparum*. PLoS pathogens. 2006;2:e57–e57.

9. Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, et al. Adaptive Copy Number Evolution in Malaria Parasites. PLOS Genetics. 2008;4:e1000243.

10. Price RN, Uhlemann A-C, Brockman A, McGready R, Ashley E, Phaipun L, et al. Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number. Lancet. 2004;364:438–47.

11. Ribacke U, Mok BW, Wirta V, Normark J, Lundeberg J, Kironde F, et al. Genome wide gene amplifications and deletions in *Plasmodium falciparum*. Molecular and Biochemical Parasitology. 2007;155:33–44.

12. Rottmann M, McNamara C, Yeung BKS, Lee MCS, Zou B, Russell B, et al. Spiroindolones, a potent compound class for the treatment of malaria. Science. 2010;329:1175–80.

13. Sidhu ABS, Uhlemann A-C, Fidock DA, Valderramos J-C, Krishna S, Valderramos SG. Decreasing *pfmdr1* Copy Number in *Plasmodium falciparum* Malaria Heightens Susceptibility to Mefloquine, Lumefantrine, Halofantrine, Quinine, and Artemisinin. The Journal of Infectious Diseases. 2006;194:528–35.

14. Singh A, Rosenthal PJ. Selection of Cysteine Protease Inhibitor-resistant Malaria Parasites Is Accompanied by Amplification of Falcipain Genes and Alteration in Inhibitor Transport. Journal of Biological Chemistry. 2004;279:35236–41.

15. Triglia T, Foote SJ, Kemp DJ, Cowman AF. Amplification of the multidrug resistance gene *pfmdr1* in *Plasmodium falciparum* has arisen as multiple independent events. Mol Cell Biol. 1991;11:5244–50.

16. Hendrickson H, Slechta ES, Bergthorsson U, Andersson DI, Roth JR. Amplification–mutagenesis: Evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. Proc Natl Acad Sci USA. 2002;99:2164.

17. Roth JR, Andersson DI. Amplification–mutagenesis—how growth under selection contributes to the origin of genetic diversity and explains the phenomenon of adaptive mutation. Research in Microbiology. 2004;155:342–51.

18. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, et al. Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses. Cell. 2012;150:831–41.

19. Anderson TJC, Patel J, Ferdig MT. Gene copy number and malaria biology. Trends in Parasitology. 2009;25:336–43.

20. Ravenhall M, Benavente ED, Sutherland CJ, Baker DA, Campino S, Clark TG. An analysis of large structural variation in global *Plasmodium falciparum* isolates identifies a novel duplication of the chloroquine resistance associated gene. Sci Rep. 2019;9:8287–8287.

21. Heinberg A, Siu E, Stern C, Lawrence EA, Ferdig MT, Deitsch KW, et al. Direct evidence for the adaptive role of copy number variation on antifolate susceptibility in *Plasmodium falciparum*. Molecular microbiology. 2013;88:702–12.

22. Cheeseman IH, Miller B, Tan JC, Tan A, Nair S, Nkhoma SC, et al. Population Structure Shapes Copy Number Variation in Malaria Parasites. Molecular biology and evolution. 2016;33:603–20.

23. Kalisky T, Quake SR. Single-cell genomics. Nature Methods. 2011;8:311–4.

24. Kalisky T, Blainey P, Quake SR. Genomic analysis at the single-cell level. Annu Rev Genet. 2011;45:431–45.

25. Lauer S, Avecilla G, Spealman P, Sethia G, Brandt N, Levy SF, et al. Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. PLOS Biology. 2018;16:e3000069.

26. Zhang L, Vijg J. Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. Annu Rev Genet. 2018;52:397–419.

27. Zong C, Lu S, Chapman AR, Xie XS. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. Science. 2012;338:1622.

28. Chronister WD, Burbulis IE, Wierman MB, Wolpert MJ, Haakenson MF, Smith ACB, et al. Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell Reports. 2019;26:825-835.e7.

29. Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, et al. Identification of genomic indels and structural variations using split reads. BMC Genomics. 2011;12:375–375.

30. Wang R, Lin D-Y, Jiang Y. SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing. bioRxiv. 2019;:594267.

31. Wang X, Chen H, Zhang NR. DNA copy number profiling using single-cell sequencing. Brief Bioinform. 2018;19:731–6.

32. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nature Reviews Genetics. 2016;17:175.

33. Macaulay IC, Voet T. Single Cell Genomics: Advances and Future Perspectives. PLOS Genetics. 2014;10:e1004126.

34. Wang Y, Navin NE. Advances and Applications of Single-Cell Sequencing Technologies. Molecular Cell. 2015;58:598–609.

35. Estévez-Gómez N, Prieto T, Guillaumet-Adkins A, Heyn H, Prado-López S, Posada D. Comparison of single-cell whole-genome amplification strategies. bioRxiv. 2018;:443754.

36. Hou Y, Wu K, Shi X, Li F, Song L, Wu H, et al. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. GigaScience. 2015;4. doi:10.1186/s13742-015-0068-3.

37. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. Annu Rev Genom Hum Genet. 2015;16:79–102.

38. Deleye L, Tilleman L, Vander Plaetsen A-S, Cornelis S, Deforce D, Van Nieuwerburgh F. Performance of four modern whole genome amplification methods for copy number variant detection in single cells. Sci Rep. 2017;7:3422–3422.

39. Duan M, Hao J, Cui S, Worthley DL, Zhang S, Wang Z, et al. Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing. Cell Research. 2018;28:359–73.

40. Hughes AEO, Magrini V, Demeter R, Miller CA, Fulton R, Fulton LL, et al. Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. PLoS Genet. 2014;10:e1004462–e1004462.

41. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nature Methods. 2015;12:519.

42. Neves RPL, Raba K, Schmidt O, Honisch E, Meier-Stiegen F, Behrens B, et al. Genomic High-Resolution Profiling of Single $CK^{pos}/CD45^{neg}$ Flow-Sorting Purified Circulating Tumor Cells from Patients with Metastatic Breast Cancer. Clin Chem. 2014;60:1290.

43. Paolillo C, Mu Z, Rossi G, Schiewer MJ, Nguyen T, Austin L, et al. Detection of Activating Estrogen Receptor Gene (ESR1) Mutations in Single Circulating Tumor Cells. Clin Cancer Res. 2017;23:6086–93.

44. Rohrback S, April C, Kaper F, Rivera RR, Liu CS, Siddoway B, et al. Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. Proc Natl Acad Sci USA. 2018;115:10804.

45. Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, et al. Sequencing thousands of single-cell genomes with combinatorial indexing. Nat Methods. 2017;14:302–8.

46. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60.

47. Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, et al. Scalable whole-genome single-cell library preparation without preamplification. Nature Methods. 2017;14:167–73.

48. Burbulis IE, Wierman MB, Wolpert M, Haakenson M, Lopes M-B, Schiff D, et al. Improved molecular karyotyping in glioblastoma. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2018;811:16–26.

49. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. 2015;31:382–92.

50. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002;419:498–511.

51. Nkhoma SC, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, et al. Resolving within-host malaria parasite diversity using single-cell sequencing. bioRxiv. 2018;:391268.

52. Trevino SG, Nkhoma SC, Nair S, Daniel BJ, Moncada K, Khoswe S, et al. High-Resolution Single-Cell Sequencing of Malaria Parasites. Genome Biol Evol. 2017;9:3373–83.

53. Lasken RS, Stockwell TB. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnology. 2007;7:19.

54. Oyola SO, Manske M, Campino S, Claessens A, Hamilton WL, Kekre M, et al. Optimized whole-genome amplification strategy for extremely AT-biased template. DNA Res. 2014;21:661–71.

55. de Bourcy CFA, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR. A Quantitative Comparison of Single-Cell Whole Genome Amplification Methods. PLOS ONE. 2014;9:e105585.

56. Ning L, Li Z, Wang G, Hu W, Hou Q, Tong Y, et al. Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. Scientific Reports. 2015;5:11415.

57. Haynes JD, Diggs CL, Hines FA, Desjardins RE. Culture of human malaria parasites *Plasmodium falciparum*. Nature. 1976;263:767–9.

58. Bei AK, Desimone TM, Badiane AS, Ahouidi AD, Dieye T, Ndiaye D, et al. A flow cytometry-based assay for measuring invasion of red blood cells by *Plasmodium falciparum*. Am J Hematol. 2010;85:234–7.

59. Brown AC, Guler JL. SLOPE: A two-part method for the enrichment of ring stage *Plasmodium falciparum* parasites. bioRxiv. 2018;:474338.

60. T. Maniatis, Sambrook J, Fritsch EF. Molecular cloning: a laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.

61. Ribaut C, Berry A, Chevalley S, Reybier K, Morlais I, Parzy D, et al. Concentration and purification by magnetic separation of the erythrocytic stages of all human *Plasmodium* species. Malaria Journal. 2008;7:45.

62. Jensen MA, Fukushima M, Davis RW. DMSO and Betaine Greatly Improve Amplification of GC-Rich Constructs in De Novo Synthesis. PLOS ONE. 2010;5:e11024.

63. Pickard AL, Wongsrichanalai C, Purfield A, Kamwendo D, Emery K, Zalewski C, et al. Resistance to antimalarials in Southeast Asia and genetic polymorphisms in *pfmdr1*. Antimicrob Agents Chemother. 2003;47:2418–23.

64. Perandin F, Manca N, Calderaro A, Piccolo G, Galati L, Ricci L, et al. Development of a real-time PCR assay for detection of *Plasmodium falciparum*, *Plasmodium vivax*, and *Plasmodium ovale* for routine clinical diagnosis. J Clin Microbiol. 2004;42:1214–9.

65. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci USA. 2002;99:5261.

66. Huckaby AC, Granum CS, Carey MA, Szlachta K, Al-Barghouthi B, Wang Y-H, et al. Complex DNA structures trigger copy number variation across the *Plasmodium falciparum* genome. Nucleic Acids Research. 2018;47:1615–27.

67. Bushnell B. BBMap. http://sourceforge.net/projects/bbmap/ (2019). Accessed 1 May 2019.

68. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12:966–8.

69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

70. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 2012;28:2678–9.

71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

72. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

73. Cheong W-H, Tan Y-C, Yap S-J, Ng K-P. ClicO FS: an interactive web-based service of Circos. Bioinformatics. 2015;31:3685–7.

74. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science. 2017;356:189.

75. B. Marwick, K. Krishnamoorthy. Cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups. R software package version 0.2.0. https://github.com/benmarwick/cvequality. Accessed 1 Oct 2019.

76. Harrell F. Hmisc: Harrell miscellaneous (R package Version 4.3-0). https://CRAN.R-project.org/package=Hmisc. Accessed 1 May 2019.

77. Chen D, Zhen H, Qiu Y, Liu P, Zeng P, Xia J, et al. Comparison of single cell sequencing data between two whole genome amplification methods on two sequencing platforms. Sci Rep. 2018;8:4963–4963.

78. Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, et al. Gplots: Various R programming tools for plotting data. R package version 3.0.1.1. https://cran.r-project.org/web/packages/gplots/index.html. Accessed 1 Oct 2019.

79. 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.

80. 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.

81. Zhang X, Liang B, Xu X, Zhou F, Kong L, Shen J, et al. The comparison of the performance of four whole genome amplification kits on ion proton platform in copy number variation detection. Bioscience Reports. 2017;37. doi:10.1042/BSR20170252.

82. Wang Y, Gao Z, Xu Y, Li G, He L, Qian P. An evaluation of multiple annealing and looping based genome amplification using a synthetic bacterial community. The Chinese Society of Oceanography. 2016;35:131–6.

83. Ohkubo S, Muto A, Kawauchi Y, Yamao F, Osawa S. The ribosomal protein gene cluster of Mycoplasma capricolum. Molecular and General Genetics MGG. 1987;210:314–22.

84. Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UCM, Podowski RM, et al. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature. 1998;396:133–40.

85. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature. 1997;390:580–6.

86. Lorenzi HA, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, et al. New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information. PLoS Negl Trop Dis. 2010;4:e716–e716.

87. Ohama T, Muto A, Osawa S. Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus a bacterium with a high genomic GC-content. Nucleic Acids Research. 1990;18:1565–9.

88. Lasken RS. Single-cell sequencing in its prime. Nature Biotechnology. 2013;31:211–2.

89. Viguera E, Canceill D, Ehrlich SD. In vitro replication slippage by DNA polymerases from thermophilic organisms. Journal of Molecular Biology. 2001;312:323–33.

90. Ignatov KB, Barsova EV, Fradkov AF, Blagodatskikh KA, Kramarova TV, Kramarov VM. A strong strand displacement activity of thermostable DNA polymerase markedly improves the results of DNA amplification. BioTechniques. 2014;57:81–7.

91. Carey MA, Covelli V, Brown A, Medlock GL, Haaren M, Cooper JG, et al. Influential Parameters for the Analysis of Intracellular Parasite Metabolomics. mSphere. 2018;3:e00097-18.

92. Waldvogel Abramowski S, Tirefort D, Lau P, Guichebaron A, Taleb S, Modoux C, et al. Cell-free nucleic acids are present in blood products and regulate genes of innate immune response. Transfusion. 2018;58:1671–81.

93. 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. Malaria Journal. 2012;11:41.

94. Jacob CG, Tan JC, Miller BA, Tan A, Takala-Harrison S, Ferdig MT, et al. A microarray platform and novel SNP calling algorithm to evaluate *Plasmodium falciparum* field samples of low DNA quantity. BMC Genomics. 2014;15:719–719.

95. 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.

96. McFeters GA, Broadaway SC, Pyle BH, Egozy Y. Distribution of bacteria within operating laboratory water purification systems. Appl Environ Microbiol. 1993;59:1410–5.

97. Kulakov LA, McAlister MB, Ogden KL, Larkin MJ, O'Hanlon JF. Analysis of bacteria contaminating ultrapure water in industrial systems. Appl Environ Microbiol. 2002;68:1548–55.

98. Nogami T, Ohto T, Kawaguchi O, Zaitsu Y, Sasaki S. Estimation of Bacterial Contamination in Ultrapure Water: Application of the Anti-DNA Antibody. Anal Chem. 1998;70:5296–301.

99. Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S, et al. Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. PLOS ONE. 2011;6:e26161.

100. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology. 2014;12:87.

101. Rand KH, Houck H. Taq polymerase contains bacterial DNA of unknown origin. Molecular and Cellular Probes. 1990;4:445–50.

102. Kil E-J, Kim S, Lee Y-J, Kang E-H, Lee M, Cho S-H, et al. Advanced loop-mediated isothermal amplification method for sensitive and specific detection of Tomato chlorosis virus using a uracil DNA glycosylase to control carry-over contamination. Journal of Virological Methods. 2015;213:68–74.

103. Hou Y, Fan W, Yan L, Li R, Lian Y, Huang J, et al. Genome Analyses of Single Human Oocytes. Cell. 2013;155:1492–506.

104. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, et al. Interactive analysis and assessment of single-cell copy-number variations. Nature methods. 2015;12:1058–60.

105. Dong X, Zhang L, Hao X, Wang T, Vijg J. SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing. bioRxiv. 2019;:535807.

106. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4.

107. McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, et al. Mosaic copy number variation in human neurons. Science. 2013;342:632–7.

108. Fu Y, Li C, Lu S, Zhou W, Tang F, Xie XS, et al. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. Proc Natl Acad Sci USA. 2015;112:11923.

109. Chen M, Song P, Zou D, Hu X, Zhao S, Gao S, et al. Comparison of Multiple Displacement Amplification (MDA) and Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) in Single-Cell Sequencing. PLOS ONE. 2014;9:e114520.

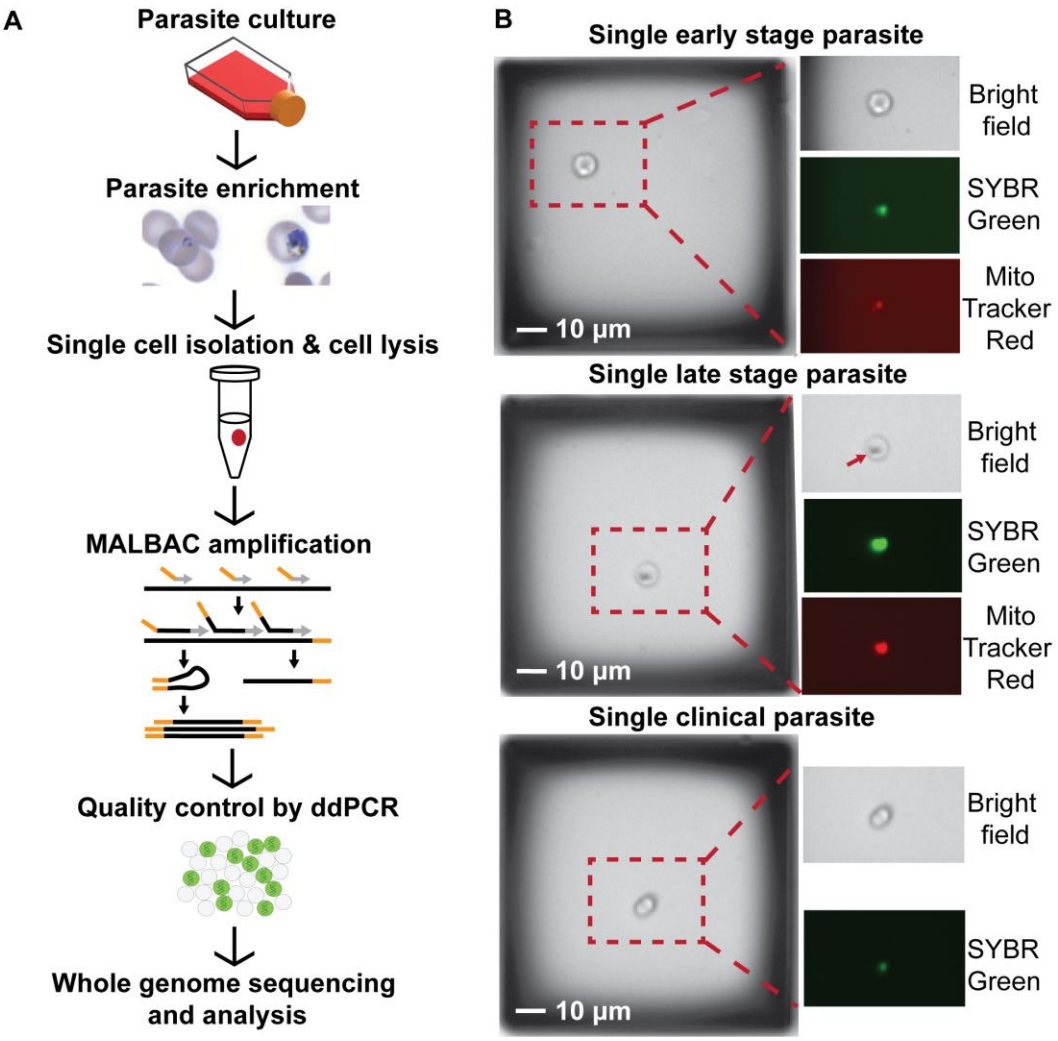

Figure 1. Single *P. falciparum*-infected erythrocytes are isolated, amplified, and sequenced.

A. **Experimental workflow.** Parasites are grown *in vitro* in human erythrocytes or isolated from infected patients. In order to limit the number of uninfected erythrocytes in the sample, infected cells are enriched using column and gradient-based methods (see *Methods*). Individual early-stage (left image) and late-stage (right image) parasite-infected erythrocytes were automatically isolated into PCR tubes using the CellRaft AIR System (Cell Microsystems, see panel B). All cells were lysed by combining a freeze–thaw step with detergent treatment prior to MALBAC amplification. MALBAC uses a combination of common (orange) and degenerate (grey) primers to amplify the genome. The quality of amplified genomes was assessed prior to library

preparation and sequencing using Droplet Digital (dd) PCR; DNA is partitioned into individual droplets to accurately measure gene copies. Suitable samples were Illumina sequenced and analyzed as detailed in Figure S2. **B. Parasite stage visualization on the CellRaft AIR System using microscopy** (10X magnification). Enriched early and late stage parasite-infected erythrocytes at low density were seeded into microwells to yield only a single cell per well (left image of each group), and identified with SYBR green and Mitotracker Red staining (indicates parasite DNA and mitochondrion, respectively). Early stage parasites exhibited lower fluorescence due to their smaller size and late stage parasites had noticeable dark spots (arrow) due to the accumulation of hemozoin pigment. Scale bar represents 10 μm.

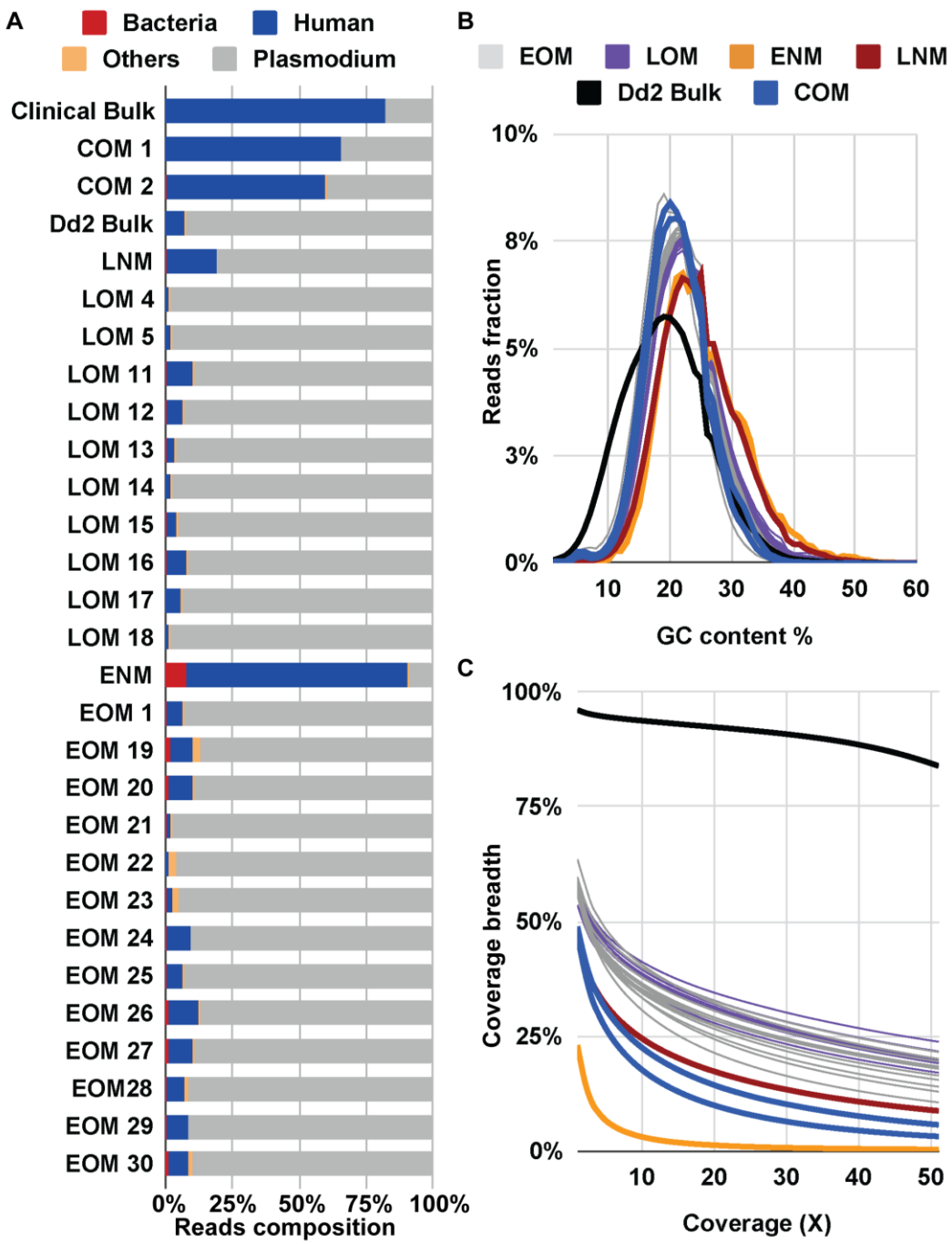

Figure 2. Sequencing statistics show benefits of optimized MALBAC. A. Contribution of reads based on organism type. A subset of 10,000 reads from each sample were randomly selected for BLAST to identify sources of DNA. Color representation: bacteria (red); human (blue); other organisms (orange); *Plasmodium* (grey). **B**. **GC-content of *P. falciparum* mapped reads.** GC-content of reads was calculated by Qualimap with default parameters. Color

representation: EOM (grey): Early stage single parasites amplified by optimized MALBAC; LOM (purple): Late stage single parasites amplified by optimized MALBAC; ENM (orange): Early stage single parasites amplified by non-optimized MALBAC; LNM (dark red): Late stage single parasites amplified by non-optimized MALBAC; Dd2 bulk genomic DNA (black); COM samples (blue): Clinical single parasites amplified by optimized MALBAC. Clinical Bulk genomic DNA is not shown here due to <1% of the genome being covered by at least one read. **C. Fraction of *P. falciparum* genome covered by >1 read.** The fraction of the genome was calculated by Qualimap with default parameters. Color representations are the same as described in panel B.

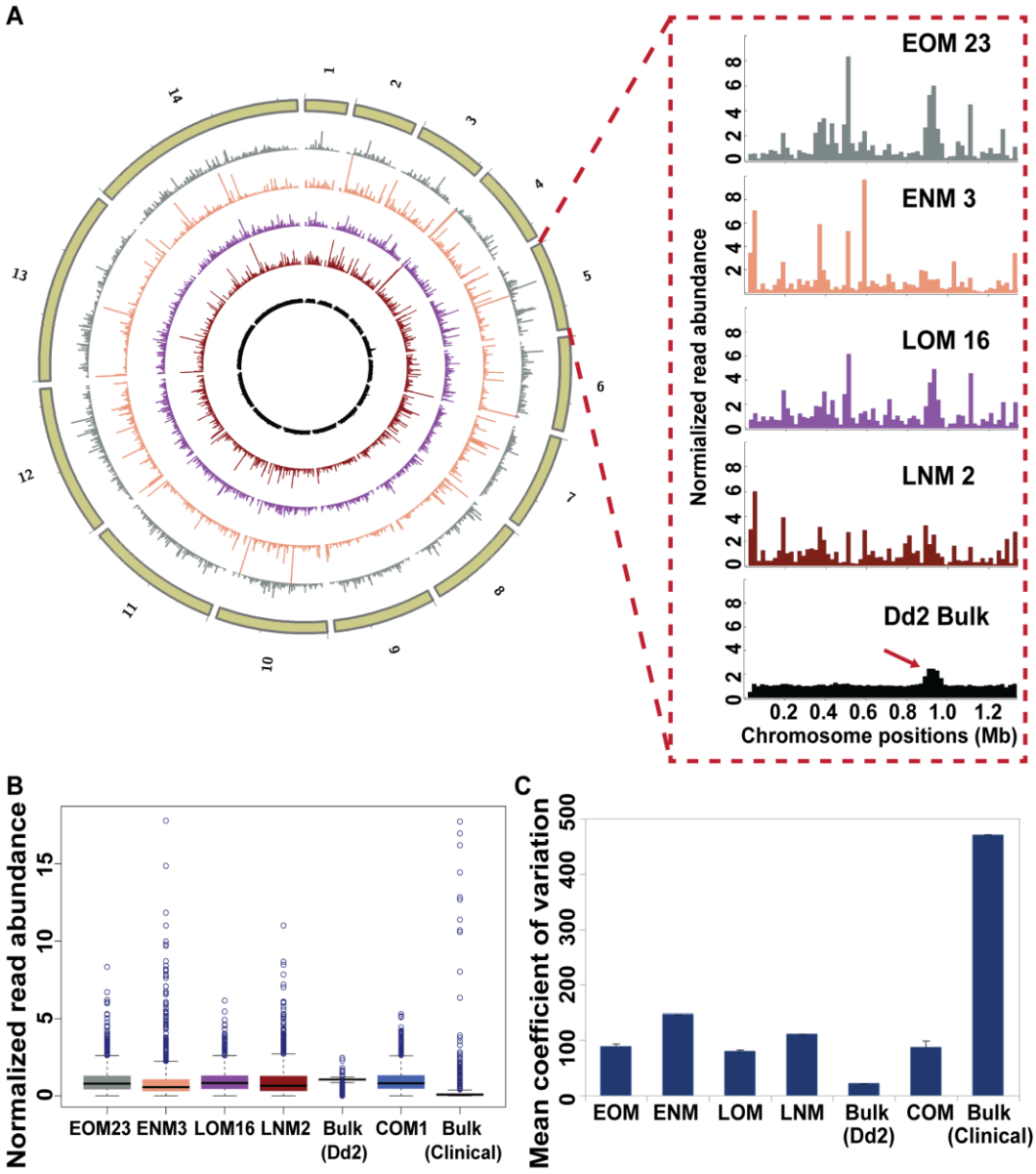

Figure 3. Samples amplified by optimized MALBAC display improved uniformity of read abundance. A. Normalized read abundance across the genome. The reference genome was divided into 20kb bins and read counts in each bin were normalized by the mean read count in each sample. The circles of the plot represent (from outside to inside): chromosomes 1 to 14 (tan); one EOM sample (#23, grey); one ENM sample (#3, orange); one LOM sample (#16, purple); one LNM sample (#2, dark red); Dd2 bulk genomic DNA (black). The zoomed panel shows the read distribution across chromosome 5, which has a known copy number variation

(arrow on Dd2 bulk sample). **B. Distribution of normalized read abundance values for all bins.** The boxes were drawn from Q1 (25th percentiles) to Q3 (75th percentiles) with a horizontal line drawn in the middle to denote the median of normalized read abundance for each sample. Outliers, above the highest point of the upper whisker (Q3 + 1.5×IQR) or below the lowest point of the lower whisker (Q1-1.5×IQR), are depicted with circles. One sample from each type is represented (see all samples in **Figure S3C**). **C. Coefficient of variation of normalized read abundance.** The average and SD (error bars) coefficient of variation for all samples from each type is represented (EOM: 13 samples; ENM: 1 sample; LOM: 10 samples; LNM: 1 sample; Dd2 Bulk: 1 sample; COM: 2 samples; Clinical Bulk: 1 sample). See *Methods* for calculation.

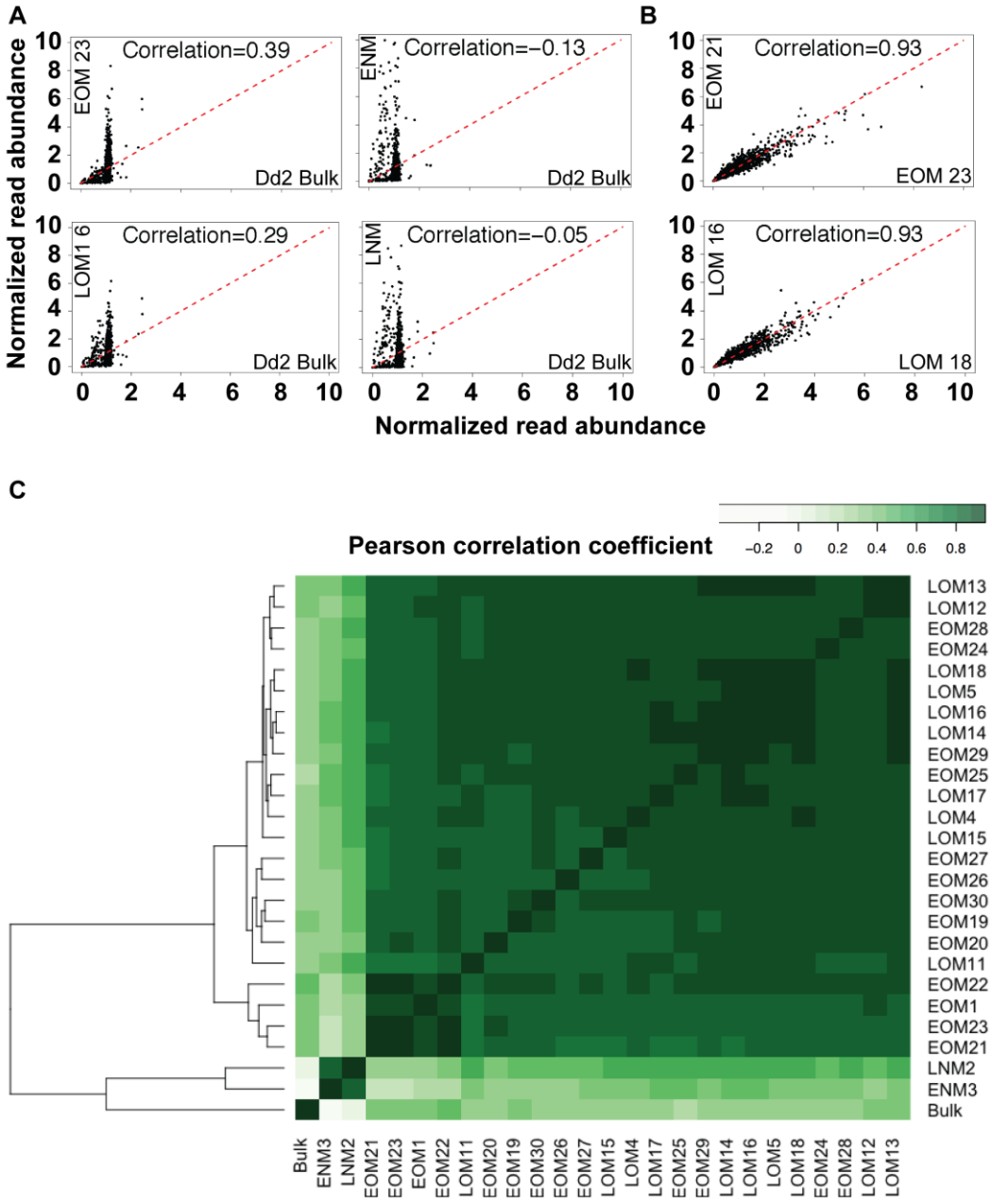

Figure 4. Correlations show reproducibility of amplification pattern by optimized MALBAC. A. Comparison of normalized read abundance (single cell versus bulk). The normalized read abundance in 20kb bins of one single cell sample is plotted against that of bulk sample. The Pearson correlation coefficient between the two samples is listed for each comparison. A narrow distribution of dots along the y = x (red dotted line) indicates a strong correlation between the patterns of read abundance. **B. Comparison of normalized read**

abundance (single cell versus single cell). Details are the same as in panel A except comparisons are between two single cells. **C. Pearson correlation coefficients between all sequenced samples.** The hierarchical clustering heatmap was generated using Pearson correlation coefficients of normalized read abundance. The color scale indicates the degree of correlation (white, low correlation < 0.3; green, high correlation >= 0.3).

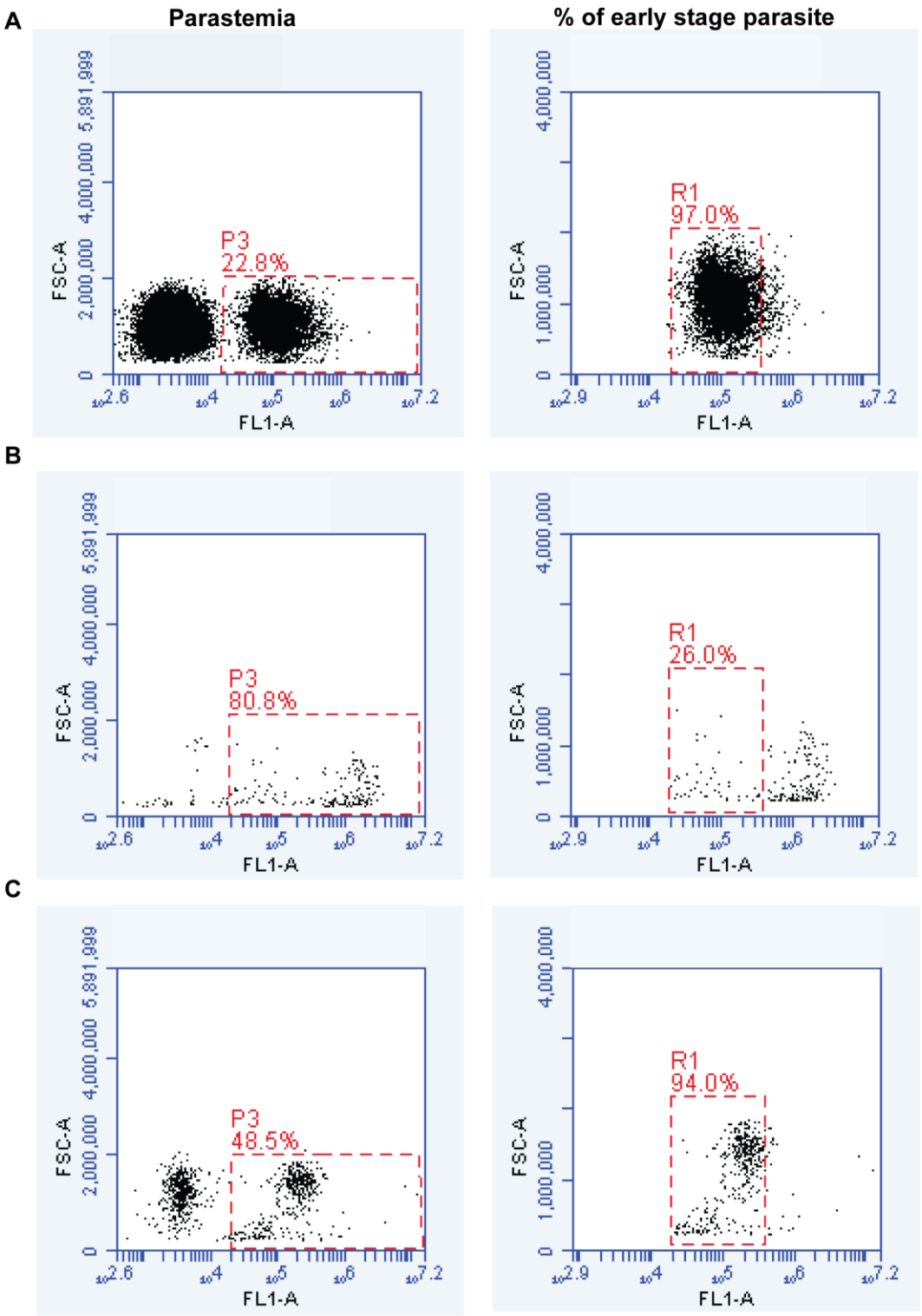

Figure S1. Confirmation of staging for enriched parasite samples. Parasitemia (proportion of infected erythrocytes; P3 gate, left plots) and proportion of early stage parasites (R1 gate, right plots) is shown for each sample. **A. Early stage laboratory parasite samples.** Early stage

parasites were enriched by harvesting the flow-through of the MACS column, which contained parasite stages that have not accumulated hemozoin, a paramagnetic crystal, along with uninfected cells. The SLOPE method was used to deplete the sample of uninfected cells to yield a high parasitemia (22.8%), predominantly early-staged population (97.0%) for single cell isolation. The flow cytometry run collected 30,485 total events. **B. Late stage laboratory parasite samples.** Late stage parasites were enriched by collecting the bound fraction from the MACS column, which contained parasite stages with high levels of paramagnetic hemozoin. This high parasitemia (80.8%), predominantly late stage population (74%), was used for single cell isolation. Late-stage parasites were further selected by higher fluorescence (due to increased DNA content and mitochondrial size) on the CellRaft microscope during isolation (see **Figure 1A**). The flow cytometry run collected 1357 total events. **C. Early stage clinical parasite samples**. Whole blood collected from a *P. falciparum*-infected patient was stored in sodium citrate for 23 hours in the hospital and incubated in RPMI for 48 hours in the laboratory (see *Methods* for details). The SLOPE method was used to deplete the sample of uninfected cells to yield a high parasitemia (48.5%), predominantly early-staged population (94.0%, confirmed by microscopy) for single cell isolation. The flow cytometry run collected 1314 total events.

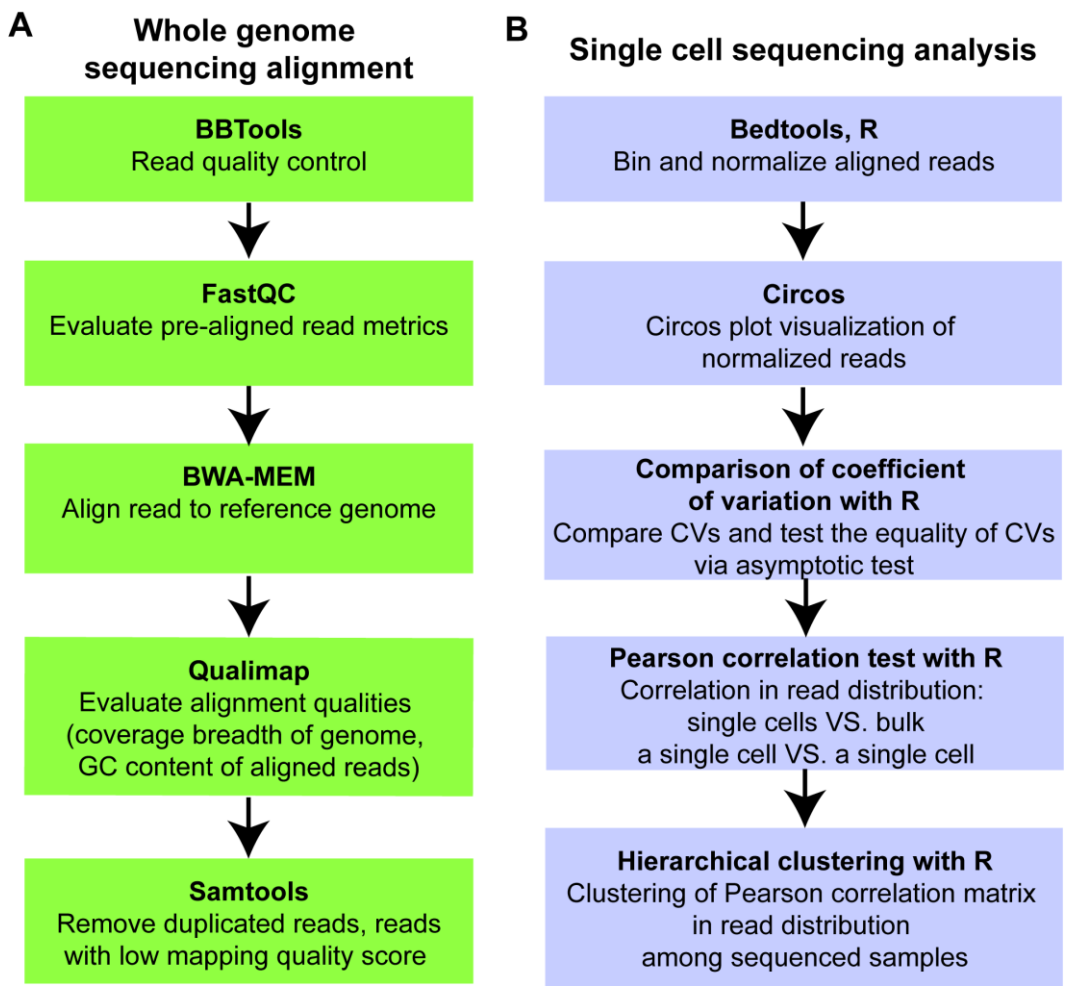

Figure S2. Bioinformatic analysis of sequencing reads. A. Whole genome sequencing analysis and alignment. Alignment of whole genome sequencing reads started with BBTools to remove low quality bases, adapter sequences, trim the common sequence of MALBAC primer and verify correct pairing of reads. The resulting "clean" paired reads were evaluated by FastQC for overrepresented sequences, per base read qualities, and read length distributions. After passing read quality control, BWA-MEM was used to align "clean" paired reads to the *3d7 Plasmodium falciparum* reference genome. Qualimap was then used to evaluate the alignments for the fraction of the genome covered by at least one read, GC-content of aligned reads, and mapping quality. Duplicated reads and reads with mapping quality score below 10 were removed for downstream analysis by Samtools. **B. Single cell sequencing analysis steps.** The reference

genome was divided into 20kb bins by Bedtools. Read counts was calculated in every 20kb bins and normalized by the mean read count in each sample with Bedtools and R. Circos was utilized to visualize the distribution of normalized read counts over 14 chromosomes. To compare variance of read counts in each sample, coefficient of variations in read counts were calculated with R and the equality of coefficient of variations was tested via Feltz and Miller's (1996) asymptotic test. Correlation coefficients between all combinations of pairs of sequenced samples were calculated via Pearson correlation test on normalized read counts in 20kb bins with R. Clustering of Pearson correlation matrix of normalized read counts in 20kb bins was conducted among sequenced samples with R.

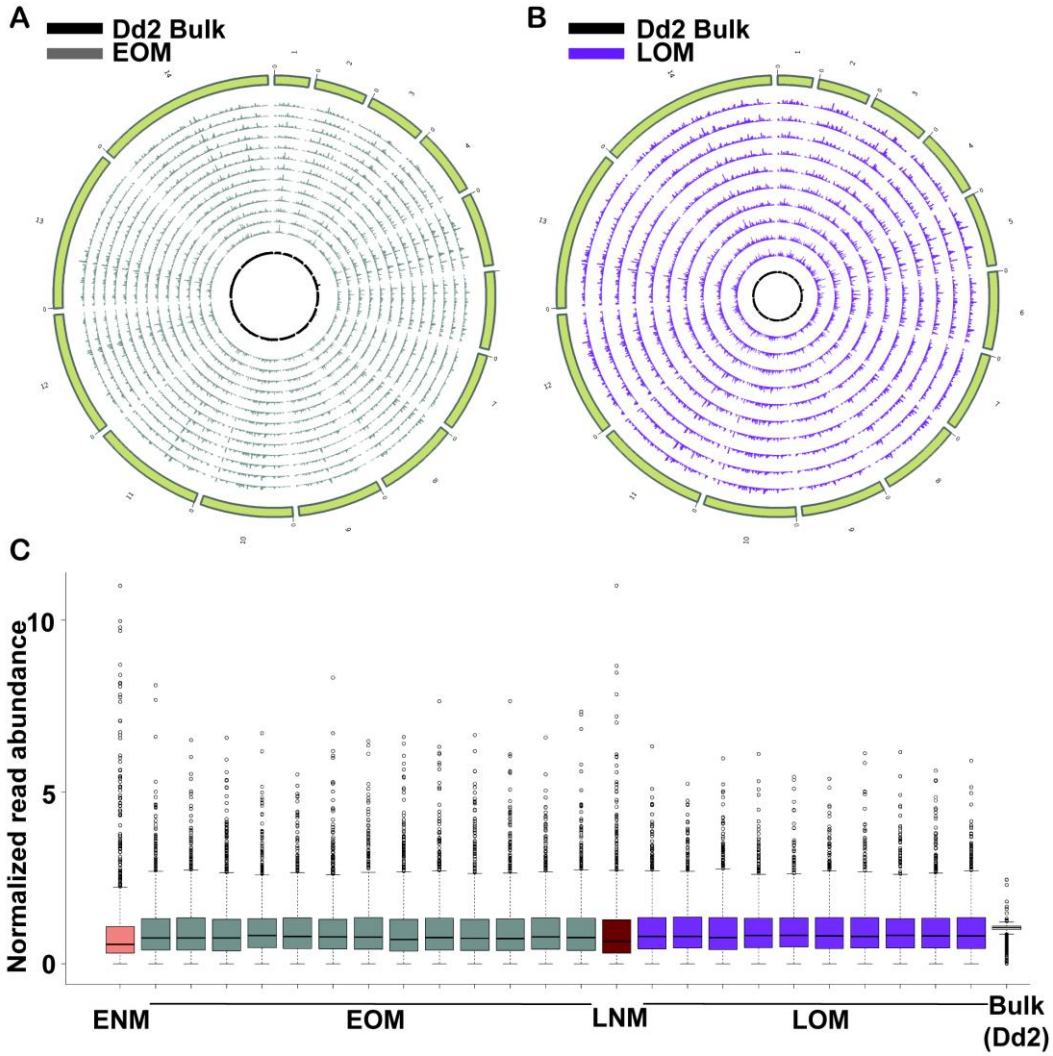

Figure S3. Uniformity of read abundance across the whole genome of all EOM and LOM samples. A & B. Normalized read abundance across the genome. Distribution of read abundance in 20kb bins in all EOM (A) and LOM (B) samples is shown. Read counts in each bin were normalized by the mean read count over the whole genome for each sample. The circles from outside to the inside represent: chromosomes 1:14 (tan); 13 EOM samples (grey) or 10 LOM samples (purple); Bulk genomic DNA (black). **C. Distribution of normalized read abundance values for all bins. N**ormalized read abundance per 20kb bins with outliers (circles) is represented.

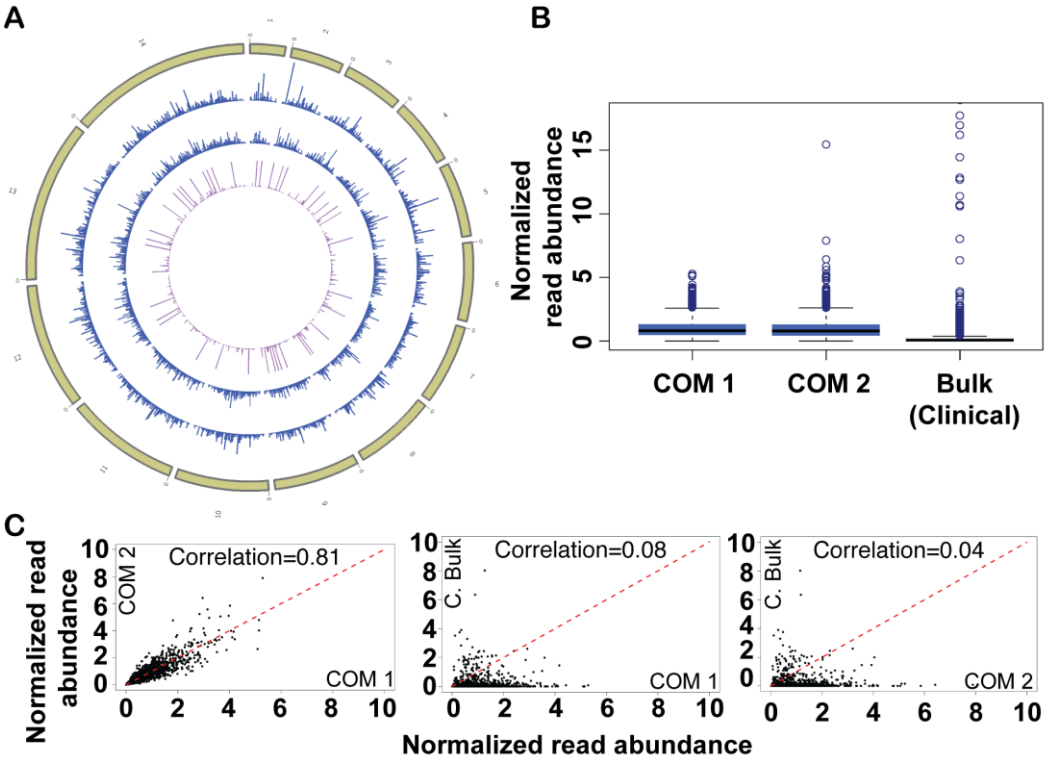

Figure S4. Uniformity of read abundance across the whole genome and correlation analysis for clinical samples. A. Normalized read abundance across the genome. Distribution of read abundance in 20kb bins of clinical samples is shown. Read counts in each bin were normalized by the mean read count over the whole genome for each sample. The circles from outside to the inside represent: chromosome 1:14 (green); 2 COM sample (blue); 1 Clinical bulk genomic DNA (pink). **B**. **Distribution of normalized read abundance values for all bins.** Box plot representing the normalized read abundance per 20kb bins, outliers are illustrated by blue dots. **C**. **Comparison of normalized read abundance (single cell vs single cell, single cell versus bulk).** The normalized read abundance in 20kb bins of single cell samples is compared to each other or that of the clinical bulk sample. The Pearson correlation coefficient between is listed for each comparison.

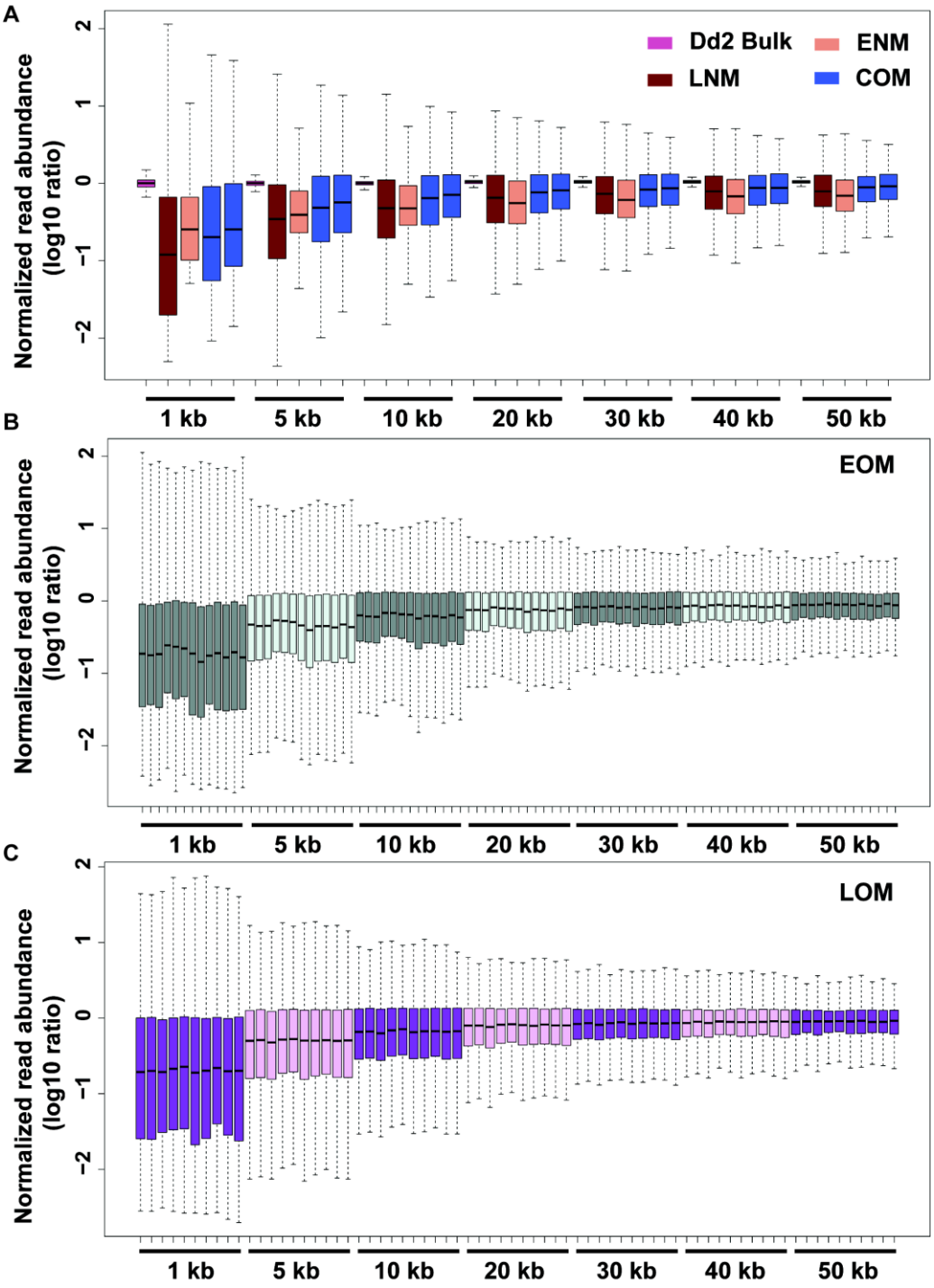

Figure S5. Distribution of normalized read counts in various bins sizes. The Log10 ratios of normalized read abundance in 1- 50kb (at intervals of 5 and 10kb) are showed for sequenced samples. The boxes indicate Q1 (25th percentiles) to Q3 (75th percentiles) with a horizontal line drawn in the middle to denote the median. Outliers, above the highest point of the upper whisker

(Q3 + 1.5×IQR) or below the lowest point of the lower whisker (Q1-1.5×IQR), are not displayed. **A. Distribution of normalized read counts in various bins sizes for select sample types.** Dd2 Bulk (purple), ENM (pink, 1 samples), LNM (maroon, 1 samples), COM (blue, 2 samples) samples. **B. Distribution of normalized read counts in various bin sizes for all EOM samples.** EOM (green, n=13). **C. Distribution of normalized read counts in various bin sizes for LOM samples.** LOM (purple, n=10).

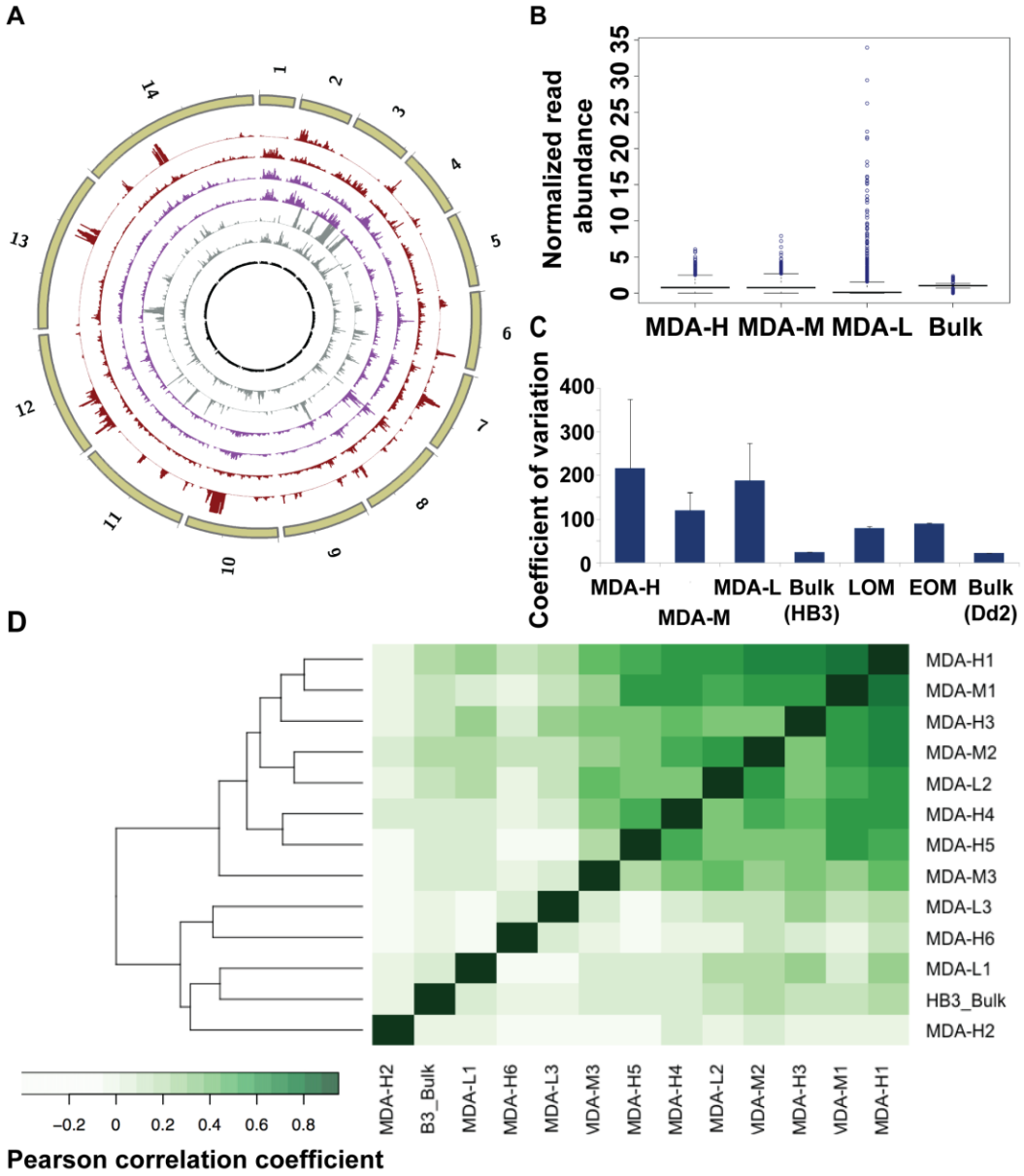

Figure S6. Uniformity of coverage and correlation analysis in MDA-amplified single cell samples. Samples for this analysis were amplified and sequence by Trevino et al., 2017 and analyzed using our pipeline (**Figure S2**). **A. Normalized read abundance across the genome.** Distribution of read abundance in 20kb bins of MDA-amplified samples is shown. Read counts in each bin were normalized by the mean read count over the whole genome for each sample. The circles from outside to the inside represent: chromosome 1:14 (tan); two MDA-H samples (dark red, HB3 parasite with high DNA content amplified by MDA); two MDA-M sample

(purple, HB3 parasite with medium DNA content amplified by MDA); two MDA-L samples (grey, parasite with low DNA content amplified by MDA); one HB3 Bulk sample. **B. Distribution of normalized read abundance values for all bins in MDA amplified samples.** MDA-H: 1 representative sample; MDA-M: 1 representative sample; MDA-L: 1 representative sample from (Trevino et al., 2017). **C. Comparison of coefficient of variation of normalized read abundance between MDA and MALBAC amplified single parasites.** The average and SD (error bars) of coefficient of variation of all samples from each type are represented (MDA-H: 6 samples; MDA-M: 3 samples; MDA-L: 3 samples; HB3 Bulk: 1 sample; LOM: 10 samples; EOM: 13 samples; Dd2 Bulk: 1 sample). **D. Pearson correlation coefficient between MDA amplified samples.** The hierarchical clustering heatmap was generated using Pearson correlation coefficients of normalized read abundance. The color scale indicates the degree of correlation (white, low correlation; green, high correlation).

Table S1. GC content of 5 base windows in the *P. falciparum* genome

GC content of 5-base windows	Regions					
	Whole genome		Genic regions		Intergenic regions	
	Count of windows	Percentage of windows	Count of windows	Percentage of windows	Count of windows	Percentage of windows
0%	8965149	38.49%	4170088	29.92%	4811633	51.39%
20%	8046176	34.54%	5180728	37.18%	2848042	30.42%
25%	6	0.00%	2761	0.02%	1713	0.02%
33%	10	0.00%	3588	0.03%	1487	0.02%
40%	4613827	19.81%	3312094	23.77%	1294355	13.83%
50%	11	0.00%	2060	0.01%	1575	0.02%
60%	1430717	6.14%	1079552	7.75%	349881	3.74%
67%	3	0.00%	39	0.00%	310	0.00%
75%	2	0.00%	23	0.00%	76	0.00%
80%	220354	0.95%	172049	1.23%	48285	0.52%
100%	16367	0.07%	12319	0.09%	4845	0.05%

Table S2. Primer and probe design for droplet digital PCR

Locus Name	Chr.	CN in *Dd2*	Ref.	Target			
				Primer/ probe	Final concentration	Sequence	Size
Pfseryl tRNA syn	7	1	[7]	F	900nM	5'-GGAACAATTCTGTATTGCTTTACC-3'	142bp
				R	900nM	5'-AAGCTGCGTTGTTTAAAGCTC-3'	
				Probe	250nM	5'-(VIC)-ACATGAAGAAATGATACAAACA-(IOWA BLACK FQ)-3'*	
Pfhsp70	8	1	-	F	600nM	5'-TGCTGTCATTACCGTTCCAG-3'	118bp
				R	600nM	5'-AGCAGCTGCAGTAGGTTCATT-3'	
				Probe	250nM	5'-(HEX)-AGATGCTGGTACAATTGCAGGA-(IOWA BLACK FQ)-3'*	
Pfdhfr	4	1	-	F	600nM	5'-CTCATATGATTGCACAAGTCTG-3'	133bp
				R	600nM	5'-GTTGGGAATGGATAGGGTATT-3'	
				Probe	250nM	5'-(FAM)-TTGCAACCTGCGCAGTTCATACAC-(IOWA BLACK FQ)-3'*	
Pfldh	13	1	[63]	F	600nM	5'-ACGATTTGGCTGGAGCAGAT-3'	85bp
				R	600nM	5'-TCTCTATTCCATTCTTTGTCACTCTTTC-3'	
				Probe	250nM	5'-(HEX)-AGTAATAGTAACAGCTGGATTTACCAAGGCCCCA-(IOWA BLACK FQ)-3'*	
Pf18S rDNA	1,11, 13	3	[64]	F	600nM	5'-CTTTTGAGAGGTTTTGTTACTTTGAGTAA-3'	99bp
				R	600nM	5'-TATTCCATGCTGTAGTATTCAAACACA-3'	
				Probe	250nM	5'-(FAM)-TGTTCATAACAGACGGGTAGTCATGATTGAGTTCA-(IOWA BLACK FQ)-3'*	
Pfmdr1	5	3	-	F	600nM	5'-TGCCCACAGAATTGCATCTA-3'	99bp
				R	600nM	5'-TCGTGTGTTCCATGTGACTG-3'	
				Probe	250nM	5'-(FAM)-ACCCTGATCGAAATGGAACCT-(IOWA BLACK FQ)-3'*	

*VIC, HEX and FAM, fluorescent reporter dyes; IOWA BLACK FQ: quencher (Integrated DNA Technologies).

Table S3. DNA yield after MALBAC amplification

Sample type	DNA yield after MALBAC and purification (ng)						
	EOM	**ENM**	**LOM**	**LNM**	**COM**	**NTC (opt)**	**NTC (Non-opt)**
Successfully amplification	**1408***	**1420**	**278**	**1480**	**1368**	Undetectable	1250
	1296	1740	**268**	1220	**1624**	Undetectable	-
	1264	2120	**241**	2425	**347.2**	Undetectable	-
	992	-	**239**	2265	**84.8**	Undetectable	-
	776	-	**232**	-	-	-	-
	568	-	**204**	-	-	-	-
	494	-	**199**	-	-	-	-
	456	-	**178**	-	-	-	-
	456	-	**172**	-	-	-	-
	436	-	**165**	-	-	-	-
	418	-	**144**	-	-	-	-
	417	-	**143**	-	-	-	-
	308	-	**142**	-	-	-	-
	152	-	**107**	-	-	-	-
	117.6	-	125	-	-	-	-
	79.6	-	119	-	-	-	-
	52.4	-	118	-	-	-	-
	164	-	105	-	-	-	-
Unsuccessful amplification	Undetectable (24 samples)	-	Undetectable (2 samples)	-	-	-	-
Success rate	43%	-	90%	-	-	-	-

*Bolded numbers represent samples being selected for ddPCR assessment.

Table S4. List of the coefficient variation of normalized read abundance in each sequenced sample

Batch	Sample name	Coefficient of Variation (CV) of normalized read abundance (%)	Mean CV in each type of samples (%)	SD of CVs in each type of samples
-	Dd2_Bulk	22	22	-
-	ENM3	147	147	-
-	EOM1	91	89	4
Batch-1	EOM19	86		
	EOM20	91		
	EOM25	95		
	EOM26	91		
	EOM28	92		
	EOM30	92		
Batch-2	EOM21	82		
	EOM22	81		
	EOM23	90		
Batch-3	EOM24	87		
	EOM27	91		
	EOM29	84		
-	LNM2	111	111	-
Same Batch	LOM11	84	79	2
	LOM12	77		
	LOM13	76		
	LOM14	80		
	LOM15	79		
	LOM16	78		
	LOM17	81		
	LOM18	79		
Same Batch	LOM4	82	87	12
	LOM5	79		
	COM1	78		
	COM2	96		
-	Clinical Bulk	472	472	-

SD, standard deviation.

Table S5. Summary of tests on the equality of CVs in normalized read abundance by sequenced samples

Sample name	Paired test on the equality of CVs between ENM3 and a EOM		Sample name	Paired test on the equality of CVs between LNM2 and a LOM	
	Test statistic	P-value		Test statistic	P-value
EOM1	66.76287	3.06E-16	LOM11	30.05426	4.20E-08
EOM19	86.53364	1.37E-20	LOM12	54.05277	1.95E-13
EOM20	68.38409	1.35E-16	LOM13	60.58646	7.04E-15
EOM21	103.6689	2.39E-24	LOM14	43.05383	5.33E-11
EOM22	110.2728	8.54E-26	LOM15	47.97997	4.31E-12
EOM23	73.3549	1.08E-17	LOM16	52.48721	4.33E-13
EOM24	81.76793	1.53E-19	LOM17	41.60157	1.12E-10
EOM25	56.81782	4.78E-14	LOM18	45.69159	1.38E-11
EOM26	67.57568	2.03E-16	LOM4	36.88771	1.25E-09
EOM27	66.63139	3.27E-16	LOM5	48.01785	4.22E-12
EOM28	65.25652	6.58E-16			
EOM29	97.36433	5.77E-23			
EOM30	65.9547	4.61E-16			

Table S6. Pearson correlation coefficient of all sequenced samples

Pearson correlation coefficient	LOM4	LOM5	LOM11	LOM12	LOM13	LOM14	LOM15	LOM16	LOM17	LOM18	EOM1	EOM19	EOM20	EOM21	EOM22	EOM23	EOM24	EOM25	EOM26	EOM27	EOM28	EOM29	EOM30	LNM2	ENM3	Dd2 Bulk	COM1	COM2	Clinical Bulk
LOM4	1.00	0.91	0.85	0.90	0.91	0.91	0.87	0.91	0.89	0.93	0.81	0.84	0.84	0.79	0.86	0.81	0.90	0.90	0.83	0.86	0.89	0.91	0.85	0.50	0.38	0.28	0.75	0.73	0.07
LOM5	0.91	1.00	0.88	0.91	0.93	0.94	0.90	0.93	0.92	0.93	0.83	0.87	0.86	0.79	0.88	0.82	0.92	0.91	0.86	0.88	0.91	0.92	0.87	0.51	0.38	0.31	0.76	0.72	0.06
LOM11	0.85	0.88	1.00	0.84	0.87	0.88	0.84	0.87	0.86	0.88	0.76	0.80	0.81	0.74	0.83	0.77	0.85	0.84	0.79	0.83	0.85	0.86	0.81	0.48	0.35	0.28	0.70	0.66	0.09
LOM12	0.90	0.91	0.84	1.00	0.95	0.92	0.89	0.91	0.92	0.92	0.86	0.87	0.86	0.80	0.89	0.84	0.91	0.91	0.87	0.86	0.91	0.91	0.89	0.45	0.31	0.33	0.77	0.74	0.06
LOM13	0.91	0.93	0.87	0.95	1.00	0.94	0.90	0.93	0.92	0.93	0.84	0.87	0.86	0.79	0.88	0.83	0.92	0.91	0.87	0.89	0.91	0.93	0.88	0.49	0.35	0.33	0.75	0.72	0.08
LOM14	0.91	0.94	0.88	0.92	0.94	1.00	0.90	0.94	0.93	0.93	0.81	0.86	0.86	0.76	0.86	0.80	0.92	0.93	0.88	0.89	0.92	0.93	0.89	0.54	0.41	0.29	0.73	0.71	0.06
LOM15	0.87	0.90	0.84	0.89	0.90	0.90	1.00	0.90	0.89	0.90	0.79	0.83	0.83	0.75	0.85	0.79	0.88	0.88	0.83	0.84	0.87	0.88	0.86	0.49	0.36	0.31	0.73	0.69	0.05
LOM16	0.91	0.93	0.87	0.91	0.93	0.94	0.90	1.00	0.93	0.93	0.82	0.87	0.86	0.78	0.86	0.81	0.92	0.92	0.87	0.88	0.92	0.93	0.88	0.55	0.43	0.29	0.74	0.72	0.04
LOM17	0.89	0.92	0.86	0.92	0.92	0.93	0.89	0.93	1.00	0.91	0.81	0.85	0.84	0.75	0.84	0.79	0.90	0.91	0.87	0.86	0.91	0.91	0.86	0.53	0.40	0.27	0.74	0.70	0.07
LOM18	0.93	0.93	0.88	0.92	0.93	0.93	0.90	0.93	0.91	1.00	0.83	0.87	0.88	0.81	0.89	0.85	0.92	0.91	0.86	0.90	0.90	0.93	0.88	0.49	0.37	0.32	0.78	0.75	0.06
EOM1	0.81	0.83	0.76	0.86	0.84	0.81	0.79	0.82	0.81	0.83	1.00	0.83	0.82	0.86	0.90	0.87	0.84	0.81	0.79	0.81	0.85	0.83	0.83	0.31	0.19	0.35	0.80	0.75	0.05
EOM19	0.84	0.87	0.80	0.87	0.87	0.86	0.83	0.87	0.85	0.87	0.83	1.00	0.85	0.82	0.88	0.83	0.85	0.85	0.84	0.83	0.87	0.85	0.86	0.41	0.30	0.32	0.76	0.73	0.06
EOM20	0.84	0.86	0.81	0.86	0.86	0.86	0.83	0.86	0.84	0.88	0.82	0.85	1.00	0.83	0.89	0.85	0.86	0.86	0.82	0.84	0.86	0.85	0.84	0.39	0.26	0.32	0.78	0.72	0.04
EOM21	0.79	0.79	0.74	0.80	0.79	0.76	0.75	0.78	0.75	0.81	0.86	0.82	0.83	1.00	0.94	0.93	0.80	0.75	0.75	0.77	0.80	0.77	0.80	0.27	0.16	0.39	0.85	0.77	0.06
EOM22	0.86	0.88	0.83	0.89	0.88	0.86	0.85	0.86	0.84	0.89	0.90	0.88	0.89	0.94	1.00	0.95	0.88	0.86	0.84	0.86	0.88	0.86	0.87	0.34	0.20	0.40	0.86	0.79	0.07
EOM23	0.81	0.82	0.77	0.84	0.83	0.80	0.79	0.81	0.79	0.85	0.87	0.83	0.85	0.93	0.95	1.00	0.84	0.80	0.78	0.82	0.83	0.82	0.82	0.29	0.17	0.39	0.84	0.77	0.05
EOM24	0.90	0.92	0.85	0.91	0.92	0.92	0.88	0.92	0.90	0.92	0.84	0.85	0.86	0.80	0.88	0.84	1.00	0.91	0.86	0.87	0.91	0.92	0.88	0.47	0.35	0.31	0.74	0.73	0.06
EOM25	0.90	0.91	0.84	0.91	0.91	0.93	0.88	0.92	0.91	0.91	0.81	0.85	0.86	0.75	0.86	0.80	0.91	1.00	0.88	0.87	0.92	0.92	0.89	0.52	0.41	0.23	0.72	0.72	0.05
EOM26	0.83	0.86	0.79	0.87	0.87	0.88	0.83	0.87	0.87	0.86	0.79	0.84	0.82	0.75	0.84	0.78	0.86	0.88	1.00	0.84	0.89	0.87	0.84	0.44	0.32	0.29	0.71	0.69	0.07
EOM27	0.86	0.88	0.83	0.86	0.89	0.89	0.84	0.88	0.86	0.90	0.81	0.83	0.84	0.77	0.86	0.82	0.87	0.87	0.84	1.00	0.88	0.89	0.85	0.46	0.34	0.30	0.74	0.72	0.06
EOM28	0.89	0.91	0.85	0.91	0.91	0.92	0.87	0.92	0.91	0.90	0.85	0.87	0.86	0.80	0.88	0.83	0.91	0.92	0.89	0.88	1.00	0.91	0.89	0.47	0.35	0.30	0.76	0.74	0.07
EOM29	0.91	0.92	0.86	0.91	0.93	0.93	0.88	0.93	0.91	0.93	0.83	0.85	0.85	0.77	0.86	0.82	0.92	0.92	0.87	0.89	0.91	1.00	0.87	0.52	0.39	0.28	0.73	0.72	0.04
EOM30	0.85	0.87	0.81	0.89	0.88	0.89	0.86	0.88	0.86	0.88	0.83	0.86	0.84	0.80	0.87	0.82	0.88	0.89	0.84	0.85	0.89	0.87	1.00	0.40	0.29	0.30	0.76	0.73	0.04
LNM2	0.50	0.51	0.48	0.45	0.49	0.54	0.49	0.55	0.53	0.49	0.31	0.41	0.39	0.27	0.34	0.29	0.47	0.52	0.44	0.46	0.47	0.52	0.40	1.00	0.83	-0.05	0.27	0.29	0.02
ENM3	0.38	0.38	0.35	0.31	0.35	0.41	0.36	0.43	0.40	0.37	0.19	0.30	0.26	0.16	0.20	0.17	0.35	0.41	0.32	0.34	0.35	0.39	0.29	0.83	1.00	-0.13	0.15	0.19	0.01
Dd2 Bulk	0.28	0.31	0.28	0.33	0.33	0.29	0.31	0.29	0.27	0.32	0.35	0.32	0.32	0.39	0.40	0.39	0.31	0.23	0.29	0.30	0.30	0.28	0.30	-0.05	-0.13	1.00	0.30	0.20	0.06
COM1	0.75	0.76	0.70	0.77	0.75	0.73	0.73	0.74	0.74	0.78	0.80	0.76	0.78	0.85	0.86	0.84	0.74	0.72	0.71	0.74	0.76	0.73	0.76	0.27	0.15	0.30	1.00	0.81	0.08
COM2	0.73	0.72	0.66	0.74	0.72	0.71	0.69	0.72	0.70	0.75	0.75	0.73	0.72	0.77	0.79	0.77	0.73	0.72	0.69	0.72	0.74	0.72	0.73	0.29	0.19	0.20	0.81	1.00	0.04
Clinical Bulk	0.07	0.06	0.09	0.06	0.08	0.06	0.05	0.04	0.07	0.06	0.05	0.06	0.04	0.06	0.07	0.05	0.06	0.05	0.07	0.06	0.07	0.04	0.04	0.02	0.01	0.06	0.08	0.04	1.00

Table S7. List of the coefficient variation of normalized read abundance in MDA amplified samples

Sample name	Coefficient of Variation (CV) of normalized read abundance (%)	Mean CV in each type of samples (%)	SD of CVs in each type of samples
HB3_Bulk	24	24	-
MDA-H1	83	217	157
MDA-H2	324		
MDA-H3	127		
MDA-H4	134		
MDA-H5	148		
MDA-H6	489		
MDA-M1	103	120	40
MDA-M2	91		
MDA-M3	166		
MDA-L1	166	188	86
MDA-L2	115		
MDA-L3	283		

SD, standard deviation.

Table S8. Pearson correlation coefficient of MDA amplified samples

Correlation coefficient	MDA-H1	MDA-H2	MDA-H3	MDA-H4	MDA-H5	MDA-H6	MDA-M1	MDA-M2	MDA-M3	MDA-L1	MDA-L2	MDA-L3	HB3_Bulk
MDA-H1	1.00	0.09	0.68	0.63	0.55	0.21	0.75	0.70	0.47	0.37	0.66	0.26	0.31
MDA-H2	0.09	1.00	0.09	0.12	-0.02	-0.02	0.09	0.14	0.03	0.06	0.09	0.00	0.05
MDA-H3	0.68	0.09	1.00	0.49	0.43	0.12	0.64	0.45	0.42	0.33	0.44	0.35	0.20
MDA-H4	0.63	0.12	0.49	1.00	0.58	0.11	0.64	0.56	0.40	0.16	0.43	0.13	0.17
MDA-H5	0.55	-0.02	0.43	0.58	1.00	0.04	0.66	0.41	0.27	0.13	0.40	0.03	0.16
MDA-H6	0.21	-0.02	0.12	0.11	0.04	1.00	0.09	0.21	0.08	-0.02	0.09	0.16	0.08
MDA-M1	0.75	0.09	0.64	0.64	0.66	0.09	1.00	0.66	0.32	0.18	0.57	0.23	0.24
MDA-M2	0.70	0.14	0.45	0.56	0.41	0.21	0.66	1.00	0.32	0.29	0.65	0.20	0.30
MDA-M3	0.47	0.03	0.42	0.40	0.27	0.08	0.32	0.32	1.00	0.18	0.47	0.16	0.13
MDA-L1	0.37	0.06	0.33	0.16	0.13	-0.02	0.18	0.29	0.18	1.00	0.27	0.04	0.18
MDA-L2	0.66	0.09	0.44	0.43	0.40	0.09	0.57	0.65	0.47	0.27	1.00	0.20	0.25
MDA-L3	0.26	0.00	0.35	0.13	0.03	0.16	0.23	0.20	0.16	0.04	0.20	1.00	0.05
HB3_Bulk	0.31	0.05	0.20	0.17	0.16	0.08	0.24	0.30	0.13	0.18	0.25	0.05	1.00

Table S9. Proportion of aligned reads containing MALBAC primer common sequence overlapped with genomic regions

Sample name	% aligned reads containing MALBAC primer along the genome	% aligned reads with MALBAC primer overlapped with genomic regions	
		Genic regions	Intergenic regions
Dd2_Bulk	0.04%	92.99%	7.01%
COM 1	37.72%	90.05%	9.95%
COM 2	35.39%	92.69%	7.31%
ENM	23.43%	84.91%	15.09%
LNM	25.87%	94.24%	5.76%
EOM 21	39.23%	93.79%	6.21%
EOM 23	38.48%	95.19%	4.81%
LOM 16	27.85%	95.41%	4.59%
LOM 18	27.60%	95.79%	4.21%

Table S10. Overall sequencing reads output and percentage aligned reads to P. falciparum genome

Sample name	Total reads	Reads aligned (Q>=10, non-duplicated)	% Reads aligned (Q>=10, non-duplicated)	Mean coverage (X)	SD coverage (X)	Coverage breadth %
Dd2_Bulk	14088518	12451854	88.38%	75.8308	39.7742	96.12%
Bulk clinical	1798478	31179	1.73%	0.0345	4.137	0.3%
COM 1	3703368	1348000	36.40%	7.4389	23.5745	47.03%
COM 2	5199958	2079068	39.98%	11.6436	40.9011	48.87%
ENM	3605934	330429	9.16%	1.4713	10.6785	23.04%
LNM	5441240	3843328	70.63%	20.4289	113.9654	47.43%
LOM 4	7716822	6981687	90.47%	39.5272	112.0864	57.42%
LOM 5	7960574	7162154	89.97%	40.3766	130.3619	57.64%
LOM 11	8435214	6572607	77.92%	36.7722	185.5418	53.56%
LOM 12	8247324	7125641	86.40%	40.1058	121.7324	58.25%
LOM 13	8440624	7491252	88.75%	41.9726	128.0252	59.33%
LOM 14	8355392	7499920	89.76%	42.309	138.4362	56.23%
LOM 15	8755272	7713286	88.10%	42.8325	135.5189	55.89%
LOM 16	9144356	7401206	80.94%	41.2309	137.0799	57.71%
LOM 17	10547490	9107689	86.35%	50.2924	180.1848	58.09%
LOM 18	10855318	9879263	91.01%	55.5314	153.4196	59.07%
EOM 1	5879596	5134985	87.34%	29.045	94.6642	57.12%
EOM 19	8522966	6990566	82.02%	37.4556	121.5732	56.71%
EOM 20	6916038	5821742	84.18%	32.6576	102.4475	56.62%
EOM 21	4381554	3989410	91.05%	21.9652	74.3972	58.59%
EOM 22	9211442	8441554	91.64%	45.0949	133.0433	63.56%
EOM 23	5557586	5004172	90.04%	26.8804	87.347	58.99%
EOM 24	8002608	6636807	82.93%	37.5702	109.667	56.46%
EOM 25	9315768	7830782	84.06%	43.9508	145.9258	55.65%
EOM 26	8492108	6808137	80.17%	37.754	132.5401	56.76%
EOM 27	9241720	7656595	82.85%	42.7219	135.2933	57.20%
EOM28	9166292	7767520	84.74%	42.704	136.5738	56.78%
EOM 29	10415486	8719933	83.72%	49.1051	138.0999	59.68%
EOM 30	8883286	7431044	83.65%	41.1568	131.6227	58.06%
HB3_bulk	17,719,090	16,189,815	91.37%	67.2147	48.8778	95.51%
HB3_L_single_cell	12,245,390	10,488,358	85.65%	41.7289	183.9036	66.40%
HB3_L_single_cell	12,278,504	11,389,251	92.76%	45.9091	69.8633	94.14%
HB3_L_single_cell	18,664,566	17,344,347	92.93%	70.812	139.5796	82.10%
HB3_M_single_cell	11,392,932	10,499,255	92.16%	42.6567	50.188	95.19%
HB3_M_single_cell	6,880,990	6,253,026	90.87%	25.2398	52.9048	86.70%
HB3_M_single_cell	6084642	5510823	90.57%	22.3326	30.0916	93.64%
HB3_H_single_cell	3,895,054	3,397,493	87.23%	13.8257	78.6302	47.35%
HB3_H_single_cell	11,111,292	9,920,995	89.29%	40.2235	73.7899	91.95%
HB3_H_single_cell	3,658,764	3,198,863	87.43%	13.0053	20.4314	84.00%
HB3_H_single_cell	8,534,604	7,957,119	93.23%	32.1219	125.2137	55.52%
HB3_H_single_cell	5,473,840	4,919,047	89.86%	20.0882	22.2623	94.20%
HB3_H_single_cell	9,999,412	8,744,829	87.45%	35.4627	58.1737	92.44%

SD, standard deviation.