Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes

Iwanka Kozarewa[1,2], Zemin Ning[1,2], Michael A Quail[1], Mandy J Sanders[1], Matthew Berriman[1] & Daniel J Turner[1]

Amplification artifacts introduced during library preparation for the Illumina Genome Analyzer increase the likelihood that an appreciable proportion of these sequences will be duplicates and cause an uneven distribution of read coverage across the targeted sequencing regions. As a consequence, these unfavorable features result in difficulties in genome assembly and variation analysis from the short reads, particularly when the sequences are from genomes with base compositions at the extremes of high or low G+C content. Here we present an amplification-free method of library preparation, in which the cluster amplification step, rather than the PCR, enriches for fully ligated template strands, reducing the incidence of duplicate sequences, improving read mapping and single nucleotide polymorphism calling and aiding *de novo* assembly. We illustrate this by generating and analyzing DNA sequences from extremely (G+C)-poor (*Plasmodium falciparum*), (G+C)-neutral (*Escherichia coli*) and (G+C)-rich (*Bordetella pertussis*) genomes.

Sequencing genomes with biased nucleotide compositions poses great technical challenges to the currently available sequencing platforms, most notably the highly (G+C)-poor genomes of *Plasmodium* species, which are difficult even for the traditional Sanger method[1–4]. In several malaria species, including *P. falciparum*, the mean exonic A+T content is >75%, and in intergenic and intronic regions it can be close to 100%[5,6].

One lane of an Illumina Genome Analyzer flowcell[7] can yield 700×10^6 bases of purity-filtered sequence data in a 36-base paired-end run, which is >30-fold coverage of the genome of the 23 megabase (Mb) reference *P. falciparum* clone 3D7 (ref. 6). To make the most of the sequencing capacity of the Genome Analyzer, it is essential to obtain as broad a representation of the genome as possible, but amplification and sampling biases during library preparation result in libraries that are lower in complexity than the genomic DNA from which they were derived. Additional sequencing runs of the same library are often not sufficient to improve coverage of regions that are poorly represented, and it becomes necessary to prepare and sequence additional libraries.

The standard Illumina library preparation pipeline is a multistep process that ends with PCR amplification before the sample is loaded into the flowcell. For the last 20 years, PCR has been used ubiquitously to amplify specific sections of DNA exponentially[8], but it is an inherently biased procedure[9–12]. To help overcome amplification biases and to reduce the formation of primer dimers, the Illumina library preparation protocol uses universal PCR primers, which allow simultaneous amplification of all loci in complex template pools[7]. There is a narrow range of conditions in the PCR that will give clean libraries with adequate representation[13], but even when performed under optimal conditions, the PCR is still sensitive to biases, particularly when the template to be amplified has a high A+T content such as *P. falciparum*.

The aim of the malaria sequencing program at the Wellcome Trust Sanger Institute is to sequence hundreds of cell lines, including clinical isolates. As a pilot study, we started with Genome Analyzer sequencing runs of *P. falciparum* 3D7, the reference genome sequenced by the Sanger dideoxy method[6], with the intention of correcting base errors in the reference. We followed this by several more sequencing runs for a variety of malaria strains. Quality of read mapping against the reference was very poor, with a high proportion of duplicate reads and uneven coverage. Such artifacts increase sequencing costs, as only a portion of the reads are useful. For the (G+C)-neutral *E. coli* and (G+C)-rich *B. pertussis* genomes, the coverage bias was far less pronounced.

Here we report an alternative method of Illumina library preparation that omits the PCR step entirely. For the extremely (G+C)-poor malaria genomes, datasets obtained from these libraries not only improved single-nucleotide polymorphism (SNP) detection, but also facilitated *de novo* assemblies using short read assemblers.

RESULTS

No-PCR library preparation

During Illumina library preparation, sample DNA is fragmented, end-repaired and A-tailed. Adapters, essentially consisting of the sequencing primer–annealing sequences, are then ligated via a

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [2]These authors contributed equally to this work. Correspondence should be addressed to D.J.T. (djt@sanger.ac.uk).

RECEIVED 10 OCTOBER 2008; ACCEPTED 17 FEBRUARY 2009; PUBLISHED ONLINE 15 MARCH 2009; DOI:10.1038/NMETH.1311

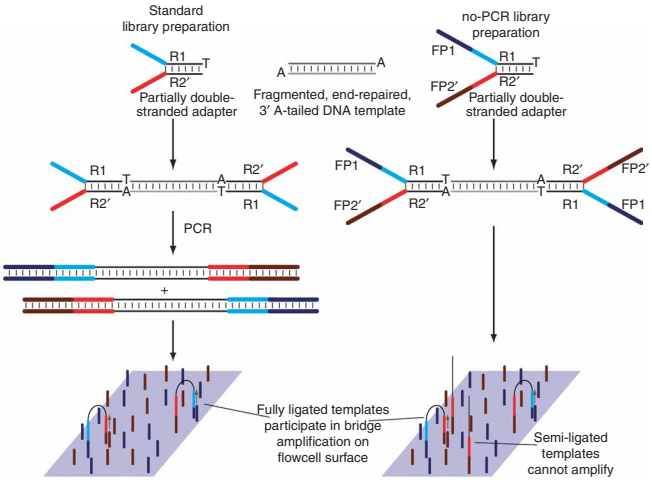

Figure 1 | No-PCR library preparation. In both standard and no-PCR library preparations, partially complementary adapters with a 3′ thymine (T) overhang are ligated onto fragmented, end-repaired, 3′ adenine (A)-tailed DNA. Whereas standard adapters consist only of sections to which read 1 and read 2 sequencing primers hybridize (R1 and R2′), no-PCR adapters also contain sequences that facilitate hybridization to oligonucleotides attached to the flowcell surface (FP1 and FP2′). The standard library preparation uses PCR to add these sections and to enrich for fully ligated templates, which then amplify on the flow cell surface. In the no-PCR approach, the flowcell itself is used to select for fully ligated template molecules. All no-PCR templates hybridize to the flowcell in the same orientation because only the FP2′ sequence is reverse-complementary to a flowcell oligonucleotide.

3′ thymine overhang. The structure of the adapters ensures that each strand receives a unique adaptor sequence at either end. Finally, ligated fragments are amplified by PCR[7]. Amplification is needed to generate sufficient quantities of template DNA to allow accurate quantification and to enrich for successfully ligated fragments. During the PCR additional adaptor sequence is added using tailed primers, resulting in template molecules capable of hybridizing to oligonucleotides on the flowcell surface. Even though the number of cycles of PCR amplification is kept low (10–12 cycles for paired-end libraries)[7], the PCR is still a source of duplicate sequences, amplification bias and struggles with (A+T)-rich base compositions[14]. Runs therefore become less efficient, and assembly, mapping and SNP detection are made more complicated than necessary.

In our no-PCR protocol, partially double-stranded adapters are also added to end-repaired template DNA with a 3′ adenine overhang, by ligation (**Fig. 1**). Unlike the standard Illumina adapters, no-PCR adapters contain additional sequences that allow hybridization of templates directly to the flowcell surface. Incompletely ligated fragments are inert in the cluster amplification step. Thus it is not necessary to retain the PCR step to enrich for properly ligated fragments, but to obtain an optimal cluster density, it is necessary to accurately quantify only those template fragments with an adaptor at either end. We achieve this by quantitative PCR, using primers that target the adaptor regions[13].

To investigate differences between standard and no-PCR library preparations, we produced four sets of 35- and 36-base paired-end *P. falciparum* 3D7 data from standard (STD) libraries STD-PF88, STD-PF2, STD-PF3 and STD-PF85, corresponding to read coverage of 174-, 114-, 96- and 21-fold, respectively (**Supplementary Table 1** online). We also produced two sets of paired-end 3D7 data from one no-PCR (NP) library: 36 bases (NP-3D7-S) and 76 bases (NP-3D7-L), corresponding to read coverage of 44- and 65-fold, respectively. We mapped reads to the 3D7 reference sequence using a modified version of SSAHA (sequence search and alignment by hashing algorithm)[15]. Data from the standard *P. falciparum* 3D7 libraries (STD-PF2, STD-PF3 and STD-PF85) and one run of a standard library made from a clinical isolate (STD-PF88) all failed to show a typical Poisson distribution with the peak around the average read depth. In contrast, the peak for both 36-base and 76-base no-PCR data agreed closely with read depth (**Fig. 2a**).

Sequence coverage and SNP calling

A plot of accumulated fractions of unmasked genome, in which Repeatmasker (Smith A.F.A., Hubley R. & Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org/) was not run to mask repetitive elements, against depth of base coverage (**Fig. 2b**) revealed that for STD-PF2 only 30% bases were covered by the mapped reads at tenfold or greater, whereas the 2.2 Gb of raw data should cover the 23 Mb genome 96 times on average. Between 4.8 and 19.9% of bases in the reference sequence were not represented in the standard Genome Analyzer sequence data (**Supplementary Table 1**). A highly uneven coverage distribution makes it difficult to identify duplicated regions. In such regions, variant nucleotides could be misinterpreted as SNPs when in reality they are paralogous sequence variants. Assuming that at least tenfold coverage is required to call SNPs reliably, the no-PCR data performed substantially better than the other four datasets, with 97% of bases covered 10 times or more (**Fig. 2b**). In contrast, for the (G+C)-neutral *E. coli* and (G+C)-rich *B. pertussis* genomes, this situation was less pronounced, with data generated from libraries produced using the standard protocol showing a Poisson-like distribution (**Fig. 2c**) and a high proportion of bases being covered by mapped reads at 30-fold or greater (**Fig. 2d** and **Supplementary Table 1**).

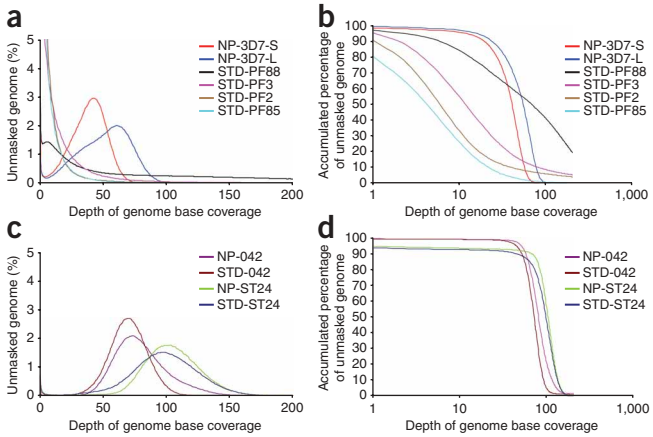

Figure 2 | Distribution of genome sequence coverage across the unmasked genomes for various datasets with or without the PCR step. (**a**,**b**) Percentage of unmasked genome versus depth of genome base coverage (**a**) and accumulated percentage of unmasked genome versus depth of genome base coverage (**b**) for standard (STD) and no-PCR (NP) library preparations of *P. falciparum* (PF) strains 2, 3, 88, 85 and 3D7 with either long (L) or short (S) reads. (**c**,**d**) Percentage of unmasked genome versus depth of genome base coverage (**c**) and accumulated percentage of unmasked genome against depth of genome base coverage (**d**) for standard (STD) and no-PCR (NP) library preparations of *E. coli* 042 and *B. pertussis* ST24.

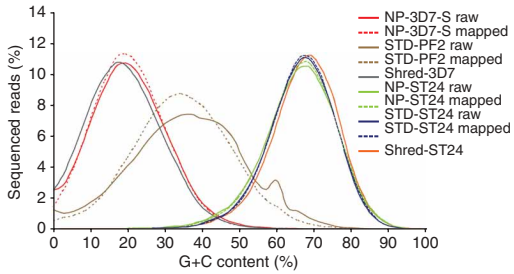

Figure 3 | Distribution of sequenced reads for different values of G+C content. G+C content profiles for raw and mapped sequence data for library preparations NP-3D7-S and STD-PF2 are shown alongside simulated data ('Shred-3D7') for comparison. G+C content was calculated in a window size of read length and therefore, the peak of fraction reads is dependent upon read length. A shift away from the simulated data curve, toward a more balanced G+C composition, is evident for the STD-PF2 sequence data, indicating severe bias. G+C content profiles for raw and mapped sequence data for library preparations STD-ST24 and NP-ST24 are shown alongside simulated data ('Shred-ST24'). The sequence data are not shifted away from the simulated data curve, indicating no bias in standard or no-PCR libraries for this genome.

We aligned no-PCR reads using SSAHA_pileup[15] and identified 2,059 SNPs. To estimate the accuracy of our SNP calls in the 3D7 data, we determined a baseline of accuracy from *E. coli* data, for which a high-quality complete sequence is available. Short-read Illumina Genome Analyzer sequence data, generated from the no-PCR library preparation, only differed at five positions in the entire 5.3 Mb genome. By *de novo* assembly we confirmed these differences to be finishing errors (data not shown). Assuming that read accuracy is similar between the *E. coli* and *P. falciparum* Illumina datasets, virtually all of the SNPs called in the *P. falciparum* dataset are actually base errors in our Sanger sequence data.

Amplification bias

To assess systematic biases in base composition introduced during the library preparation and sequencing procedures, it is necessary to evaluate how closely the sequence data represent the base composition of the original genome and to identify any major shifts in G+C content by comparison with a reference sequence. We divided reference sequences of *P. falciparum* 3D7, *E. coli* 042 and *B. pertussis* ST24 into tiled fragments corresponding to the read length used in the different sequencing runs. For each fragment, we calculated percent G+C content and used this information to plot theoretical G+C content profiles for these genomes, with which we compared the sequence data (**Fig. 3**). The coverage for this 'shredded' data for *P. falciparum* 3D7 and *B. pertussis* ST24 sequences were both 70-fold, though as read fraction is independent of read depth, we should not see any changes in G+C profiles at different depths.

For both the raw and mapped datasets of the *P. falciparum* 3D7 STD-PF2 library sequence data, there was an appreciable shift away from the theoretical shredded data, with the modal G+C content value of ~35% rather than the expected ~20%, indicating severe bias against A+T content in the sequences. Although both mapped and raw data were shifted in this way, the G+C profile of mapped STD-PF2 library data was smoother, suggesting the presence of some low-quality and (G+C) content–biased reads in the raw sequencing data could not be aligned against the reference. We observed a similar pattern for the other standard *P. falciparum* libraries (**Supplementary Fig. 1** online). In each case, a shift of the G+C profile toward higher G+C content indicated poor performance of these standard libraries. In contrast, both the raw and mapped datasets of the 36-base no-PCR library were in good agreement with the shredded data profile, showing that the base composition of the sequence data represents that of the original genome (**Fig. 3**). For *B. pertussis*, G+C content profiles of both standard and no-PCR libraries correlated well with the simulated data (**Fig. 3**). For *E. coli*, the standard library agreed closely with the simulated data, whereas the no-PCR library G+C content profile was shifted to the left, indicating a slight bias in this library (**Supplementary Fig. 1**).

Duplicate sequences

Duplicate sequences are a major concern in Illumina sequencing. We defined duplicate reads as those sharing exactly the same start and end locations, and counted these reads to determine the extent of duplication. In addition to PCR duplicates, duplicate sequences arise from adaptor dimers created during library preparation, sequencing artifacts such as poly(A) and undetermined sequences (poly(N) reads), noise in the cluster detection and analysis software (data not shown), and potentially from genomic DNA shearing at the same position in different molecules. Reducing PCR duplicates would be beneficial, both in lowering costs and allowing improved read mapping. We assessed the frequency of duplicate sequences in our no-PCR libraries by mapping the data to the reference sequence.

The frequency of duplicate sequences was high for STD-PF3 and STD-PF2 libraries (**Fig. 4a**): tails on the distribution curves for data from these standard libraries extended far, indicating that for such an (A+T)-rich genome, PCR duplicates are the major source of duplicate sequences. Notably, the duplication rate appeared to be high for the STD-PF88 library for a *P. falciparum* clinical isolate, but was, in fact, normal judged from the distribution of read duplication. The duplication distribution curve had a relatively

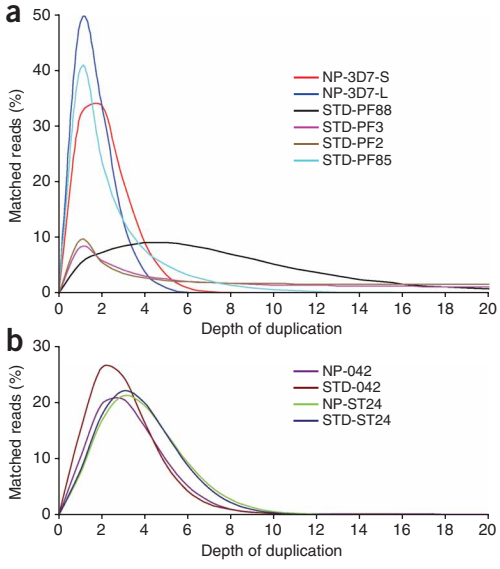

Figure 4 | Frequencies of duplicate sequences. (**a**,**b**) Duplication frequencies for *Plasmodium* libraries (**a**) and for *E. coli* and *B. pertussis* libraries (**b**) prepared both with and without a PCR step.

Table 1 | Summary of sequence data for the no-PCR and standard libraries

Library[a]	Organism	Genome size (Mb)	Insert size (bp)	Read length (bp)	Number of reads	Fold raw read coverage	Number of assembled bases	Contig coverage (%)	Number of contigs >100 bp	Contig N50
NP-3D7-S	*P. falciparum* 3D7	23	200	36	28,009,122	43	19,025,823	82.7	26,920	1,456
NP-3D7-L	*P. falciparum* 3D7	23	200	76	19,556,224	64	21,092,855	91.7	22,839	1,621
STD-PF88	*P. falciparum* 3D7	23	200	37	110,939,984	174	NA[b]	NA	NA	NA
STD-PF3	*P. falciparum* 3D7	23	200	37	75,083,768	114	NA[b]	NA	NA	NA
STD-PF2	*P. falciparum* 3D7	23	200	37	62,802,164	96	NA[b]	NA	NA	NA
STD-PF85	*P. falciparum* 3D7	23	200	37	13,530,194	21	NA[b]	NA	NA	NA
NP-042	*E. coli* 042	5.3	200	36	14,110,696	95	5,362,633	99.9	186	91,605
STD-042	*E. coli* 042	5.3	200	37	10,719,672	75	5,309,673	99.9	177	95,860
NP-ST24	*B. pertussis* ST24	4.0[c]	200	36	12,549,138	113	3,821,094	95.5	306	17,808
STD-ST24	*B. pertussis* ST24	4.0[c]	200	37	11,756,654	109	3,763,213	94.0	386	14,200

[a]No-PCR libraries have the prefix NP, and standard libraries have the prefix STD. Suffixes L and S indicate long and short sequencing runs performed on the same library. [b]No assembly was possible on data generated from the standard *P. falciparum* libraries. NA, not applicable. [c]Approximate size of the *B. pertussis* ST24 genome; in the absence of a finished assembly, it is only possible to estimate this.

short tail and a peak value at ~5.0-fold, which is close to the theoretical value of 3.4-fold, obtained by dividing mean coverage by read length. In contrast, the abundance of duplicate sequences in the no-PCR and standard libraries of *E. coli* and *B. pertussis* did not differ appreciably, suggesting that a greater proportion of these two genomes can amplify in the PCR (**Fig. 4b** and **Supplementary Table 1**). The duplication rate was low for the 36-base NP-3D7-S dataset, but not the 76-base NP-3D7-L dataset, even though we used the same library for both sequencing runs. Trimming the 76-base sequence data back to 36 bases revealed that although the base composition of mapped sequences agreed well with the theoretical data, the raw data had a tail shifting away from the theoretical predictions (**Supplementary Fig. 2** online), showing that the data from the 76-base run had greater bias than that from the 36-base run, indicating a problem with the longer sequencing run itself.

De novo assembly

The low bias of the no-PCR *P. falciparum* datasets makes *de novo* assembly possible, whereas standard libraries do not permit this, owing to uneven coverage and inadequate representation of the genome (**Table 1**). From the 36-base dataset with approximately 14 million paired end reads, we obtained an assembly of 15.5 Mb with N50 = 1.38 kilobases (kb) (that is, 50% of all bases are contained within contigs of 1.38 kb or longer). Using the 76-base data, we produced an assembly of 20.8 Mb with N50 = 1.28 kb from 9.8 million paired-end reads.

The standard *B. pertussis* library yielded an assembly with N50 = 10.6 kb from 6 million 36-base paired-end reads, whereas we obtained an N50 of 20.5 kb with the no-PCR library, also from 6 million paired-end reads (**Table 1**). No finished quality reference sequence is available from this organism; the genome has a very high mean G+C content (68%) which, coupled with a complicated repeat structure, makes assembly more difficult than for (G+C)-neutral genomes such as *E. coli*.

For *E. coli* strain 042, a 5.35 Mb genome with 50.5% G+C content, a finished sequence obtained by Sanger sequencing is available for comparison (http://www.sanger.ac.uk/Projects/Escherichia_Shigella/). Using 7 million 36-base paired-end reads from a standard library, we assembled the genome into contigs with N50 = 146 kb, compared to N50 of 71.7 kb with reads from a no-PCR library. The poorer assembly of the no-PCR library of the *E. coli* compared to the standard library is presumably due to variation in read quality, combined with a very limited effect of the no-PCR library preparation on this (G+C)-neutral genome.

DISCUSSION

By ligating adapters that consist of all sections required for sequencing primer annealing and attachment to the flowcell surface, we can avoid the requirement of a PCR step in the preparation of a sequencing library. The quantity of template DNA generated in this way is lower than when PCR is used, but library quantification by quantitative PCR[13] showed that from 5 μg of starting DNA, sufficient amount of 200 base pair (bp) no-PCR library can be obtained for >400 high-density Genome Analyzer lanes, more than enough for most sequencing purposes. Starting with lower quantities of DNA, for example, 500 ng of genomic DNA, we can obtain sufficient amount of library with a 200-bp insert size for about 12 lanes on a Genome Analyzer. Inserts of 500 bp resulted in a lower yield than shorter fragment libraries, presumably because of fewer fragments present in the same mass of DNA.

As with standard Illumina adapters, the structure of no-PCR adapters ensures that all fully ligated template strands receive the unique adaptor sequence complementary to the flow-cell adapters at their 5′ and 3′ end (**Fig. 1**). Because the efficiency of ligation is not 100%, many template strands will receive no adapters or will only be partially ligated. However, Illumina cluster amplification can only amplify template strands that have a different adaptor at either end and thus the cluster amplification step performs the enrichment that is otherwise provided in the PCR.

We demonstrated that for genomes of extreme G+C composition, the sequence coverage provided by the no-PCR approach is more even than the standard, PCR-based Illumina library preparation, contains very few duplicates, aids mapping and SNP calling, and makes assembly more straightforward. This is best illustrated by the *P. falciparum* genome, which until now has resisted attempts at *de novo* assembly from short-read data. The differences between the short- and long-read malaria assemblies are not large as the average fragment size for the no-PCR *P. falciparum* 3D7 library is only 170 bp, close to the long paired-read length of 152 bp (2 × 76 bases).

P. falciparum genomes are extremely difficult to assemble even using 600–700 base Sanger sequence reads: assembly of clinical isolates from sixfold Sanger sequencing coverage, yielded a contig

N50 of only 7 kb (data not shown). Although it seems unlikely that assemblies from short-read data alone will ever generate N50 values in the 7 kb range, we believe that we will be able to increase our malaria strain N50 beyond this by combining short-read data with Sanger reads.

Approximately 2% of the *P. falciparum* 3D7 reference sequence is not covered by the NP-3D7-S sequence data as we placed reads only to their best location and did not place repetitive reads. In contrast, 4.8–19.9% of bases were not covered by mapping for the standard *P. falciparum* libraries. Using an alternative alignment tool, MAQ[16], which places repetitive reads to a random location, the uncovered regions were reduced to just 5,585 bases for NP-3D7-S, indicating that 99.98% of the *P. falciparum* 3D7 genome is represented in the sequence data.

Anecdotally, sequences with a G+C content exceeding 80% are difficult to sequence on a Genome Analyzer. The genome of *B. pertussis* has a mean G+C content of 68%, and only a small proportion of sequence reads would have >80% G+C content. Nevertheless, both standard and no-PCR *B. pertussis* libraries revealed G+C content profiles that were almost identical to simulated data, with no decrease in representation as a function of greater G+C content (**Fig. 3**), indicating that the standard library preparation protocol finds no difficulty with G+C content within this range. If there are difficulties in sequencing organisms with a higher G+C content than *B. pertussis* on a Genome Analyzer, our data indicate that these are not the result of PCR artifacts, though it is conceivable that biases are introduced at other stages in the sequencing process, such as cluster growth[7]. However, the high G+C content of *B. pertussis* ST24 has hindered the generation of a finished standard reference sequence by Sanger sequencing: the assembly still contains 115 contigs, of which some are vector contamination, and this prevents a more thorough analysis.

Because of the absence of the PCR step, our method is quicker to perform than the standard Illumina library preparation[7], and we believe that it should be used routinely to prepare libraries for Genome Analyzer sequencing.

METHODS

Adaptor preparation. We obtained two oligonucleotides purified by high-performance liquid chromatography (Sigma): A_adapter_t and A_adapter_b. We phosphorylated 40 μM oligos at the 5′ end by 1 unit μl^{-1} of T4 polynucleotide kinase in 1× T4 ligase buffer (New England Biolabs) for 30 min at 37 °C in a thermocycler (MJ Research). We then denatured the kinase by heating and annealed the oligos by cooling to 20 °C by 0.1 °C every 2 s. We divided adaptor oligos into single-use aliquots and stored them at −20 °C.

Additional methods. Descriptions of DNA preparation, adapter ligation, adapter sequences, quantification and sequencing, standard library preparation, read alignment, SNP calling and *de novo* assembly are available in **Supplementary Methods** online.

URLs. Alignment software is available at http://www.sanger.ac.uk/software/analysis/SSAHA2/. All computer codes on the detection of SNPs and short insertion-deletions are available at ftp://ftp.sanger.ac.uk/pub/zn1/ssaha_pileup/. All raw Illumina reads and assemblies are available at ftp://ftp.sanger.ac.uk/pub/zn1/PCR_free/. Additional information regarding assemblies is available at ftp://ftp.sanger.ac.uk/pub/zn1/PCR_free/README.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust (grant WT079643). We thank C. Newbold and S. Kyes (University of Oxford) for providing DNA from *P. falciparum* 3D7.

AUTHOR CONTRIBUTIONS

I.K. planned and performed experiments; Z.N., M.J.S. and M.B. analyzed data; M.A.Q. prepared standard sequencing libraries; I.K. and D.J.T. devised the project; D.J.T., Z.N. and I.K. wrote the manuscript.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Goman, M. *et al.* The establishment of genomic DNA libraries for the human malaria parasite *Plasmodium falciparum* and identification of individual clones by hybridisation. *Mol. Biochem. Parasitol.* **5**, 391–400 (1982).
2. Camargo, A.A., Fischer, K., Lanzer, M. & del Portillo, H.A. Construction and characterization of a *Plasmodium vivax* genomic library in yeast artificial chromosomes. *Genomics* **42**, 467–473 (1997).
3. de Bruin, D., Lanzer, M. & Ravetch, J.V. Characterization of yeast artificial chromosomes from *Plasmodium falciparum*: construction of a stable, representative library and cloning of telomeric DNA fragments. *Genomics* **14**, 332–339 (1992).
4. Triglia, T. & Kemp, D.J. Large fragments of *Plasmodium falciparum* DNA can be stable when cloned in yeast artificial chromosomes. *Mol. Biochem. Parasitol.* **44**, 207–211 (1991).
5. Pollack, Y., Katzen, A.L., Spira, D.T. & Golenser, J. *The genome of Plasmodium falciparum*. I: DNA base composition. *Nucleic Acids Res.* **10**, 539–546 (1982).
6. Gardner, M.J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
7. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
8. Saiki, R.K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).
9. Day, D.J. *et al.* Identification of non-amplifying CYP21 genes when using PCR-based diagnosis of 21-hydroxylase deficiency in congenital adrenal hyperplasia (CAH) affected pedigrees. *Hum. Mol. Genet.* **5**, 2039–2048 (1996).
10. Barnard, R., Futo, V., Pecheniuk, N., Slattery, M. & Walsh, T. PCR bias toward the wild-type k-ras and p53 sequences: implications for PCR detection of mutations and cancer diagnosis. *Biotechniques* **25**, 684–691 (1998).
11. Hahn, S., Garvin, A.M., Di Naro, E. & Holzgreve, W. Allele drop-out can occur in alleles differing by a single nucleotide and is not alleviated by preamplification or minor template increments. *Genet. Test.* **2**, 351–355 (1998).
12. Ogino, S. & Wilson, R.B. Quantification of PCR bias caused by a single nucleotide polymorphism in SMN gene dosage analysis. *J. Mol. Diagn.* **4**, 185–190 (2002).
13. Quail, M.A. *et al.* A large genome centre's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).
14. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
15. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
16. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).