DNA Research, 2016, 23(4), 339–351
doi: 10.1093/dnares/dsw022
Advance Access Publication Date: 26 June 2016
Full Paper

OXFORD

Full Paper

Complete telomere-to-telomere *de novo* assembly of the *Plasmodium falciparum* genome through long-read (>11 kb), single molecule, real-time sequencing

Shruthi Sridhar Vembar[1,2,3,*], Matthew Seetin[4], Christine Lambert[4], Maria Nattestad[5], Michael C. Schatz[5,6], Primo Baybayan[4], Artur Scherf[1,2,3], and Melissa Laird Smith[4,*]

[1]Unité Biologie des Interactions Hôte-Parasite, Département de Parasites et Insectes Vecteurs, Institut Pasteur, Paris 75015, France, [2]CNRS, ERL 9195, Paris 75015, France, [3]INSERM, Unit U1201, Paris 75015, France, [4]Pacific Biosciences, Menlo Park, CA, USA, [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA, and [6]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

*To whom correspondence should be addressed. Tel. +33145688622. Email: shruthi-sridhar.vembar@pasteur.fr (S.S.V.); Tel. +1-650-521-8240. Email: mlaird@pacificbiosciences.com (M.L.S.)

Edited by Dr Yuji Kohara

Received 24 November 2015; Accepted 10 May 2016

Abstract

The application of next-generation sequencing to estimate genetic diversity of *Plasmodium falciparum*, the most lethal malaria parasite, has proved challenging due to the skewed AT-richness [~80.6% (A + T)] of its genome and the lack of technology to assemble highly polymorphic subtelomeric regions that contain clonally variant, multigene virulence families (Ex: *var* and *rifin*). To address this, we performed amplification-free, single molecule, real-time sequencing of *P. falciparum* genomic DNA and generated reads of average length 12 kb, with 50% of the reads between 15.5 and 50 kb in length. Next, using the Hierarchical Genome Assembly Process, we assembled the *P. falciparum* genome *de novo* and successfully compiled all 14 nuclear chromosomes telomere-to-telomere. We also accurately resolved centromeres [~90–99% (A + T)] and subtelomeric regions and identified large insertions and duplications that add extra *var* and *rifin* genes to the genome, along with smaller structural variants such as homopolymer tract expansions. Overall, we show that amplification-free, long-read sequencing combined with *de novo* assembly overcomes major challenges inherent to studying the *P. falciparum* genome. Indeed, this technology may not only identify the polymorphic and repetitive subtelomeric sequences of parasite populations from endemic areas but may also evaluate structural variation linked to virulence, drug resistance and disease transmission.

Key words: *Plasmodium falciparum*, AT-biased, long-read sequencing, *de novo* assembly, structural variation

1. Introduction

Plasmodium falciparum, the causative agent of the most lethal form of malaria, uses a complex developmental programme to propagate within the human host and mosquito vector.[1] Disease pathogenesis in humans correlates with asexual development when the parasite resides within mature erythrocytes and mitotically replicates its haploid genome. In addition to meiotic recombination during diploid stages in mosquitoes,[2] it is during mitosis that the parasite expands its genetic diversity via homologous recombination, leading to the acquisition of new variants of virulence-associated surface adhesion molecules such as erythrocyte membrane protein 1 [*P. falciparum* Erythrocyte Membrane Protein 1(PfEMP1); encoded by *var* genes].[3–5] Importantly, this is the stage at which the parasite evolves drug resistance.[6,7] In fact, it has been estimated that for each cycle of intraerythrocytic replication, laboratory-adapted parasites can tolerate a single nucleotide polymorphism (SNP) mutation rate of approximately $0.5–1 \times 10^{-9}$ per base,[3,4] and that nearly 0.2% of daughter parasites carry a new chimeric PfEMP1 molecule.[4] Furthermore, the heterogeneity within the human host can be amplified by multiple mosquito bites, which may harbour genetically diverse parasite populations. There is, therefore, great interest in evaluating the genetic complexity of the infectious pool of parasites in malaria-endemic regions, with the specific aims of improving surveillance and intervention strategies.

The ~23 Mb *P. falciparum* genome is organized into 14 chromosomes that range in size from 0.65 to 3.4 Mb. The draft sequence of the *P. falciparum* genome, which was first reported in 2002 using shotgun-sequencing methods for the laboratory-adapted strain 3D7,[8] revealed that the genome has an overall (A + T) composition of 80.6%, making it one of the most AT-rich genomes identified to date. The complexity of the genome is further underscored by the presence of extended tracts of As, Ts and TAs, in introns, intergenic and centromeric regions [up to 99% (A + T) content]; subtelomeric, hypervariable multigene virulence families, including the ~60-member *var* gene family[9]; and large segments of repetitive sequences, especially in subtelomeric regions. Given these unique features, not only has accurate sequencing of *P. falciparum* DNA presented a technical challenge for most next-generation sequencing (NGS) technologies,[10–14] it has been suggested that the use of the current 3D7 genome sequence as a reference for clinical isolates results in incomplete estimates of genetic diversity.[15,16]

To date, researchers have primarily used PCR-based whole genome amplification (WGA) methods to prepare short read sequencing libraries of *P. falciparum* laboratory-adapted and clinical strains,[3,4,6,17–21] with Nair et al. applying this to single cell sequencing.[22] More recently, Oyola et al. used φ29 DNA polymerase-based multiple displacement alignment (MDA), in the presence of the detergent tetramethylammonium chloride, to analyse the *P. falciparum* genome and showed that very low quantities of genomic DNA (~10 pg) were sufficient to generate multiplexed Illumina libraries.[11] However, the introduction of errors and bias during PCR-based WGA,[23] and the subsequent alignment-based mapping of short reads to the reference 3D7 genome[8] (http://genedb.org; 23.3 Mb assembly) may have led to an overestimation of SNPs in the sequencing data. Indeed, Oyola et al. observed that MDA introduces several per cent (2–6%) of *de novo* SNP calls as compared to a non-amplified library.[11] Furthermore, none of these studies analysed larger structural variants, except for Bopp et al.[3] and Claessens et al.[4] Therefore, we have a fragmented view of *P. falciparum* genome plasticity, in which SNPs are evaluated at high frequency but polymorphisms such as insertions and deletions, copy number variants, chromosomal rearrangements and structural variants in hypervariable and highly repetitive regions, are often underestimated or largely ignored.[16]

One solution that may overcome all of these caveats is the utilization of amplification-free long-read NGS technologies to sequence the *P. falciparum* genome. Single molecule real-time (SMRT) sequencing, which was the first such technology described,[24] generates long-reads with little to no sequence context bias,[13,14,25] with the most recent version of the DNA polymerase (P6) combined with C4 sequencing chemistry producing reads of average length 10–15 kb.[25] Numerous studies have shown that by oversampling a genome, structural variants can be detected with confidence,[26,27] and *de novo* assembly can be performed with high accuracy.[28–30] Attempts to analyse *P. falciparum* genomic DNA with early SMRT sequencing chemistry (P1-C1) generated ~700–1,500 base long-reads, which did not allow for complete *de novo* assembly or evaluation of structural variations.[13,14,31] Therefore, to develop a robust long-read sequencing and *de novo* assembly protocol to analyse the *P. falciparum* genome, we utilized the Pacific Biosciences RS II System with P6-C4 chemistry. Accordingly, we sequenced the genome of the strain 3D7 (Supplementary Fig. S1), and generated sequencing reads that had an average read length of 11–13 kb (maximum 45–50 kb), comprising over 5.26 Gb of data. The resulting sequences were assembled *de novo* into a highly accurate *P. falciparum* genome using the Hierarchical Genome Assembly Process (HGAP),[29] with all 14 chromosomes resolved into single contigs. Even extremely AT-rich regions, including the centromeres, were resolved with uniform coverage and for the first time, subtelomeric regions of all chromosomes were successfully assembled in a single run. We present an initial analysis of the *de novo*-assembled *P. falciparum* genome and discuss the advances that can now be made with regards to estimating *P. falciparum* genetic diversity using long-read sequencing technologies.

2. Materials and methods

2.1. Growth of *P. falciparum*

Blood stages of the *P. falciparum* laboratory strain 3D7 were grown according to Trager and Jensen[32] with a few changes. Briefly, a mixed stage culture of *P. falciparum* was grown in white blood cell (WBC)-free O+ human erythrocytes (prepared from whole blood by treatment with leucocyte-specific filters) at a haematocrit of 4% in Roswell Park Memorial Institute 1640 medium containing L-glutamine (Invitrogen) supplemented with 10% v/v Albumax II (Invitrogen) and 200 µm hypoxanthine (C.C.Pro). The cultures were grown in a gas environment of 5% CO_2, 1% O_2 and 94% N_2 to a parasitaemia of 3–8% before harvesting for downstream analysis. For synchronization, knob-positive parasites were selected by gelatin flotation using Plasmion (Fresenius Kabi)[33] and after re-invasion, treated twice with 5% sorbitol (Sigma)[34] to obtain parasites that were synchronized within a window of approximately 6 h, as evaluated by Giemsa staining.

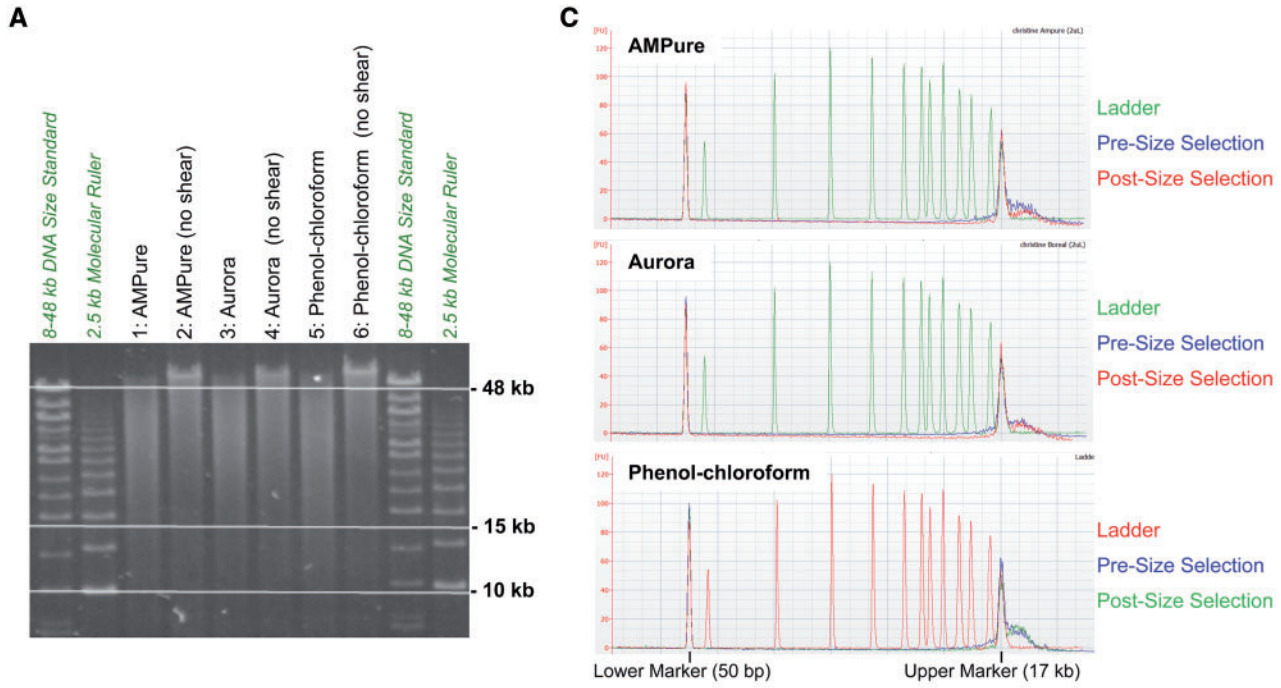
2.2. Genomic DNA isolation

Infected human erythrocytes at different parasitaemia were harvested and 1 ml or 5 ml aliquots were frozen at − 20 °C. Subsequently, genomic DNA was prepared using the DNeasy Blood and Tissue kit (Qiagen) or the Genomic Tip kit (Qiagen) according to manufac-

turer's instructions. For the DNeasy kit, free parasites that were obtained from the infected human erythrocyte pellet by saponin lysis[35] were resuspended in phosphate-buffered saline (PBS) and treated as per manufacturer's instructions. For the Genomic Tip kit, nuclei were prepared directly from the infected human erythrocyte pellet as per manufacturer's instructions.

2.3. DNA purification prior to library preparation

To remove heme and its derivatives from *P. falciparum* genomic DNA, the DNA was purified using one of the three independent methods: (i) magnetic bead-based cleanup; (ii) electrophoretic DNA extraction using the Aurora platform (Boreal Genomics) or (iii) phenol-chloroform extraction. The starting amount of DNA used for each method is indicated in Fig. 1B. For (i), AMPure PB magnetic beads (Pacific Biosciences) were mixed with *P. falciparum* genomic DNA at a 1:1 (vol:vol) ratio and incubated for 20 min at room temperature (RT) with gentle end-over-end rotation. Following this, the beads were washed twice with 1.5 ml 70% ethanol and allowed to dry briefly at RT before elution with 100 µl of Pacific Biosciences Elution Buffer (EB). For (ii), *P. falciparum* DNA was electrophoretically purified using the Aurora System (Boreal Genomics), following manufacturer's instructions in the 'Experienced User Protocol'. For (iii), *P. falciparum* genomic DNA was initially mixed with 500 µl buffer containing 1 M NaCl and 2 mM ethylenediaminetetraacetic acid (EDTA). Next, an equal volume of phenol:chloroform:isoamyl alcohol (24:23:1) was added to the DNA mixture, inverted to mix and spun in a microcentrifuge at 10,000 g for 10 min at RT. The upper aqueous phase was transferred to a new microcentrifuge tube, mixed with an equal volume of chloroform:isoamyl alcohol (24:1) and spun at 10,000 g for 10 min at RT. Then, to remove excess polysaccharide, 0.3 volumes of 99.99% ethanol was added to the upper aqueous phase, the mixture was inverted and spun at 10,000 g for 15 min at RT. Finally, DNA present in the upper aqueous phase was precipitated by adding 1.7 volumes of 99.99% ethanol and spinning at 10,000 g for 15 min at RT, followed by two 70% ethanol washes.



B

DNA Purification Strategy	Input (µg)	Purification Yield (µg)	% Recovery from Purification	Input for SMRTBell Library Prep (µg)	Total Library Yield (µg)	% Library Yield	Blue Pippin Size Selection of Library	Total Yield after Size Selection (ng)	% Recovery after Size Selection	Mean Insert Size (kb)
AMPure	11	6	54	3.9	1.9	48	20 kb	532	28	21.4
Aurora	11	4	36	3.9	1.7	43	20 kb	621	37	20.6
P-C*	100	33	33	4.1	1.9	46	20 kb	835	44	21.0

* P-C: Phenol-chloroform

Figure 1. Comparison of SMRTbell library preparation efficiency from *P. falciparum* genomic DNA purified using three different methods. (A) High molecular weight *P. falciparum* genomic DNA prepared from an asynchronous culture using the Genomic tip kit was purified by three different methods: (i) AMPure PB magnetic bead-based clean up (Lanes 1 & 2), (ii) electrophoretic DNA extraction using the Aurora System (Lanes 3 & 4) or (iii) phenol-chloroform extraction (Lanes 5 & 6), and sheared as described in Materials and Methods. Quality and size distribution of the sheared DNA (140 ng) was assessed using field-inversion gel electrophoresis (Pippin Pulse System, Sage Science). Size markers included CHEF 8-48 kb DNA Size Standard (Bio-Rad) and 2.5 kb Molecular Ruler (Bio-Rad). (B) SMRTbell libraries prepared using the indicated amount of purified genomic DNA was subjected to size selection on the BluePippin System using a 15 kb cut-off. The DNA yield and % recovery of various steps, library preparation efficiency and size-selection distribution (based on Fig. 2C) of the three DNA purification methods were compared. (C) Size, quantity and quality of SMRTbell libraries before and after size-selection were assessed using the Agilent DNA 12000 kit on the Agilent 2100 Bioanalyzer System.

The DNA pellet was allowed to air-dry at RT for up to 5 min and resuspended in 100 µl of Pacific Biosciences EB.

2.4. SMRTbell library preparation and sequencing

Three SMRTbell libraries were constructed for genomic DNA purified with each of the methods described above (Supplementary Fig. S1). Each library was constructed using ~4 µg of purified DNA and the SMRTbell Template Prep Kit 1.0, according to the protocol described in 'Procedure & Checklist—20 kb Template Preparation Using BluePippin™ Size-Selection System' (Pacific Biosciences). Briefly, *P. falciparum* DNA was sheared for 5 min at 3000g or 5,500 rpm using a g-TUBE (Covaris), concentrated with AMPure PB beads and subjected to DNA damage repair and ligation of SMRTbell adapters. Following ligation, extraneous DNA was digested with exonucleases and the SMRTbell library was cleaned and concentrated with AMPure PB beads. The libraries were then subjected to a 20 kb DNA size-selection step using the BluePippin System (SageScience) to remove shorter DNA inserts with a size cut-off of 15 kb. Library quality and quantity were assessed using the Agilent 12000 DNA Kit and 2100 Bioanalyzer System (Agilent Technologies), as well as the Qubit dsDNA Broad Range Assay kit and Qubit Fluorometer (Thermo Fisher). Sequencing primer and P6 polymerase were annealed and bound, respectively, to the SMRTbell libraries as recommended by the manufacturer (Pacific Biosciences). To identify the library concentration that would achieve optimal Poisson loading on the SMRT Cell (i.e. ~40% of zero mode waveguides loaded with a single DNA molecule), loading titrations were performed for each library. Based on this analysis, polymerase-bound SMRTbell libraries were loaded at a concentration of 200 pM for libraries cleaned up using magnetic beads and electrophoretic extraction, and at 100 pM for the phenol-chloroform extracted library to achieve comparable sequencing efficiencies. SMRT sequencing was performed on the Pacific Biosciences RS II System using the C4 sequencing kit (Pacific Biosciences), with magnetic bead loading and 240 min movies. Three SMRT cells were run per library type, providing a total of nine SMRT cells worth of data to be used for downstream analysis. Prior to data analysis, raw reads with a predicted polymerase read quality less than 0.80 were filtered out.

2.5. Data analysis

De novo assembly of the *P. falciparum* genome was carried out using the RS_HGAP_Assembly.3 protocol within Pacific Biosciences' SMRT Analysis Portal 2.3.0.p2 as previously described[29] (Supplementary Fig. S1). All parameters were set at their default values with the following exceptions: (i) the minimum subread length was set to 13,000 based on the size distribution of the reads (Fig. 2A) and for computational expediency, so as to not exceed 100-fold coverage going into the analysis; (ii) the minimum seed read length was increased to 20,000 from a default of 6,000—this was relative to the value of the minimum subread length; (iii) the genome size was set to 24,000,000 and (iv) the target coverage parameter was increased to 30 from a default of 25 to enhance coverage in the preassembly process (described below), in turn improving the quality of the finished assembly. The first step in RS_HGAP_Assembly.3, i.e. preassembly, utilizes a directed acyclic graph-based consensus procedure to align shorter reads to the longest reads in the sequencing data, thus generating corrected, continuous preassembled reads.[29] Next, the preassembled reads are assembled into larger contigs with the AssembleUnitig algorithm (Pacific Biosciences; https://github.com/PacificBiosciences/SMRT-Analysis/wiki/SMRT-Pipe-Reference-Guide), which incorporates elements of Celera Assembler[36] for overlapping and layout. Finally, the filtered reads are mapped back to the contigs using Blasr,[37] error-corrected with Quiver[29] to generate a high-quality consensus (referred to here as the SMRT assembly), and then visualized using SMRT View 2.3.0 (Pacific Biosciences).

To assemble the ~32 kb *P. falciparum* apicoplast genome, reads that did not align to contigs in the SMRT assembly (which comprised only the nuclear genome) were reanalysed in SMRT Analysis Portal 2.3.0.p2 resulting in an apicoplast contig. For the ~6 kb *P. falciparum* mitochondrial genome, Blasr was used to align the raw sequencing data to the M76611 mitochondrial sequence (http://plasmodb.org/plasmo/showRecord.do?name=SequenceRecordClasses.SequenceRecordClass&project_id=PlasmoDB&primary_key=M76611) and assembly performed with reads that presented partial or full homology to M76611 mitochondrial DNA.

Dot plots comparing the SMRT assembly to the *P. falciparum* reference genome, Pf3D7_v3.0 (http://genedb.org), were rendered with Gepard[38] using default parameters and a word length of 30. To identify structural variations in the genome, we used our new algorithm Assemblytics.[39] Briefly, Assemblytics analyses the whole genome alignment computed by MUMmer[40] and applies a unique length filtering approach to robustly identify structural variations in six classes of variants—insertions, deletions, tandem expansions, tandem contractions, repeat expansions and repeat contractions. Based on the size of the *P. falciparum* genome, the 'minimum size of variant' was set to 2 bp and the 'minimum unique sequence length to anchor an alignment' was set to 10 kb, which also becomes the maximum variant size to avoid calling variants above the size of the uniquely mapped sequence. Next, the presence of poly-A, poly-T and poly-AT in the smaller insertions (less than 10 bp) was determined by the following rules: (i) insertion must only contain the repeat sequence (for instance ATA for poly-AT) and (ii) either the 10 bp on the left or the 10 bp on the right must contain at least 6 bp of the repeat (for example, ATATAT in the case of the poly-AT repeat, AAAAAA in the case of poly-A). Finally, to determine whether a variant overlapped with a genomic feature, we performed a left outer join intersect using BEDTools[41] against Pf3D7_v3.0, annotation release 13.0 from plasmodb.org.

Direct alignment of the raw sequencing reads to Pf3D7_v3.0 was carried out using Blasr[37] with default settings. The resulting coverage obtained was 166-fold.

3. Results

3.1. High molecular weight genomic DNA was prepared from different *P. falciparum* intraerythrocytic stages

Critical to obtaining long-reads with Pacific Biosciences' SMRT sequencing technology is the extraction of high quality, high molecular weight genomic DNA, with a recommended size distribution of 50–150 kb (http://www.pacb.com/wp-content/uploads/2015/09/User-Bulletin-Guidelines-for-Preparing-20-kb-SMRTbell-Templates.pdf). To achieve this, we prepared genomic DNA from the *P. falciparum* laboratory strain 3D7 cultured in human blood using two different Qiagen kits: the DNeasy Blood and Tissue kit, which is routinely used by malaria researchers,[35] and the Genomic Tip kit. As shown in Supplementary Fig. S2A, the size distribution of genomic DNA prepared using the DNeasy kit was between 33.5 and 48.5 kb, in contrast to a size distribution of >50 kb for genomic DNA prepared with the Genomic Tip kit. We also determined the output of the

A

SMRT Cell	Purification strategy	On-plate conc. (pM)	Total Mb	Total Number of Reads	Mean Read Length	Read Length N50
1	AMPure	200	492	47,693	12,427	15,718
2	AMPure	200	669	63,846	12,865	16,337
3	AMPure	200	632	60,912	13,122	16,375
4	Aurora	200	495	46,069	12,253	16,041
5	Aurora	200	336	33,096	12,416	16,077
6	Aurora	200	552	52,603	12,773	16,226
7	P-C*	100	750	81,949	10,823	13,943
8	P-C*	100	678	71,458	11,265	14,080
9	P-C*	100	652	68,370	11,223	14,348
	Mean		***584***	***58,444***	***12,130***	***15,461***
	SD*		***127***	***15,004***	***822***	***1026***

* P-C: Phenol-chloroform
** SD: Standard Deviation

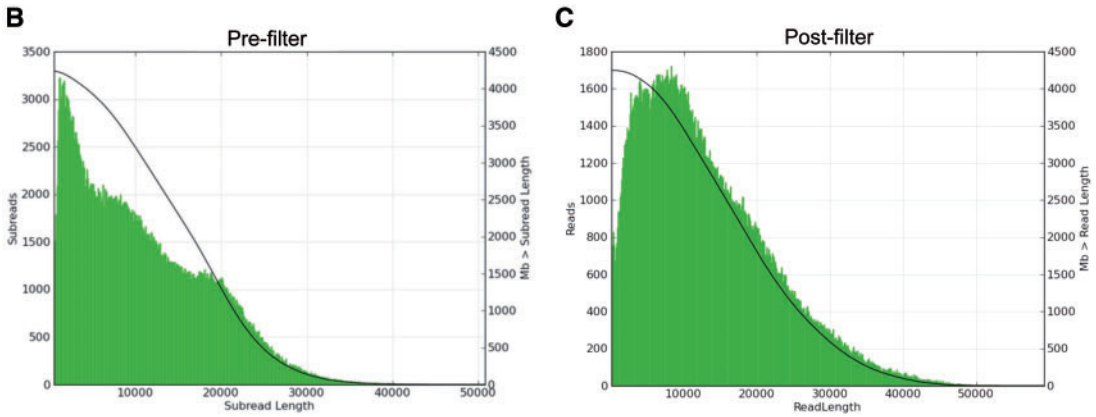


Figure 2. SMRT sequencing of *P. falciparum* genomic DNA yields >500,000 reads of average length 12.31 kb, with a read length N50 of 15.46 kb. (A) SMRTbell libraries were analysed on nine SMRT cells (three cells per DNA purification method) in a Pacific Biosciences RS II Sequencing System using 4 h-long movies. Sequencing metrics are shown for each SMRT cell. Sequencing data from nine SMRT cells were pooled and analysed for read length distribution (B) pre- and (C) post-filtering. The *x*-axis represents read length in bases, while the *y*-axis represents number of reads (gray columns) and megabases (Mb) greater than read length (black curve).

Genomic Tip kit for 3D7 parasites synchronized at ring (6–18 h post-invasion; non-replicating), trophozoite (28–35 h post-invasion, when DNA replication is initiated) and schizont stages (38–45 h post-invasion, when DNA replication is complete) and found that 3. 1×10^8 ring stage parasites yielded ~2.5 µg of genomic DNA, while >15 µg of genomic DNA could be extracted from 0.8×10^8 schizonts (Supplementary Fig. S2B). Extrapolating from these results, we conclude that >10 µg of high molecular weight genomic DNA can be extracted from a *P. falciparum* schizont culture growing in 500 µl of blood at a parasitaemia of ~2% with Qiagen's Genomic Tip kit, which is sufficient for downstream library preparation, size-selection and sequencing.

3.2. SMRT sequencing of *P. falciparum* genomic DNA yielded reads of average length ~12 kb

Genomic DNA prepared from blood stage *P. falciparum* parasites is routinely contaminated with heme and its derivatives. Because these contaminants adversely affect downstream analyses such as PCR and sequencing,[42–44] we examined the efficiency of three different purification methods to clean up *P. falciparum* genomic DNA isolated from an asynchronous 3D7 culture, prior to SMRTbell library preparation (Supplementary Fig. S1). These included: (i) magnetic bead-based cleanup, using a 1:1 ratio of AMPure PB beads, (ii) electrophoretic DNA extraction using the Aurora System from Boreal Genomics and (iii) phenol-chloroform extraction. We found that all three purification methods efficiently removed heme and other contaminants and did not affect the size distribution of the genomic DNA (Fig. 1A; 'no shear'). Notably, the magnetic bead-based purification method demonstrated the highest per cent DNA recovery values (54% versus 33 and 36% for phenol-chloroform extraction and Aurora System clean up, respectively) (Fig. 1B). Subsequently, we prepared sequencing libraries and observed that DNA purification method did not impact shearing (Fig. 1A; 'shear'), SMRTbell library yield (Fig. 1B) and BluePippin 20 kb size-selection (Figs 1B and C), and efficiently generated size-selected SMRTbell libraries with average insert lengths of ~21 kb (Figs 1B and C). Of note, the SMRTbell libraries prepared from phenol-chloroform extracted DNA showed the highest recovery (44%) after size-selection (Fig. 1B).

Thereafter, we used the Pacific Biosciences P6 polymerase, C4 sequencing chemistry and the RS II Sequencing System to analyse the size-selected *P. falciparum* SMRTbell libraries (Supplementary Fig. S1). For libraries derived from each purification method, we ran three SMRT cells for 4 h, resulting in 525,996 raw sequencing reads that totalled 5.26 Gb from nine pooled SMRT cells (Fig. 2A). SMRTbell libraries derived from phenol-chloroform extracted DNA

provided greater sequencing yields at lower loading requirements, although average read lengths were slightly shorter (not statistically significant) than libraries generated from other purification methods. On average, 58,444 reads were generated per SMRT cell, with a read length of 12.1 kb (Fig. 2A), and a tail in the read length distribution reaching close to 50 kb (Fig. 2B). Moreover, 50% of the sequenced bases originated from reads that were ≥15.5 kb long (Fig. 2A; read N50), suggesting that the Pacific Biosciences P6 polymerase may be capable of sequencing through long stretches of highly AT-rich genomic content, confirming prior reports of little to no sequence-context bias of SMRT sequencing.[13,14,25]

3.3. End-to-end *de novo* assembly of all 14 *P. falciparum* chromosomes

We next performed *de novo* assembly of filtered reads (4.26 Gb data comprising 325,565 reads of N50 18.168 kb; Fig. 2C) using the HGAP protocol in SMRT Portal 2.3.0.p2, with changes made to the default settings as described in Materials and Methods. The resulting assembly (hereafter referred to as the SMRT assembly) had a total genome size of 23.6 Mb (Fig. 3A) and produced a total of 21 contigs (Fig. 3A and Supplementary Fig. S3A–U; Supplementary Table S1). Thirteen of the 21 contigs were complete, individual *P. falciparum* chromosomes (chr. 2–14) ranging in length from 0.943 to 3.286 Mb (Fig. 3B), while the remaining eight contigs represented a 620 kb portion of chromosome 1 (contig 20) and 20–60 kb repeat regions that aligned to chromosome ends. One of these repeats corresponded to the left arm of chromosome 1 (contig 64; discussed in detail below), but the remaining six contigs could not be correctly assigned to a single chromosome end (Supplementary Table S1; Supplementary Figs. S3P–U). We consider these to be spurious contigs that may have arisen from high coverage of repetitive regions and do not have any discernible protein-coding or non-coding RNA (ncRNA) features.

Each of the 13 fully resolved chromosomes featured distinct edges at either end, where coverage abruptly stopped and beyond which reads did not extend (Fig. 3B and Supplementary Fig. S3A–M). These ends are highly homologous to each other, both within the same chromosome and between different chromosomes, and contain several copies of the telomeric repetitive element CCCTNAA, indicating complete telomere-to-telomere assembly of chr. 2–14. A representative coverage plot for contig 12, i.e. chr. 7, is shown (Fig. 3C). In the case of chr. 1, we observed that the 620 kb contig 20 began with a gradual increase in coverage (left end) and ended with a distinct edge containing telomeric repeats (right end) (Fig. 3B and Supplementary Fig. S3O). We, therefore, searched for a contig that could be contiguous with contig 20 and identified the 37 kb contig 64 (Fig. 3B and Supplementary Fig. S3N), which had a distinct telomeric repeat-containing edge at its left end and a gradual drop off in coverage at its right end. The manual joining of these two contigs yielded full-length chr. 1. Finally, because the genomes of the mitochondria [~6 kb; 68.4% (A + T) content] and apicoplast [~32 kb; 85.8% (A + T) content] were absent from the SMRT assembly, we selected reads that did not align to the nuclear chromosomes from the raw sequencing data and reanalysed them with independent HGAP calculations (detailed in Materials and Methods). In doing so, we completely assembled the genomes of these key organelles (Supplementary Figs S3W and S3X). Of note, the mitochondrial genome was assembled as a ~38 kb contig (Supplementary Fig. S3W), indicating that it exists as circular tandem duplications of itself: in this case, as six copies of the 6 kb genome.

We compared our SMRT assembly to other *de novo* assemblies of the *P. falciparum* genome that were generated using data from Roche 454 pyrosequencing,[45] Sanger shotgun sequencing[46] or Illumina-based sequencing by synthesis[47] of different *P. falciparum* strains (Table 1). We found that our assembly was the most complete: it comprised 23.6 Mb and contained the least number of contigs, with an average contig N50 of 1.71 Mb. Moreover, in comparison to previous studies that performed SMRT sequencing of *P. falciparum* genomic DNA,[13,14,31] we generated over 10-fold longer reads with an average genome coverage of × 94; indeed, we believe that our HGAP-derived SMRT assembly is the first of its kind for *P. falciparum*.

3.4. Analysis of centromeres and subtelomeric regions

Given the telomere-to-telomere nature of the SMRT assembly, we were specifically interested in the coverage of two challenging *P. falciparum* genomic regions: (i) centromeres, which have elevated AT-richness [between 90 and 99% (A + T) content][8,48,49] and (ii) subtelomeric regions, which contain six varieties of telomere-associated repeats, multigene families such as *var, rifin* and *stevor*,[8] and where much of the chromosomal length variation occurs.[50] We observed that in the case of all centromeres, depth of coverage was ×80–140 (Supplementary Table S1 and Supplementary Fig. S3—see the green line that indicates centromere position): for example, the region of chr. 14 that is marked by the centromeric histone PfCenH3, PF3D7_14:1070851-1075311,[48] was sequenced with a coverage of ~×140, as was the core ~2.5 kb centromere (Fig. 4A). This indicated that the highly AT-rich nature of centromeres, as well as the presence of atypical repeats such as AATTAA,[48] did not impede the processivity of the polymerase during sequencing. Similarly, we observed even coverage of most telomeric and subtelomeric ends (×80–140; Fig. 4B and Supplementary Fig. S3), suggesting that the >11 kb length of the SMRT reads is sufficient to differentiate between regions of very high homology.

3.5. Novel genomic features and structural variants were resolved in the SMRT assembly

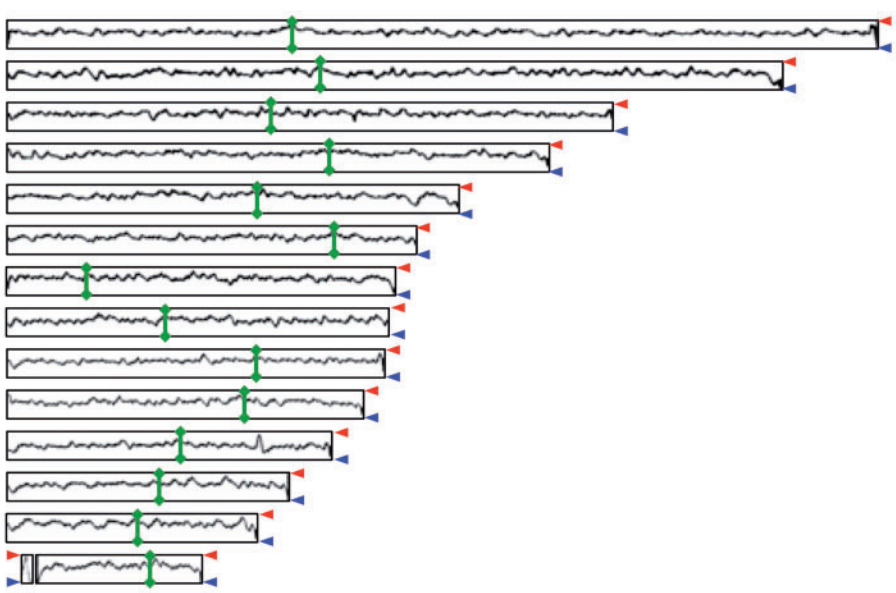
To determine if this new SMRT assembly could enhance our understanding of parasite genome organization, we compared our data to the reference 3D7 genome[8] and noted that the total size of the chromosomal contigs in the SMRT assembly, i.e. 23,294,534 bp, was comparable to the 23,292,622 bp resolved in the reference genome (Supplementary Table S1). Moreover, the sizes of most chromosomes were similar between the two assemblies with the maximum increase in size presented by the stitched SMRT chr. 1 at 3% and the maximum decrease in size presented by chr. 8 at 1.2% (Supplementary Table S1). When we visualized the comparison using dot plots, as shown in the representative dot plot for chr. 7 (Fig. 5A) and other chromosomes (Supplementary Fig. S4), we found that while the majority of assembled sequences aligned well with the reference genome, the increase in chr. 1 size in the SMRT assembly was due to the lengthening of its left arm (Supplementary Fig. S4A), and the decrease in chr. 8 size was due to a shortening of its right arm (Supplementary Fig. S4H). Upon closer analysis, we concluded that both of these size discrepancies originated from changes in the lengths of subtelomeric repeat sequences.

Furthermore, in select cases, the longer reads generated by SMRT sequencing allowed for the resolution of genomic features that were not previously described for the reference. These included large duplications and insertions (Figs 5B–D). For example, at position

A

	SMRT Cells	Assembly Size (Mb)	Total Number of contigs	Assembly size relative to reference (%)	Contig N50	Concordance with Sanger reference	QV
P. falciparum 3D7 (SMRT Assembly)	9	23.6	21	103%	1.71 Mb	~99%	50

B



Contig(s)	Size	Chromosome
0	3.3 Mb	14
1	2.9 Mb	13
2	2.3 Mb	12
3	2.0 Mb	11
6	1.7 Mb	10
11	1.54 Mb	9
8	1.43 Mb	8
12	1.42 Mb	7
13	1.41 Mb	6
14	1.36 Mb	5
15	1.21 Mb	4
18	1.08 Mb	3
19	0.95 Mb	2
64 + 20	35 kb + 0.62 Mb	1

C

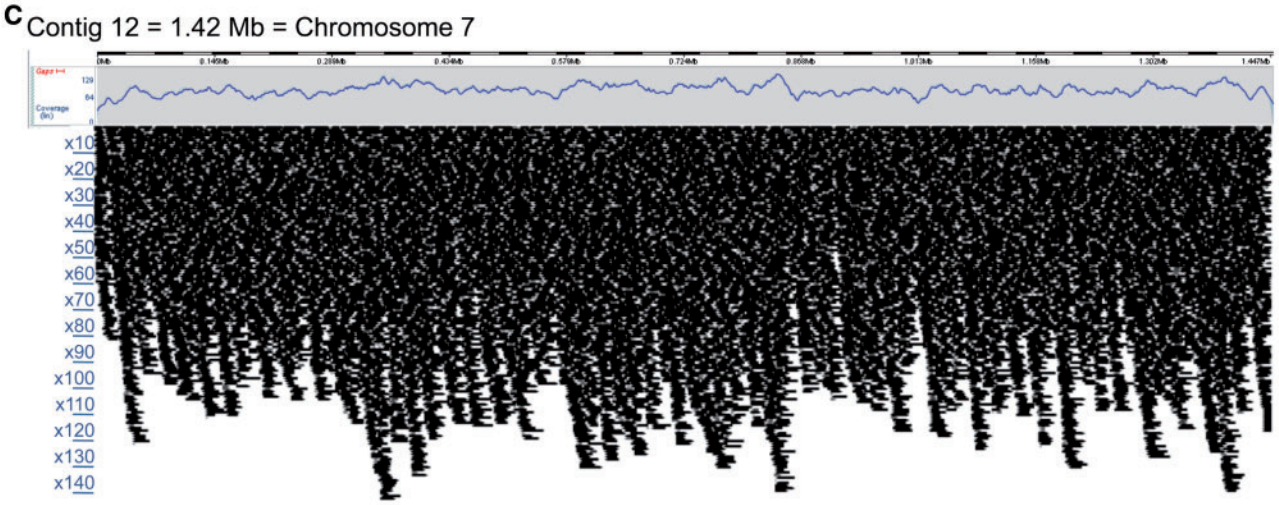


Figure 3. *De novo* assembly using HGAP resolves all 14 *P. falciparum* chromosomes. (A) Summary of the SMRT assembly metrics for the *P. falciparum* genome. Contig N50 is a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value, and QV50 indicates an estimated accuracy of 99.999% for the assembly. (B) All 14 *P. falciparum* chromosomes were assembled end-to-end using HGAP, without any gaps. The only exception was chr. 1, whose left arm was assembled as separate contig of 0.35 kb. The contig name in the SMRT assembly, its size, the corresponding chromosome and its coverage are indicated. The scale of coverage is between ×0 (blue arrowhead) and 150 (red arrowhead) except for contig 64, where the scale is from 0 to 100. The centromere position in each contig is indicated with the green line. (C) Contig 12 of the SMRT assembly, which aligns to *P. falciparum* chr. 7, was completely assembled using HGAP and showed a depth of coverage between ×70 and ×140.

1,612,060 of chromosome 10, a 14 kb duplication was apparent in the SMRT assembly relative to the reference (Fig. 5B), which results in the addition of three more *rifin* genes to the *P. falciparum* genome; PfRifins are virulence molecules that are expressed on the surface of infected erythrocytes, mediate the binding of infected erythrocytes to the vasculature[51] and undergo antigenic variation.[52] Another example is position 939,044 of chromosome 4, where a 29 kb insertion relative to the reference results in the duplication of the following features to the *P. falciparum* genome: three *var* genes, one ncRNA-encoding open reading frame (ORF), one *rif* and one gene encoding a conserved protein of unknown function (Fig. 5C). However, this region shows a substantial pileup of extra coverage, indicative of highly repetitive sequences, and may contain additional features that were not fully resolved in the SMRT assembly. One way to address this would be to analyse 40 kb size-selected SMRTbell libraries of *P. falciparum* DNA. Furthermore, adjacent to this insertion, we

Table 1. A comparison of *de novo P. falciparum* genome assemblies compiled using different NGS technologies

	Sanger sequencing[a]		Illumina sequencing[b]		454 pyrosequencing[c]		SMRT sequencing (this study)
Parasite strain	Dd2	HB3	NP-3D7-S	NP-3D7-L	7C126	SC05	3D7
Read length (bases)	600–700		36	76	3,000(paired-end)		12,130
Number of contigs	4,511	2,971	26,920	22,839	9,452	9,597	21
N50 contig size (kb)	11.6	20.6	1.5	1.6	3.3	3.3	1,710
Largest contig (kb)	79.2	111.9	29.1	24.0	36.7	34.4	3,290
Number of assembled bases (Mb)	19.5	23.4	19.0	21.1	20.8	21.1	23.6
Average coverage	×7.8	×7.1	×43	×64	×33	×36	×94

[a]Volkman et al., 2006.
[b]Kozarewa et al., 2009; NP: No PCR; S and L indicate short and long sequencing runs performed on the same library.
[c]Samarakoon et al., 2011.

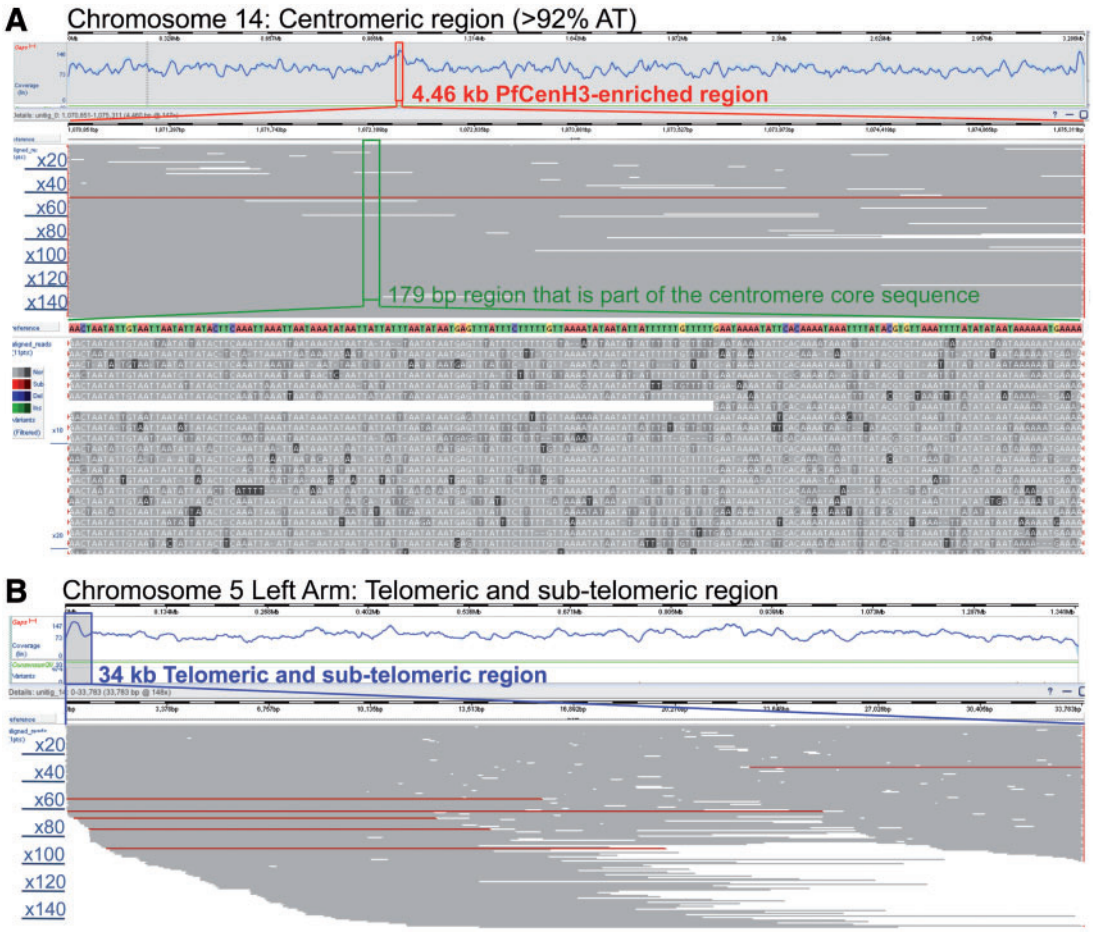


Figure 4. The depth of coverage of the SMRT assembly and length of SMRT reads are sufficient to resolve centromeres and subtelomeric regions. (A) The depth of coverage of the ~4.5 kb PfCenH3-occupied region of chromosome 14, which averages 92% (A + T), is shown. Zooming in, the raw sequences obtained for a 176 bp fragment of the core centromere are shown. (B) The depth of coverage of the ~34.6 kb telomeric/subtelomeric region of chromosome 5 is shown, where each grey line represents a single read. Note that the horizontal black lines represent reads that do not map uniquely to the assembly and have a mapping QV of zero.

observed a 1.5 kb stretch of DNA that was novel compared to the corresponding 2.5 kb stretch in the reference (Fig. 5C; 'zoom—novel') indicating that this subtelomeric region of chromosome 4 may be more complex than previously annotated. As a third example, we detected a 16 kb duplication at position 1,707,528 of chromosome 12 of the SMRT assembly (Fig. 5D), which results in the addition of 2 *var* genes and one ncRNA-encoding ORF to the *P. falciparum* annotation.

In addition to the features described above that span several kilobases (>10 kb), we utilized our assembly comparison tool called Assemblytics[39] to identify additional structural variants in our SMRT assembly relative to the reference genome. As summarized in

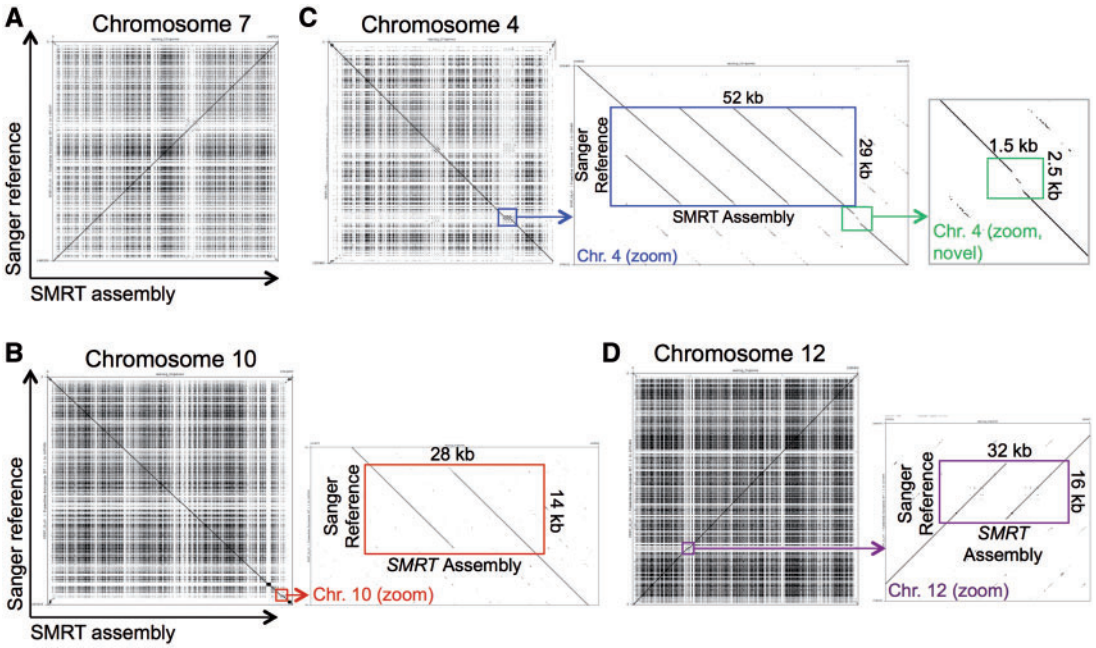


Figure 5. Dot plots comparing the SMRT assembly to the reference 3D7 genome identify large genomic variants. Dot plots were generated from Gepard nucleotide alignments of chromosomal contigs in the SMRT assembly (*x*-axis) and chromosomes assembled in the reference 3D7 genome (*y*-axis; Sanger reference[8]). Each dot is a grey-scale representation of nucleotide identity within a 30-nucleotide window centred on that position. The main diagonal line shows the alignment between the SMRT assembly and the reference genome. Off-diagonal lines parallel to the main diagonal indicate parallel duplications in the chromosome while off-diagonal perpendicular lines indicate inversions. (A) Chromosome 7. (B) Chromosome 10. Zooming into position 1,612,060, a 14 kb tandem duplication was resolved in the SMRT assembly relative to the reference. (C) Chromosome 4. Zooming into position 939,044, a 29 kb insertion was resolved in the SMRT assembly relative to the reference. Further zooming into a position immediately downstream of this insertion, a 1.5 kb stretch was detected in the SMRT assembly that showed very low homology to the corresponding 2.5 kb stretch present in the reference genome, and hence labelled 'novel'. (D) Chromosome 12. Zooming into position 1,707,528, a 16 kb tandem duplication was resolved in the SMRT assembly relative to the reference. Plots not drawn to scale.

Table 2. Size distribution of structural variants in the SMRT assembly relative to the reference genome

Size range (bp)	Variant type													
	Insertion		Deletion		Tandem expansion		Tandem contraction		Repeat expansion		Repeat contraction		Homo-polymer tract expansion[a]	
	Count	Total (bp)	Count	Total (bp)	Count	Total (bp)	Count	Total (bp)	Count	Total (bp)	Count	Total (bp)	Type	Count
2–10	10,112	26,499	335	1,148	0	0	0	0	1	8	0	0	poly-A	2,535
10–50	130	2,410	82	1,786	2	90	3	114	0	0	0	0	poly-T	2,517
50–100	1	66	6	401	2	142	2	123	1	77	1	97	poly-TA	1,108
100–1,000	4	647	3	507	15	5,446	3	865	4	2,123	1	808	poly-TG	3
1,000–10,000	1	1,019	0	0	3	8,294	0	0	2	7,526	1	1,018	poly-AG and poly-TC	1 each
Total	10,248	30,641	426	3,842	22	13,972	8	1,102	8	9,734	3	1,923	Total	6,164

[a]Length of expansion considered is between 2 and 10 bp.

Table 2 and Supplementary Table S2, we identified 10,248 insertions, 426 deletions, 22 tandem expansions, 8 tandem contractions, 8 repeat expansions and 3 repeat contractions of size 2 bp to 10 kb in the SMRT assembly; the size distribution of these structural variations is depicted in Figs 6A and B. Interestingly, homopolymer tract expansions, i.e. poly-A, poly-T and poly-AT expansions, of < 10 bp in size account for ~61% of all insertions identified (Table 2 and Supplementary Table S3). Some examples of variants include a 1,019 bp insertion at position 2,328,302 of chr. 13; 280 and 117 bp deletions at positions 1,452,348 of chr. 9 and 754,796 of chr. 13, respectively; a 4,131 bp tandem expansion at position 1,543,910 of chr. 10; a 378 bp tandem contraction at position 17,963 of chr. 4; a 5,100 bp repeat expansion at position 973,156 of chr. 12; and a 1,018 bp repeat contraction at position 965,633 of chr. 4 (Supplementary Table S2). These changes affect genic, intergenic and subtelomeric regions of chromosomes: for example, the 117 bp deletion at position 754,796 of chromosome 13 is within the intronless gene *PF3D7_1318300*, which encodes a conserved *Plasmodium*

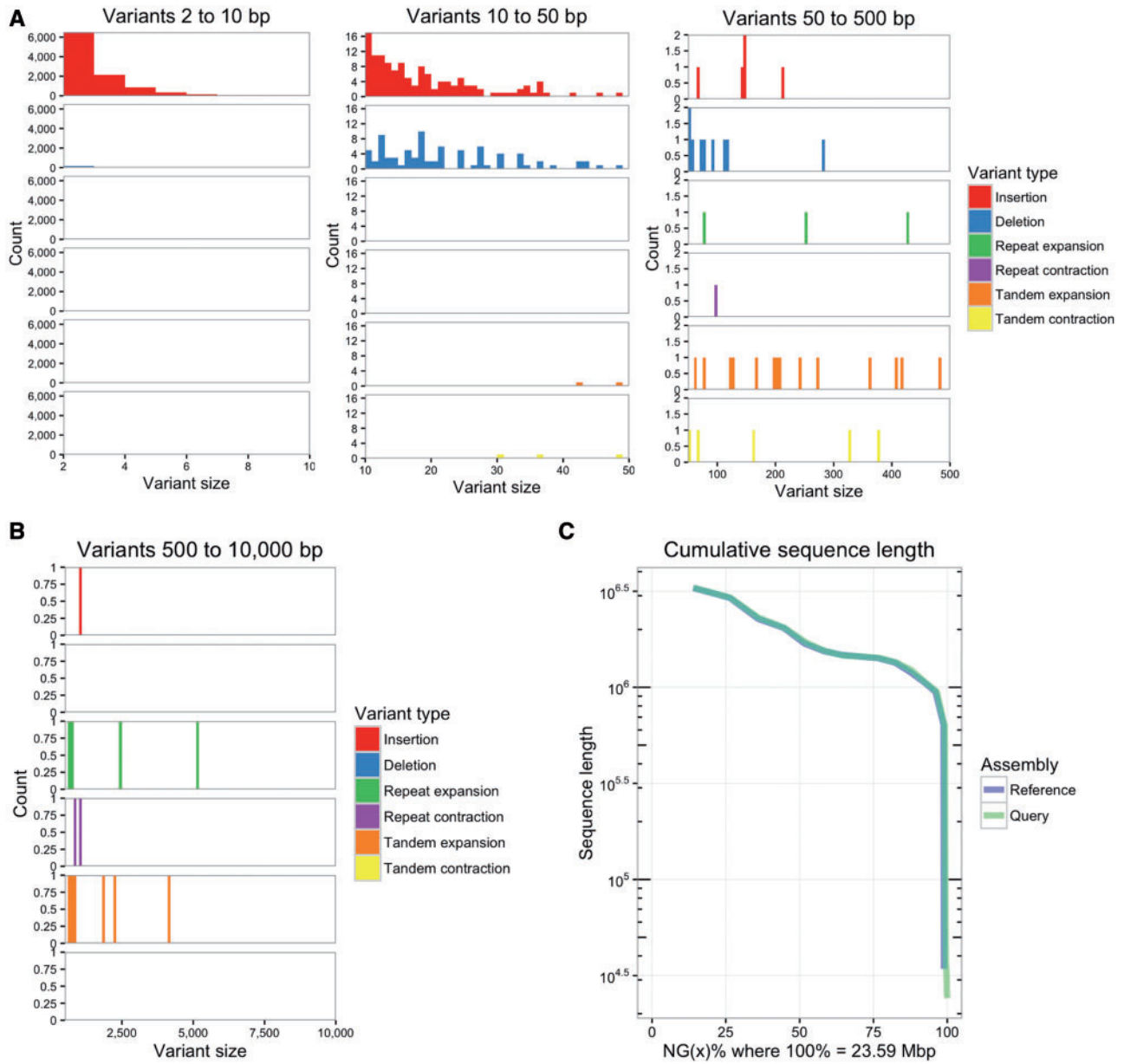


Figure 6. The majority of variations in the SMRT assembly relative to the reference are small insertions. Variants ranging from 2 to 10 kb in size were called using Assemblytics.[39] (A) Size distribution analysis of variants from 2 to 500 bp in size showed that the majority of insertions are <50 bp in size while the majority of deletions are between 10 and 500 bp in size. The *x*-axis represents variant size in bp and the *y*-axis represents a variant number. (B) Large structural variants from 500 bp to 10 kb in size are depicted with the *x*-axis representing variant size in bp and the *y*-axis representing variant number. (C) Cumulative sequence length plot showing the nearly identical contiguity and the total size of the SMRT assembly (query; in green) versus the reference (in blue). The length of each individual sequence is indicated on the *y*-axis with the cumulative sum of sorted sequence lengths on the *x*-axis. The N50 of the reference (50% on the *x*-axis) is 1.688 Mb, while the N50 of the SMRT assembly is 1.712 Mb.

protein of unknown function, while the 280 bp deletion in chr. 9 affects an intergenic region. Similarly, the 4,131 bp tandem expansion at position 1,543,910 of chr. 10 is within the intronless gene *PF3D7_1038400*, which encodes the gametocyte-specific protein Pf11-1, while the 378 bp tandem contraction in chr. 4 is in a subtelomeric region. Therefore, these variants will need to be rigorously curated to understand their impact on parasite biology. Given that the SMRT assembly and the reference match well in contiguity and have nearly the same total sequence length (Fig. 6C and Supplementary Table S1), and given its high-quality score (Fig. 3A; QV50), we can conclude that our SMRT assembly and the reference genome are of equal quality and that the structural variants identified here are significant to understanding the complexity of the parasite genome.

4. Discussion

In malaria-endemic areas, *P. falciparum* parasite genomes are constantly in flux.[16,53] However, our knowledge of *P. falciparum* genomic variability, which is built on PCR-based genotyping, microarray analyses and short-read NGS, is largely restricted to SNPs.[16] To characterize other genomic variants, particularly those that occur in

multigene virulence families and regulatory regions characterized by low sequence complexity such as subtelomeric repeats, there is an urgent need to apply new methods to analyse the *P. falciparum* genome. Therefore, we adapted the amplification-free, single molecule, long-read sequencing technology developed by Pacific Biosciences to analyse *P. falciparum* genomic DNA and compiled a new HGAP-derived 23.6 Mb genomic assembly that comprised 14 contiguous end-to-end chromosome sequences. In particular, we accurately resolved repetitive and polymorphic chromosome ends and identified large duplication and insertion events in subtelomeric regions and intra-chromosomal regions (adjacent to coding sequences) with high confidence, suggesting a greater complexity of the *P. falciparum* genome. In addition, we identified smaller insertions, deletions and tandem expansions in coding and non-coding regions of the genome, all of which can be used to correct and complete the *P. falciparum* reference genome[8]. It is noteworthy that these variants would have been missed if we had directly aligned the SMRT reads to the reference genome, thereby strongly supporting a hypothesis-free *de novo* assembly approach, not bound by reference bias.

In the past five yrs, the adaptation of NGS to study *P. falciparum* clinical isolates in Africa and South-East Asia has provided genome-wide estimates of allele frequency distribution, population structure and linkage disequilibrium in malaria-endemic regions.[18,20] In addition, NGS analysis of the genetic architecture of drug-resistant subpopulations of *P. falciparum* in Cambodia has identified SNPs that show high levels of differentiation, and other genomic correlates of the spread of resistance.[19] Complementarily, longitudinal studies that measured the genomic plasticity of laboratory-adapted *P. falciparum* parasites have determined the rates at which *var* gene polymorphisms are generated during intraerythrocytic development.[3,4] However, all of these studies used PCR-based WGA combined with short-read NGS and relied on the alignment-based mapping of sequencing reads to the reference 3D7 genome[8] to assess genetic diversity. While this method will prove invaluable to analysing SNPs in clinical settings where researchers mostly dispose of low sample volumes, we believe that the read length limitations of these approaches will make the identification of other critical structural variants extremely difficult. Given the lack of sequence bias in SMRT sequencing,[24] length of SMRT reads (>11 kb; compared to the ~8 kb size of a *var* ORF) and our complete resolution of chromosomal ends and copy number variants of virulence genes such as *var* and *rifin*, we argue that long-read sequencing will provide complete reference-grade genomes for *P. falciparum* field isolates. Such analyses, in combination with other approaches to document SNP variation, will comprehensively determine the geographical diversity of *P. falciparum* populations and monitor longitudinal population changes of both *P. falciparum* and other human malaria species.

We recognize that one limitation of our whole genome sequencing approach may be the amount of starting material (~10 μg) that is necessary to prepare high quality, size-selected SMRT sequencing libraries for *de novo* assembly. For example, if a malaria patient presents with a parasitaemia of >3% rings, around 4 ml of blood will be required to yield ~10 μg of *P. falciparum* genomic DNA, after eliminating human genomic DNA contamination.[54] In contrast, for cases with lower parasitaemia, ~200–500 μl of patient blood will have to be frozen and recultured with fresh WBC-free blood for one or two generations, or until the cultures reach a parasitaemia of 2% schizonts, before being processed for DNA extraction. Nonetheless, because several studies have shown that clinical isolates can be cultured *in vitro* from frozen stocks for short periods of time without losing their multiplication and erythrocyte invasion properties,[55,56] we believe that the latter option is feasible to compile SMRT sequencing-based whole genomes of *P. falciparum* clinical isolates. Furthermore, as sequencing technologies progress and develop, we anticipate that system requirements will change rapidly, consequently leading to a reduction in initial sample input, cost and time to project completion.

In conclusion, our study emphasizes the value of long-read sequencing technologies and *de novo* genome assembly to fully resolve pathogen genome architecture and complexity, paving the way for comprehensively assessing genetic variation at all size scales. In the specific case of *Plasmodium*, not only will it provide information about within-host *P. falciparum* diversity, but it will fill existing gaps in the genome sequences of laboratory-adapted strains as well as of field isolates. Furthermore, the resolution and understanding of *P. falciparum* structural variants, in particular of multigene families involved in adhesion and immune evasion (Ex: PfEMP1 and PfRifin), and genes involved in erythrocyte invasion (vaccine candidates such as PfAMA1 and PfRh proteins) and sexual stage biology (vaccine candidates such as P48/45, P25 and P28), will provide genetic correlates of parasite virulence and transmissibility. Given the recent assembly of the haploid human genome using SMRT sequencing,[28] we propose that long-read sequencing technologies will be crucial to combining *Plasmodium* genomic epidemiology with human population genetics to obtain a comprehensive view of parasite behaviour, genetic predisposition of humans to malaria and the co-evolution of parasite and host at a molecular level. Finally, SMRT technology can be used to identify epigenomic variations (Ex: genome-wide DNA methylation patterns) based on the kinetics of the DNA synthesis reaction performed by the polymerase, concurrent with sequencing.[57] Because this method is being utilized to examine both bacterial and eukaryotic epigenomes[58–60] and given the tractable size of the *P. falciparum* genome, future work could investigate if dynamic changes in DNA methylation[61] of *P. falciparum* clinical isolates are linked to varying degrees of severe malaria.

Acknowledgements

We are grateful to Cameron Macpherson and Sebastian Baumgarten for helpful scientific discussions and Jonas Korlach for critical reading of the manuscript. We also acknowledge PlasmoDB, a member of pathogen-databases that are housed under the NIAID-funded EuPathDB Bioinformatics Resource Center (BRC) umbrella, for regular updates and ready access to *P. falciparum* genome information.

Conflict of interest

C.L., M.S., P.B. and M.L.S. are full-time employees at Pacific Biosciences, a company that commercializes single molecule, real-time sequencing technologies.

Supplementary Data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by a European Research Council Advanced Grant (ERC AdG PlasmoEscape 250320) and the French Parasitology consortium ParaFrap (ANR-11-LABX0024) to A. S. S.S.V. was supported by the European Molecular Biology Organization Long-Term Fellowship, the Marie Sklodowska-Curie International Incoming Fellowship (FP7-MC-IIF-302451) and the Institut Pasteur Bourse Roux post-doctoral fellowship. Funding for open access charge is provided by the grant ERC 2014 AdG PlasmoSilencing awarded to A.S.

Availability of Supporting Data

The raw data generated in this study have been deposited in the National Centre for Biotechnology Information (NCBI) Sequence Read Archive under the accession number SRA360189 (http://www.ncbi.nlm.nih.gov/sra/?term=SRA360189; BioProject accession number = PRJNA313199 and BioSample accession number = SRS1315102 last accessed on June 8, 2016). The SMRT assembly has been deposited in the European Nucleotide Archive under the study accession number PRJEB11803 (http://www.ebi.ac.uk/ena/data/view/PRJEB11803 last accessed on June 8, 2016). Note that in the fastq file for the final assembly, the term Unitig is used instead of Contig.

References

1. Miller, L.H., Ackerman, H.C., Su, X.Z. and Wellems, T.E. 2013, Malaria biology and disease pathogenesis: insights for new treatments, *Nat. Med.*, **19**, 156–67.
2. Su, X., Hayton, K. and Wellems, T.E. 2007, Genetic linkage and association analyses for trait mapping in *Plasmodium falciparum*, *Nat. Rev. Genet.*, **8**, 497–506.
3. Bopp, S.E., Manary, M.J., Bright, A.T., et al. 2013, Mitotic evolution of *Plasmodium falciparum* shows a stable core genome but recombination in antigen families, *PLoS Genet.*, **9**, e1003293.
4. Claessens, A., Hamilton, W.L., Kekre, M., et al. 2014, Generation of antigenic diversity in *Plasmodium falciparum* by structured rearrangement of Var genes during mitosis, *PLoS Genet.*, **10**, e1004812.
5. Freitas-Junior, L.H., Bottius, E., Pirrit, L.A., et al. 2000, Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*, *Nature*, **407**, 1018–22.
6. Ariey, F., Witkowski, B., Amaratunga, C., et al. 2014, A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria, *Nature*, **505**, 50–5.
7. Rottmann, M., McNamara, C., Yeung, B.K., et al. 2010, Spiroindolones, a potent compound class for the treatment of malaria, *Science*, **329**, 1175–80.
8. Gardner, M.J., Hall, N., Fung, E., et al. 2002, Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, **419**, 498–511.
9. Su, X.Z., Heatwole, V.M., Wertheimer, S.P., et al. 1995, The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes, *Cell*, **82**, 89–100.
10. Oyola, S.O., Otto, T.D., Gu, Y., et al. 2012, Optimizing illumina next-generation sequencing library preparation for extremely AT-biased genomes, *BMC Genomics*, **13**, 1.
11. Oyola, S.O., Manske, M., Campino, S., et al. 2014, Optimized whole-genome amplification strategy for extremely AT-biased template, *DNA Res.*, **21**, 661–71.
12. Quail, M.A., Otto, T.D., Gu, Y., et al. 2012, Optimal enzymes for amplifying sequencing libraries, *Nat. Methods*, **9**, 10–11.
13. Quail, M.A., Smith, M., Coupland, P., et al. 2012, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, **13**, 341.
14. Ross, M.G., Russ, C., Costello, M., et al. 2013, Characterizing and measuring bias in sequence data, *Genome Biol.*, **14**, R51.
15. Carlton, J.M., Sullivan, S.A., Le Roch, K.G., Carlton, J.M., Perkins, S.L. and Deitsch, K.W. 2013, Plasmodium genomics and the art of sequencing malaria parasite genomes. In: J.M., Carlton, S.L., Perkins and K.W., Deitsch (eds.), *Malaria parasites: comparative genomics, evolution, and molecular biology*, Norfolk, UK: Caister Academic Press, pp. 35–58.
16. Kwiatkowski, D. 2015, Malaria genomics: tracking a diverse and evolving parasite population, *Int. Health*, **7**, 82–4.
17. Kamau, E., Campino, S., Amenga-Etego, L., et al. 2015, K13-propeller polymorphisms in *Plasmodium falciparum* parasites from sub-Saharan Africa, *J. Infect. Dis.*, **211**, 1352–5.
18. Manske, M., Miotto, O., Campino, S., et al. 2012, Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing, *Nature*, **487**, 375–9.
19. Miotto, O., Almagro-Garcia, J., Manske, M., et al. 2013, Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia, *Nat. Genet.*, **45**, 648–55.
20. Miotto, O., Amato, R., Ashley, E.A., et al. 2015, Genetic architecture of artemisinin-resistant *Plasmodium falciparum*, *Nat. Genet.*, **47**, 226–34.
21. Mobegi, V.A., Duffy, C.W., Amambua-Ngwa, A., et al. 2014, Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity, *Mol. Biol. Evol.*, **31**, 1490–9.
22. Nair, S., Nkhoma, S.C., Serre, D., et al. 2014, Single-cell genomics for dissection of complex malaria infections, *Genome Res.*, **24**, 1028–38.
23. Beerenwinkel, N., Günthard, H.F., Roth, V. and Metzner, K.J. 2012, Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data, *Front. Microbiol.*, **3**, 329.
24. Eid, J., Fehr, A., Gray, J., et al. 2009, Real-time DNA sequencing from single polymerase molecules, *Science*, **323**, 133–8.
25. Chaisson, M.J., Wilson, R.K. and Eichler, E.E. 2015, Genetic variation and the *de novo* assembly of human genomes, *Nat. Rev. Genet.*, **16**, 627–40.
26. Dilernia, D.A., Chien, J.T., Monaco, D.C., et al. 2015, Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing, *Nucleic Acids Res.*, **43**, e129.
27. English, A.C., Salerno, W.J., Hampton, O.A., et al. 2015, Assessing structural variation in a personal genome-towards a human reference diploid genome, *BMC Genomics*, **16**, 286.
28. Chaisson, M.J., Huddleston, J., Dennis, M.Y., et al. 2015, Resolving the complexity of the human genome using single-molecule sequencing, *Nature*, **517**, 608–11.
29. Chin, C.S., Alexander, D.H., Marks, P., et al. 2013, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat. Methods*, **10**, 563–9.
30. Koren, S., Schatz, M.C., Walenz, B.P., et al. 2012, Hybrid error correction and *de novo* assembly of single-molecule sequencing reads, *Nat. Biotechnol.*, **30**, 693–700.
31. Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C. and DePristo, M.A. 2012, Pacific biosciences sequencing technology for genotyping and variation discovery in human data, *BMC Genomics*, **13**, 375.
32. Trager, W. and Jensen, J.B. 1976, Human malaria parasites in continuous culture, *Science*, **193**, 673–5.
33. Lelièvre, J., Berry, A. and Benoit-Vical, F. 2005, An alternative method for Plasmodium culture synchronization, *Exp. Parasitol.*, **109**, 195–7.
34. Lambros, C. and Vanderberg, J.P. 1979, Synchronization of *Plasmodium falciparum* erythrocytic stages in culture, *J. Parasitol.*, **65**, 418–20.
35. Moll, K., Kaneko, A., Scherf, A., Wahlgren, M. 2013 Methods in Malaria Research, 6th edition.
36. Myers, E.W., Sutton, G.G., Delcher, A.L., et al. 2000, A whole-genome assembly of Drosophila, *Science*, **287**, 2196–204.
37. Chaisson, M.J. and Tesler, G. 2012, Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory, *BMC Bioinformatics*, **13**, 238.
38. Krumsiek, J., Arnold, R. and Rattei, T. 2007, Gepard: a rapid and sensitive tool for creating dotplots on genome scale, *Bioinformatics*, **23**, 1026–8.
39. Nattestad, M. and Schatz, M.C. 2016, Assemblytics: a web analytics tool for the detection of assembly-based variants, *bioRxiv*. doi: http://dx.doi.org/10.1101/044925.
40. Kurtz, S., Phillippy, A., Delcher, A.L., et al. 2004, Versatile and open software for comparing large genomes, *Genome Biol.*, **5**, R12.
41. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
42. Akane, A., Matsubara, K., Nakamura, H., Takahashi, S. and Kimura, K. 1994, Identification of the heme compound copurified with deoxyribonucleic acid (DNA) from bloodstains, a major inhibitor of polymerase chain reaction (PCR) amplification, *J. Forensic Sci.*, **39**, 362–72.
43. Kermekchiev, M.B., Kirilova, L.I., Vail, E.E. and Barnes, W.M. 2009, Mutants of Taq DNA polymerase resistant to PCR inhibitors allow DNA

amplification from whole blood and crude soil samples, *Nucleic Acids Res*., **37**, e40.
44. Zhang, Z., Kermekchiev, M.B. and Barnes, W.M. 2010, Direct DNA amplification from crude clinical samples using a PCR enhancer cocktail and novel mutants of Taq, *J. Mol. Diagn*., **12**, 152–61.
45. Samarakoon, U., Regier, A., Tan, A., et al. 2011, High-throughput 454 resequencing for allele discovery and recombination mapping in *Plasmodium falciparum*, *BMC Genomics*, **12**, 116.
46. Volkman, S.K., Sabeti, P.C., DeCaprio, D., et al. 2007, A genome-wide map of diversity in *Plasmodium falciparum*, *Nat. Genet*., **39**, 113–9.
47. Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M. and Turner, D.J. 2009, Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes, *Nat. Methods*, **6**, 291–5.
48. Hoeijmakers, W.A., Flueck, C., Francoijs, K.J., et al. 2012, *Plasmodium falciparum* centromeres display a unique epigenetic makeup and cluster prior to and during schizogony, *Cell. Microbiol*., **14**, 1391–401.
49. Kelly, J.M., McRobert, L. and Baker, D.A. 2006, Evidence on the chromosomal location of centromeric DNA in *Plasmodium falciparum* from etoposide-mediated topoisomerase-II cleavage, *Proc. Natl. Acad. Sci. U. S. A*., **103**, 6706–11.
50. Figueiredo, L.M., Freitas-Junior, L.H., Bottius, E., Olivo-Marin, J.C. and Scherf, A. 2002, A central role for *Plasmodium falciparum* subtelomeric regions in spatial positioning and telomere length regulation, *EMBO J*., **21**, 815–24.
51. Goel, S., Palmkvist, M., Moll, K., et al. 2015, RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria, *Nat. Med*., **21**, 314–7.
52. Kyes, S.A., Rowe, J.A., Kriek, N. and Newbold, C.I. 1999, Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*, *Proc. Natl. Acad. Sci. U. S. A*., **96**, 9333–8.
53. Ghansah, A., Amenga-Etego, L., Amambua-Ngwa, A., et al. 2014, Monitoring parasite diversity for malaria elimination in sub-Saharan Africa, *Science*, **345**, 1297–8.
54. Oyola, S.O., Gu, Y., Manske, M., et al. 2013, Efficient depletion of host DNA contamination in malaria clinical sequencing, *J. Clin. Microbiol*., **51**, 745–51.
55. Bei, A.K., Patel, S.D., Volkman, S.K., et al. 2014, An adjustable gas-mixing device to increase feasibility of *in vitro* culture of *Plasmodium falciparum* parasites in the field, *PLoS One*, **9**, e90928.
56. Ribacke, U., Moll, K., Albrecht, L., et al. 2013, Improved *in vitro* culture of *Plasmodium falciparum* permits establishment of clinical isolates with preserved multiplication, invasion and rosetting phenotypes, *PLoS One*, **8**, e69781.
57. Korlach, J. and Turner, S.W. 2012, Going beyond five bases in DNA sequencing, *Curr. Opin. Struct. Biol*., **22**, 251–61.
58. Davis, B.M., Chao, M.C. and Waldor, M.K. 2013, Entering the era of bacterial epigenomics with single molecule real time DNA sequencing, *Curr. Opin. Microbiol*., **16**, 192–8.
59. Greer, E.L., Blanco, M.A., Gu, L., et al. 2015, DNA methylation on N6-Adenine in *C. elegans*, *Cell*, **161**, 868–78.
60. Wu, T.P., Wang, T., Seetin, M.G., et al. 2016, DNA methylation on N(6)-adenine in mammalian embryonic stem cells, *Nature*, **532**, 329–333.
61. Ponts, N., Fu, L., Harris, E.Y., et al. 2013, Genome-wide mapping of DNA methylation in the human malaria parasite *Plasmodium falciparum*, *Cell Host Microbe*, **14**, 696–706.