Hackflex: low cost Illumina sequencing library construction for high sample counts

Daniela Gaio[1], Joyce To[1], Michael Liu[1], Leigh Monahan[1], Kay Anantanawat[1], Aaron E. Darling[1*]

[1]The ithree institute, University of Technology Sydney, Sydney, NSW, Australia

*Corresponding author: Aaron E. Darling, aaron.darling@uts.edu.au

Keywords: library preparation; multi-plex; Illumina sequencing; metagenomics; Nextera Flex; high-throughput sequencing

ABSTRACT

We developed Hackflex, a low-cost method for the production of Illumina-compatible sequencing libraries that allows up to 11 times more libraries for high-throughput Illumina sequencing to be generated at a fixed cost. We call this new method Hackflex. Quality of library preparation was tested by constructing libraries from *E. coli* MG1655 genomic DNA using either Hackflex, standard Nextera Flex or a variation of standard Nextera Flex in which the bead-linked transposase is diluted prior to use. We demonstrated that Hackflex can produce high quality libraries and yields a highly uniform coverage, equivalent to the standard Nextera Flex kit. Using Hackflex, we were able to achieve a per sample reagent cost of library prep of A$8.66, which is 8.23 times lower than the Standard Nextera Flex protocol at advertised retail price. An additional simple modification to the protocol enables a further price reduction of up to 11 fold or about A$6.50/sample. This method will allow researchers to construct more libraries within a given budget, thereby yielding more data and facilitating research programs where sequencing large numbers of libraries is beneficial.

INTRODUCTION

The original Nextera protocol provided an easy to use and flexible means for generating Illumina-compatible shotgun libraries. When applied at scale, however, the Nextera reagents at list price could become prohibitively expensive for projects with large sample counts and low sequencing requirements per-sample. Previous work demonstrated that it was possible to dilute the Nextera reagents and with custom buffers, the per-sample library cost could be greatly reduced [1,2], thereby facilitating the processing of large sample batches. In 2017 Illumina introduced a new type of Nextera kit, called Nextera Flex, and subsequently discontinued the original Nextera kits for which dilution strategies had been developed. The Nextera Flex kits use bead-linked transposases to fragment and tag DNA with adapter sequences. The tagmentation technique allows the incorporation of defined adapter sequences, enabling barcoded primers to anneal and be extended through tagmented DNA fragments in subsequent PCR amplification and sequencing reactions [3]. The new Nextera Flex kits have been shown to yield greatly improved data quality relative to the original Nextera and Nextera XT kits [4]. Unfortunately, due to several chemistry differences, the existing dilution-based protocols can not be directly applied with the new Nextera Flex kits.

In this work, we introduce an ultra low cost variant of the Nextera Flex protocol that we call “Hackflex”. In addition to diluting the bead-linked transposases, our protocol replaces all other reagents with components readily available from third party sources, driving the cost per sample to A$8.66 (**Fig. 1**; **S1**). In this study we present our Hackflex protocol and compare the quality of the resulting data to the standard Nextera Flex kit protocol in terms of uniformity of coverage, sequence accuracy, GC coverage bias, and uniformity of barcode counts.

METHODS

Genomic DNA preparation

Genomic DNA of three different bacteria were used in this study: *Escherichia coli* strain MG1655, *Staphyloccus aureus* strain ATCC25923 and *Psudomonas aeruginosa* strain PAO1. For E. coli MG1655 strain, the reference genome used in this study differs from the original *E. coli* MG1655 strain sequenced by Blattner *et al* [4], most notably as it contains a pBAD plasmid. For this reason, we generated an independent reference assembly of our strain using both Illumina and Oxford Nanopore sequencing data (see below). For DNA extraction, high molecular weight gDNA was extracted from freshly cultivated cells of this strain using the Qiagen DNeasy UltraClean Microbial Kit according to the manufacturer’s instructions. Briefly, twenty milliliters of overnight culture was centrifuged at 3200 g for 5 min to obtain a cell pellet. The pellet was then washed with 5 mL sterile 0.9% sodium chloride solution, and then resuspended in 300 µL PowerBead solution before continuing with the kit ‘s manufacturer protocol. The final gDNA was eluted with 50 µL elution buffer pre-warmed to 42°C. The concentration of isolated DNA was measured using Qubit 2.0 (Thermo Fisher Scientific, USA) and diluted in water.

Barcode design

Sample index (barcode) design using a previously introduced method [5] yielded a set of 96 x 8 bp sequences (**S2**). Each i5 barcode was the reverse complement sequence of the corresponding i7 barcode (tandem complement design). Barcode sequences were designed such that no barcode contained 3 or more identical bases in a row, and the mean GC content was 0.499, max 0.875 and min 0.125 (**S2**). Note that tandem complement barcode combinations can not be used on the Illumina NovaSeq system, therefore, only 9120 of the 9216 possible barcode combinations are viable when creating libraries intended for sequencing on that system. See http://sapac.support.illumina.com/bulletins/2017/08/recommended-strategies-for-unique-dual-index-designs.html for further details on this limitation of the NovaSeq system.

Nextera Flex sequencing libraries preparation

We first created a library using the standard protocol of Nextera Flex (referred to as "Standard Flex"). The Standard Flex library was constructed using all standard kit reagents from the Nextera DNA Flex Library Prep kit (Illumina, USA), following the manufacturer's protocol. Briefly, 200 ng input DNA in 10 ul nuclease free water was tagmented by adding 10 ul of Bead Link Transponsase (BLT) and 10 ul of TB1 solution. The sample was then incubated in the thermocycler at 55C for 15 mins, then held at 10C. After the incubation, 10 ul of TSB solution was added into the tagmentation reaction, and the sample was incubated at 37C for 15 mins, then held at 10C. The sample was then transferred to the magnet to isolate the DNA-BLT complex. The DNA-BLT complex was washed with 100 ul of TWB solution three times. The PCR reaction for library amplification was prepared by mixing 20 ul of Enhanced PCR Mix (EPM) with 20 ul of nuclease free water. The mixture was added into the DNA-BLT complex. 5 ul of each i5 and i7 adapter was added into the PCR reaction. The final volume of the PCR reaction is 50 ul. The condition of PCR was 68C for 3 mins, 98C for 3 mins, followed by 5 cycles of [98C for 30 sec - 62C for 30 sec - 68C for 2 mins], 68C for 1 mins and held at 10C. After library amplification, the sample tube was placed onto the magnet. Forty-five ul of the PCR supernatant was mixed with 85 ul of diluted SPB (45 ul of PB solution diluted in 40 ul of RSB solution), and incubated at room temperature for 5 mins. The sample tube was then placed on the magnet, and 125 ul of supernatant was transferred into a new sample tube containing 15 ul of undiluted PB. The sample was mixed and incubated at room temperature for 5 mins. Then, the tube was placed on the magnet. The supernatant was discarded, and the bead was washed with 200 ul of fresh 80% ethanol twice. The bead was left to air-dry at room temperature, and were resuspended in 32 ul of RSB solution. The bead was incubated at room temperature for 2 mins. The sample tube was placed on the magnet, and finally 30 ul of eluted library was transferred into a new sample tube. The concentration of eluted library and the library size were measured using Qubit High Sensitivity dsDNA kit (Thermo Fisher Scientific, USA) and the High Sensitivity Bioanalyzer chip (Agilent, USA), respectively.

We also created a library using 1:50 diluted BLT beads (referred to as "1:50 Flex"). The 1:50 Flex library was obtained by following the standard Nextera Flex protocol using the standard reagents, except for the BLT beads which were diluted 1:50 with nuclease free water prior to use. Only 10 ng of input DNA was used and the cycle number for library amplification PCR was adjusted to 12.

Both Standard Flex and 1:50 Flex libraries were purified, pooled in equal volumes, diluted to 4 nM and QC on the Bioanalyzer (Agilent Technologies, USA). The pool was sequenced on Illumina MiSeq platform 2x300 bp using MiSeq Reagent Kit V3 (600 cycles PE) cartridge (Illumina, USA).

Tagmentation and Hackflex sequencing library preparation

For Hackflex libraries, ninety-six libraries were prepared using laboratory-made and adapted reagents from the Nextera DNA Flex Library Prep kit (Illumina; **S1**). All incubation temperature and time used in the Hackflex protocol were the same as in the Standard Flex protocol except the PCR amplification step. Briefly, BLT beads were diluted 1:50 with nuclease free water (Invitrogen). 10 ng of input gDNA in 10 ul ultrapure water (Invitrogen) was mixed with 10 ul of 1:50 diluted BLT, and 25 ul of 2x laboratory-made tagmentation buffer (20 mM Tris (pH 7.6) (Chem-Supply), 20 mM MgCl (Sigma), and 50% (v/v) Dimethylformamide (DMF) (Sigma)). The final volume of the tagmentation reactions was 45 ul. After tagmentation, 10 ul of 0.2% of sodium dodecyl sulphate (SDS; Sigma) was added into the sample to stop tagmentation, instead of using TSB. Then, instead of TWB, the beads were washed three times using 100 ul of washing solution (0.22 µm MF-Millipore™ membrane filtered solution of 10% polyethylene glycol (PEG) 8000 (Sigma), 0.25M NaCl (Chem-Supply) in Tris-EDTA buffer (TE) (Sigma)). For library amplification, EPM master mix was replaced with the PrimeSTAR GXL DNA Polymerase kit (Takara), following the manufacturer protocol Each PCR reaction contains 10 ul of 5x GXL buffer, 4 ul of 25 mM dNTPs, 2 ul of PrimeStar GXL polymerase, 19 ul of nuclease free water. The PCR mix was added into washed BLT beads. Then, 5 ul of each custom synthesized 96-well plate Illumina Adapter Oligos i5 and i7 (i7: IDT plate#: 11680765; i5: IDT plate#: 11680754) (**S2**) were added to a final concentration of 0.555 uM to each reaction. The final volume for the PCR reaction is 45 ul. Library amplification was performed with different conditions from the manufacturer's recommended protocol, as follows: 3 min at 68C, 3 min at 98C, 12 cycles of [45 sec at 98C – 30 sec at 62C – 2 min at 68C], 1 min at 68C and hold at 10C. Then, size selection and purification of the library followed, replacing reagents SPB and RSB with equal volumes of SPRIselect beads (Beckman Coulter) and ultrapure water (Invitrogen) respectively. Reactions were then pooled in equal volumes. The concentration of the pooled library was measured with Qubit HS dsDNA kit (Thermo Fisher Scientific). Fragment size distribution was assessed using the High Sensitivity DNA kit on the Bioanalyzer (Agilent Technologies). The final library was diluted and denatured following manufacturer's instructions, then 4 pM of the pooled library with 5% PhiX v3 control (Illumina) was loaded onto an Illumina MiSeq instrument and sequenced using MiSeq V2 chemistry, generating 2 x 150 bp paired-end reads with a cluster density of 471 K/mm^2 (cluster passing filter 92%).

Preparation of additional Illumina libraries

During development of the Hackflex protocol (described above), we measured the effect of different polymerases used in the library amplification step, in particular the standard EPM master mix included in the Nextera DNA Flex kit and KAPA Master Mix (2xKAPA HiFi HotStart ReadyMix #KK2602; KAPA Biosystem, USA), on library yield and GC coverage bias. We measured the effect using genomic DNA from *S. aureus* strain ATCC 25923 (Sa) and *P. aeruginosa* strain POA1 (Pa). To do this, four different types of libraries were prepared for each gDNA sample: 1) Standard Flex with EPM master mix for library amplification (SF_1), 2) 1:50 Flex with EPM master mix (SF_1:50), 3) Hackflex with KAPA master mix (KAPA_1:50), 4) Hackflex but 1:20 BLT beads with KAPA master mix (KAPA_1:20). There were eight libraries in total: Sa_SF_1, Sa_SF_1:50, Sa_KAPA_1:50, Sa_KAPA_1:20, Pa_SF_1, Pa_SF_1:50, Pa_KAPA_1:50 and Pa_KAPA_1:20. The name of the library indicates the source of gDNA used and the library preparation. For example, the library Sa_SF_1 was the library generated from *S. aureus* ATCC25923 using Standard Flex with EPM master mix, and Pa_SF_1:50 was the library generated from *P. aeruginosa* POA1 using 1:50 Flex with EPM master mix. using different genomic DNA samples. Each library preparation condition is shown schematically in Supplementary Table 3 (**S3**). All additional libraries, except Sa_SF_1 and Pa_SF_1, were prepared using 10 ng input gDNA and 12 PCR cycles for library amplification. For Sa_SF_1 and Pa_SF_1, the libraries were prepared using 200 ng of input DNA and 5 PCR cycle for library amplification.

After library preparation, the concentration of each library was measured using Qubit HS dsDNA kit (Thermo Fisher Scientific) and the fragment size was analyzed with the High Sensitivity DNA kit on the Bioanalyzer (Agilent Technologies). Libraries were sequenced on Illumina MiSeq instrument, using MiSeq V3 chemistry, generating 2 x 300 bp paired-end reads.

Nanopore library preparation and sequencing

For long-read MinION sequencing, libraries were prepared using the 1D ligation sequencing kit (SQK-LSK108) from Oxford Nanopore Technologies (ONT) with modifications to the standard ONT protocol as described previously [6]. The sample was barcoded using the Native Barcoding Expansion kit (EXP-NBD103) and barcoded templates were then pooled together with two other samples from an unrelated project. The final library was loaded onto a ONT MinION instrument with a FLO-MIN106 (R9.4) flow cell and run for 48 h as per the manufacturer's instructions. Live base-calling was not performed during the run. Nanopore sequence data was combined with Illumina sequence data into a hybrid genome assembly using the Unicycler software, version 0.46. Unicycler was run with default parameters in "normal" mode.

Data analysis

All the data analysis methods are described below and represented schematically in Supplementary Figure S4 (**S4**).

Barcode demultiplexing

Hackflex reads were demultiplexed with Bcl2fastq (Bcl2Fastq 2.18.0.12, Illumina, Inc.) software with default settings, allowing one mismatch per index. Barcode cross-contamination was quantified with the PhyloSift command `demux`. Barcode counts were retrieved from the demultiplexing statistics output of Bcl2fastq and histograms representing barcode distribution were generated with R Studio, version 1.1.463 (RStudio: Integrated Development Environment for R, Boston, Massachusetts).

Processing of libraries before mapping

Raw reads were assessed for quality with FastQC version 0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and using the Bioconductor package Rsubread [7]. PHRED scores were plotted in R studio. Trimming and normalization of the libraries were performed using the bbtools package (http://jgi.doe.gov/data-and-tools/bbtools). Due to operational constraints, Standard Flex and 1:50 Flex libraries were sequenced on a different flow cell than the Hackflex library, producing 300 bp reads instead of 150 bp reads (as with Hackflex). For this reason, reads from Standard Flex and 1:50 Flex were first right-end trimmed to 150 bp with BBDuk (http://jgi.doe.gov/data-and-tools/bbtools).

Quality trimming, PhiX and adapters removal were performed on all the libraries with BBDuk (parameters: `trimq=20, maxgc=0.98 qtrim=r entropy=0.5 minavgquality=25`), removing 58.59%, 62.75%, and 66.95% of the reads from Standard Flex, 1:50 Flex and Hackflex, respectively. The final read counts of the cleaned libraries were: 776728, 690930 and 818478 reads for Standard Flex, 1:50 Flex and Hackflex, respectively (**Table 1; S4**). With 1:50 Flex having the lowest number of reads, Standard Flex and Hackflex were downsampled with the bbmap script *reformat.sh* to match the total number of bases of 1:50 Flex (parameter `-samplebasestarget=100118301`). Quality trimming, removal of adapters and downsampling were performed likewise for all additional libraries (see above).

Hybrid assembly of Nanopore and Illumina data

The nanopore sequence data was co-assembled with Nextera Flex data using Unicycler v0.4.6 with default parameter settings. The resulting assembly was fragmented (10 contigs of sizes: 4466582 nt, 110260 nt, 59518 nt, 4535 nt, 1797 nt, 1356 nt , 476 nt, 292 nt, 184 nt, 181 nt).

Short read mapping and coverage analysis

Short read mapping of Standard Flex, 1:50 Flex and Hackflex was computed against all 10 contigs of the fragmented reference assembly with samtools[8] version 1.7 (http://samtools.sourceforge.net/) and the bwa-mem[9] (http://bio-bwa.sourceforge.net/) alignment software (**S4**). After mapping, PCR duplicates were removed using the samtools command `markdup`. For simplicity, only coverage data from the largest hybrid assembly contig (4.46Mbp, 96.1% of the total assembly) was analyzed and is taken to be representative of the genomic coverage as a whole. To avoid potential problems associated with mapping reads near contig boundaries, coverage data were trimmed to remove 200 nt from each end of the contig. Reads that failed mapping were analysed with Kraken [10], using the command `krakenhll` and default parameters (**S4**). Histograms representing the coverage distribution were generated by using the mapped depth of coverage information as computed by samtools `mpileup` for each position on the *E. coli* MG1655 reference genome and plotting the frequency of each level of coverage with R Studio, version 1.1.463 (RStudio: Integrated Development Environment for R, Boston, Massachusetts). Low coverage regions were visualized with the Integrative Genomic Viewer (IGV)[11] (**S4**). Short read mapping and coverage analysis were performed for 4 additional libraries in the same manner as described above. Libraries Sa_SF_1:50 and Sa_SF_1 were mapped against *S. aureus* ATCC 25923 reference genome (CP009361.1). Libraries Pa_SF_1:50 and Pa_SF_1 were mapped against *P. aeruginosa* PAO1 reference genome, which was obtained co-assembling libraries 3 and 4 with A5-miseq [12].

GC content

Coverage by GC content was obtained by binning the *E. coli* MG1655 reference genome into 102 bins and calculating the GC content of each bin with ALFRED [13]. Likewise, coverage by GC content of additional libraries Sa_SF_1:50, Sa_SF_1, Pa_SF_1:50, and Pa_SF_1 were obtained by binning the *S. aureus* ATCC 25923 reference genome (libraries Sa_SF_1:50 and Sa_SF_1) or the *P. aeruginosa* PAO1 (libraries Pa_SF_1:50

and Pa_SF_1) into 102 bins and calculating the GC content of each bin with ALFRED[13] (**S4**). Coverage of each bin for each library was plotted with R Studio, version 1.1.463 (RStudio: Integrated Development Environment for R, Boston, Massachusetts). (**S4**)

Data availability

All sequence data has been deposited in NCBI under accession number PRJNA549801.

RESULTS

Hackflex library preparation and sequencing

Our customised library preparation protocol, Hackflex, involves several modifications to the Nextera Flex method, including the use of a 1:50 dilution of the bead-linked transposase and the replacement of several kit components with alternative reagents to greatly expand the total number of libraries that can be produced from a single kit. To evaluate the performance of the Hackflex protocol, we performed in parallel a sequencing library preparation with the standard Nextera Flex protocol (which we refer to as "Standard Flex") and with an adapted version of the Nextera Flex protocol using the diluted transposase beads but using standard kit components (we refer to as "1:50 Flex"). Libraries Standard Flex, 1:50 Flex and Hackflex were prepared from genomic DNA of *E. coli* strain MG1655. Libraries were sequenced on an Illumina MiSeq. Read counts obtained, read- and library metrics are shown in **Table 1**.

Barcode distribution and quality

Ninety-six i5 and i7 barcodes (8 bp) were designed for this study to provide a resource for high throughput multiplexing of Hackflex libraries. To this end, performance of the 96 designed barcodes was evaluated by subjecting *E. coli* MG1655 DNA to 96 independent library constructions with Hackflex reagents, each library with a different barcode combination (**S2**). In order to assess uniform performance of individual barcodes, barcode calling analysis was performed. Barcode analysis with the `phylosift demux` command showed a high rate of correctly paired barcoded reads (99.84%), with a remaining 0.16% which can be attributed either to oligonucleotide cross-contamination or errors during sequencing such as the presence of multiple overlapping clusters on the flowcell surface, base miscalls, and image processing errors during cluster calling. With the exception of two failed libraries (with 6 and 9 reads), the relative abundance of barcodes across the 96 libraries was homogenous. The average barcode count is 5517 barcodes per sample and 50% of the samples contain between 4292 and 7128 barcodes. (**Fig. 2**) Ninety-eight percent (94 out of 96) of the libraries fall within a 6.8-fold range of relative abundance (**S5**). Coefficient of variation is 0.38 with, and 0.34 without the two outliers.

The GC content of the oligos designed for this study was measured and plotted against the read count obtained from each oligo pair. A lower yield was obtained from libraries constructed with higher GC content barcodes (**S6**), possibly due to the tandem complement design of the barcode pairs.

The two failed libraries have a normal GC content (38% and 50%), falling within the GC range of the Hackflex barcodes where 79.2% of the barcodes (76/96) have between a 30% and 70% GC content. The estimated deltaG of primer dimers for the two outliers did not differ from that estimated for the other oligo pairs.

Quality of raw reads

Quality scores from the three libraries were obtained using the Bioconductor package Rsubread [7]. PHRED scores increased from <30-34 to 36-38 within the first 25 nt of each read, independent of the library preparation method (**Fig. 3**). Reads from the Hackflex library started from a lower score (29-32) and reached the same PHRED score of reads from Standard Flex and 1:50 Flex (36-38) within the first 25 nt. All reads from each library reached a maximum PHRED score in the first 25 nucleotides, to decrease slightly after the first 25 nt from 36-38 to 34-36 for Hackflex, while reads from Standard Flex and 1:50 Flex decreased from 36-38 to 32-34 across the first 150 nt of each read. (**Fig. 3**)

Cleaning

Quality filtering, PhiX DNA and adapter removal with BBDuk resulted in the removal of 58.59%, 62.75%, and 66.95%, leaving a total of 776728, 690930 and 818478 reads for Standard Flex, 1:50 Flex and Hackflex, respectively. Median read length was 151 nt for all libraries. (**Table 2**) With 1:50 Flex being the library with fewest reads, Standard Flex and Hackflex were randomly subsampled to the same number of reads as 1:50 Flex.

Fragment size

Quality filtered and trimmed reads were mapped against the reference sequence with samtools and bwa mem. The .bam output was converted to text with samtools to report fragment size. The text file was analyzed with R studio and fragment size versus read density was plotted for Standard Flex, 1:50 Flex and Hacklex libraries (**Fig. 4**). The observed fragment size distribution appeared uniform for all three libraries, with 1:50 Flex reads being more skewed to the left, representing a higher density of <250 bp fragments, Standard Flex reads being slightly skewed to the right, representing a higher density of larger fragments compared to the other two libraries. Hackflex showed a centered distribution, with the highest density of fragments being between >250-300 bp. (**Fig. 4**)

Coverage

In order to assess performance of the library prep methods, reads were aligned with BWA-MEM [9] and samtools [8] to the *E. coli* MG1655 genome. The mapping files of the libraries were converted to .tsv and analyzed with R studio for coverage. A mapped fraction of 0.999, 0.999 and 0.998 was measured for Standard Flex, 1:50 Flex and Hackflex, respectively (**Table 2; S7**). Unmapped reads from the Hackflex library appear to derive from an unknown source of contamination, as suggested by Kraken [10] analysis (**S8**). The libraries showed a normal distribution (**Fig. 5**) with a median read count of 17, 18 and 18 reads and an average coverage depth of 16.75, 18.04 and 18.01 reads for Standard Flex, 1:50 Flex and Hackflex, respectively. The correlation in per-site coverage between libraries was high, with Standard Flex and 1:50 Flex showing the highest similarity (correlation coefficient: 0.4289; p-value: < 2.2e-16), followed by 1:50 and Hackflex (correlation coefficient: 0.3773; p-value: < 2.2e-16) and Standard Flex and Hackflex (correlation coefficient: 0.3674; p-value: < 2.2e-16).

GC content

In order to assess the correlation between GC coverage bias and sequence coverage, the GC content of the *E. coli* MG1655 reference genome and reads from Standard Flex, 1:50 Flex and Hackflex libraries were assessed. To this end the mapping (.bam) output of each library was converted to .tsv with the samtools command `mpileup` [8] as above and analyzed with ALFRED. The output was loaded into R studio and reference genome GC content (102 bins) against coverage was plotted for the three libraries (**Fig. 6**). A significant negative correlation between GC content and coverage was seen for Standard Flex, 1:50 Flex and Hackflex libraries where the GC content ranged between 30% and 70%. All three libraries showed to a certain extent a bias at extreme GC content areas as it would be expected [12,13]. The extent of bias was highest for 1:50 Flex (ρ = -0.959; p-value: 1.04e-112), lower for Standard Flex (ρ= -0.950; p-value: 6.62e-100) and lowest for Hackflex (ρ= -0.770; p-value: 1.847e-42). All tests of correlation were carried out using weighted Pearson's correlation coefficient on the observed/expected read count ratios, using the read counts from the 102 GC bins as weights.

Low coverage regions

Standard flex, 1:50 Flex and Hackflex each produced 5, 10 and 11 regions, respectively, with low coverage (<3 reads per site) and no sites with zero coverage (**S10-11**). The low coverage regions were overlapping to a certain extent among the three libraries, possibly indicating a common feature of these regions that biases against their sequencing with Illumina chemistry. The size of the low coverage regions and their position on the reference genome are displayed in Supplementary Figure **S10**, where it is possible to notice to what extent each low coverage region was exclusive to a library or common to two or all three of the libraries. The low coverage in those areas does not appear to be associated with GC content extremes (**S10-S11**).

Additional libraries: yield and GC coverage bias

Additional libraries were prepared in order to test the performance of 2x KAPA HiFi HotStart ReadyMix #KK2602 (KAPA) polymerase with the Hackflex reagents in terms of yield and GC coverage bias. The yield using 2x KAPA HiFi HotStart ReadyMix #KK2602 (KAPA) with Hackflex reagents ranged from 23.3 to 34.6 nM compared to that of EPM with Standard Flex reagents of 18.5 to 23.1 nM (**S3**) and to that of PrimeSTAR with Hackflex reagents of 43.1 to 52.4 nM.

As expected, a higher GC coverage bias was seen from libraries produced with 12 rather than 5 PCR cycles (**S9**).

DISCUSSION

The Hackflex library prep workflow we have introduced is as time effective as the Standard Nextera Flex method and yields significant savings in terms of reagent costs (from 1.66-fold for 96x2 samples to 8.23-fold for 96x50 samples). This study demonstrates that data of comparably high quality can be obtained with Hackflex as could be generated by the existing Nextera Flex method.

Two out of ninety-six libraries made with Hackflex yielded a low read count, which could be attributed to human error during the error prone step of sample indexing prior to amplification. We suggest as a further improvement to the Hackflex protocol, the automation of the sample indexing step using liquid handling robots so as to eliminate human error.

It is worth noting that the libraries Standard Flex, 1:50 Flex and Hackflex in this study have been constructed with different polymerases. Standard Flex and 1:50 Flex were constructed with EPM from the Nextera Flex kit (Illumina), while PrimeSTAR GXL (Takara) was used for Hackflex. Therefore, the small differences in coverage observed in this study may be attributable to the different polymerases used. Before opting for PrimeSTAR GXL (Takara), we tested the performance of KAPA HiFi HotStart ReadyMix #KK2602 (KAPA) as used by Lamble et al [14], but we observed half the yield from these PCR reactions (**S3**) as with PrimeSTAR GXL. This stands in contrast to the behavior of KAPA HiFi when coupled with the original Nextera and Nextera XT protocols, where it produces high library yields.

Additionally, PrimeSTAR GXL can be decreased to 1.25 units per reaction (50% of the amount used in this study) as per manufacturer's instructions, without compromising the quality of the library (data not shown), further reducing the costs of Hackflex from 1.75-fold for 96x2 to 11-fold for 96x50 samples (**S1**).

Although there is no indication from the present study of Hackflex performing worse than Nextera Flex with genomes having extreme GC content and with lower DNA inputs, this remains to be tested more comprehensively in future work.

Caveats and limitations

The i5 and i7 barcodes we describe in this work have a tandem complement design, where the corresponding wells of the i7 barcode oligo plate have the reverse complement of the barcode of the corresponding well in the i5 plate. It has been noted by Illumina that the use of tandem complement barcodes on the current generation NovaSeq instruments can lead to significantly reduced quality scores for the i5 index read (http://sapac.support.illumina.com/bulletins/2017/08/recommended-strategies-for-unique-dual-index-designs.html). In this study our samples were sequenced on an Illumina MiSeq instrument, which does not appear to be prone to the tandem complement barcode limitation. Although the index read 2 quality of our data was not noticeably impacted when compared to Standard Flex, we did measure a trend towards lower read count of libraries made from higher GC content oligos (**S6**). Possibly this effect is due to the tandem complement design of the oligo pairs.

On a positive note, the self-designed barcode oligos produced by IDT using their standard oligo plate manufacturing process appeared to yield a relatively low cross-contamination rate.

CONCLUSION

Here we have developed and characterised an alternative method of library construction for Illumina sequencing which by reducing the library prep expenses, allows users to process from 1.75-fold to 11-fold more samples at the same reagent cost. Comparison with the existing Nextera Flex method demonstrates that Hackflex is a valid and cost-effective alternative to construct libraries at a large scale.

Conflicts of interest

A.D. and L.M. have a commercial interest in Longas Technologies Pty Ltd, which is developing synthetic long read sequencing technologies for short read sequencing platforms.

Funding information

This work was funded in part by ARC Linkage Project LP15100912, Lead CI: Steven P. Djordjevic.

Author contributions

Conceived and designed the experiments: DG ML JT KA LM AD. Performed the experiments: DG ML JT LM KA. Analyzed the data: DG AD. Contributed reagents/materials/analysis tools: JT ML KA LM. Wrote the paper: DG AD LM.

REFERENCES

1. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
2. Baym, M. *et al.* Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE* **10**, e0128036 (2015).
3. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
4. Bruinsma, S. *et al.* Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics* **19**, 722 (2018).
5. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
6. Monahan, L. G. *et al.* High contiguity genome sequence of a multidrug-resistant hospital isolate of Enterobacter hormaechei. *Gut Pathog.* **11**, 3 (2019).
7. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47–e47 (2019).

8. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

9. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

10. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

11. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

12. Coil, D., Jospin, G. & Darling, A. E. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinforma. Oxf. Engl.* **31**, 587–589 (2015).

13. Rausch, T., Hsi-Yang Fritz, M., Korbel, J. O. & Benes, V. Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics* doi:10.1093/bioinformatics/bty1007

14. Lamble, S. *et al.* Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnol.* **13**, 104 (2013).

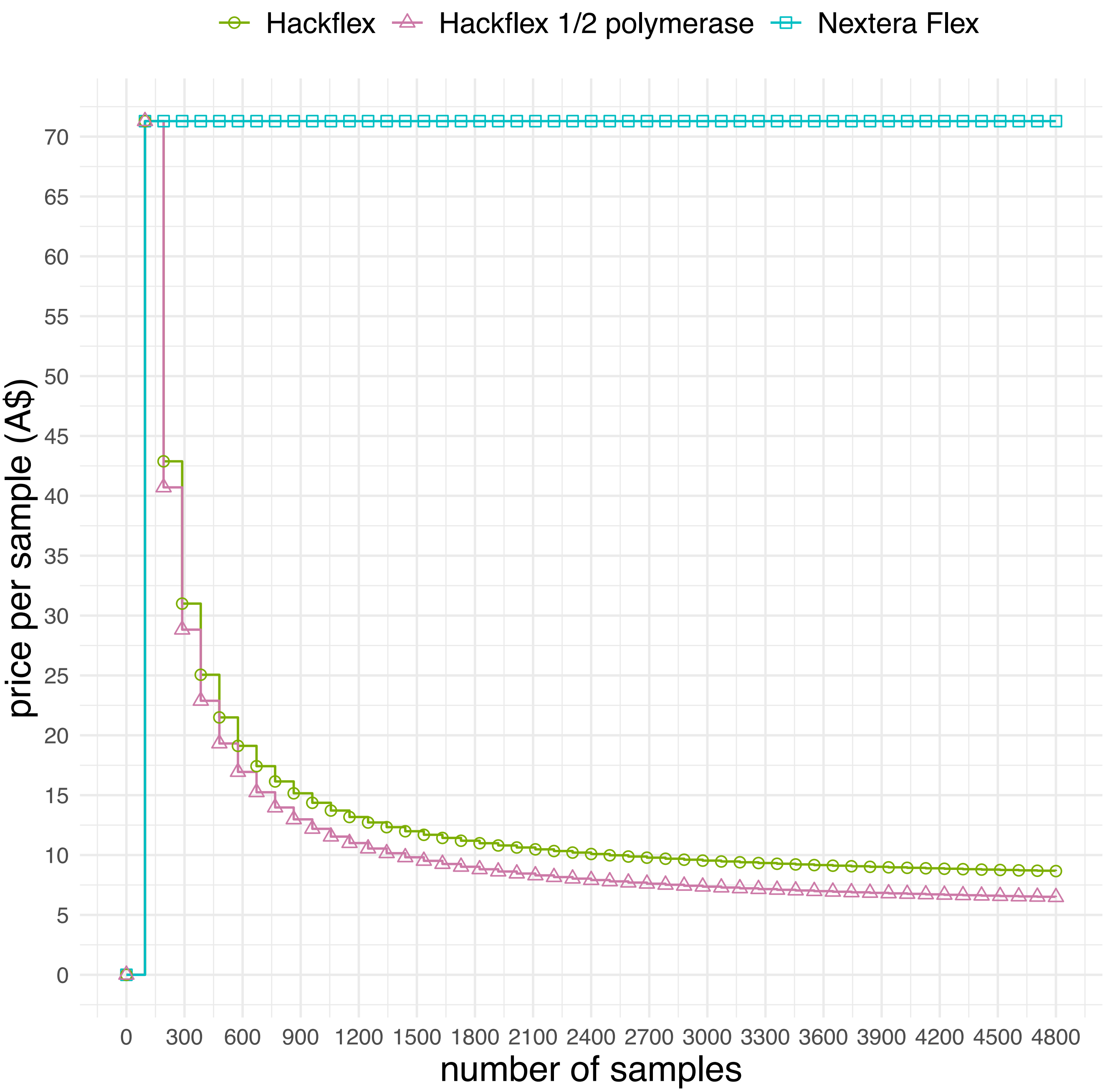


Figure 1.
Price per sample (in AUD) when processing up to 96x50 (4800) samples with Nextera Flex (blue) or Hackflex reagents using 2.5 units (green) or 1.25 units (pink) of PrimeSTAR GXL per reaction.

Table 1.

Overview of protocol used, number of reads obtained and filtered, with the three library construction methods.

Library	Protocol	Input DNA	Polymerase	PCR cycles	Total reads generated	Reads removed in cleaning steps (%)	Total number of reads after cleaning	Median read length (nt)
Standard Flex	Nextera Flex kit	200 ng	EPM (Nextera Flex kit)	5	1,325,670	58.59%	776,728	151
1:50 Flex	Nextera Flex kit (1:50 BLT dilution)	10 ng	EPM (Nextera Flex kit)	12	1,101,114	62.75%	690,930	151
Hackflex	Adapted protocol	10 ng	PrimeSTAR (Takara)	12	1,222,468	66.95%	818,478	151

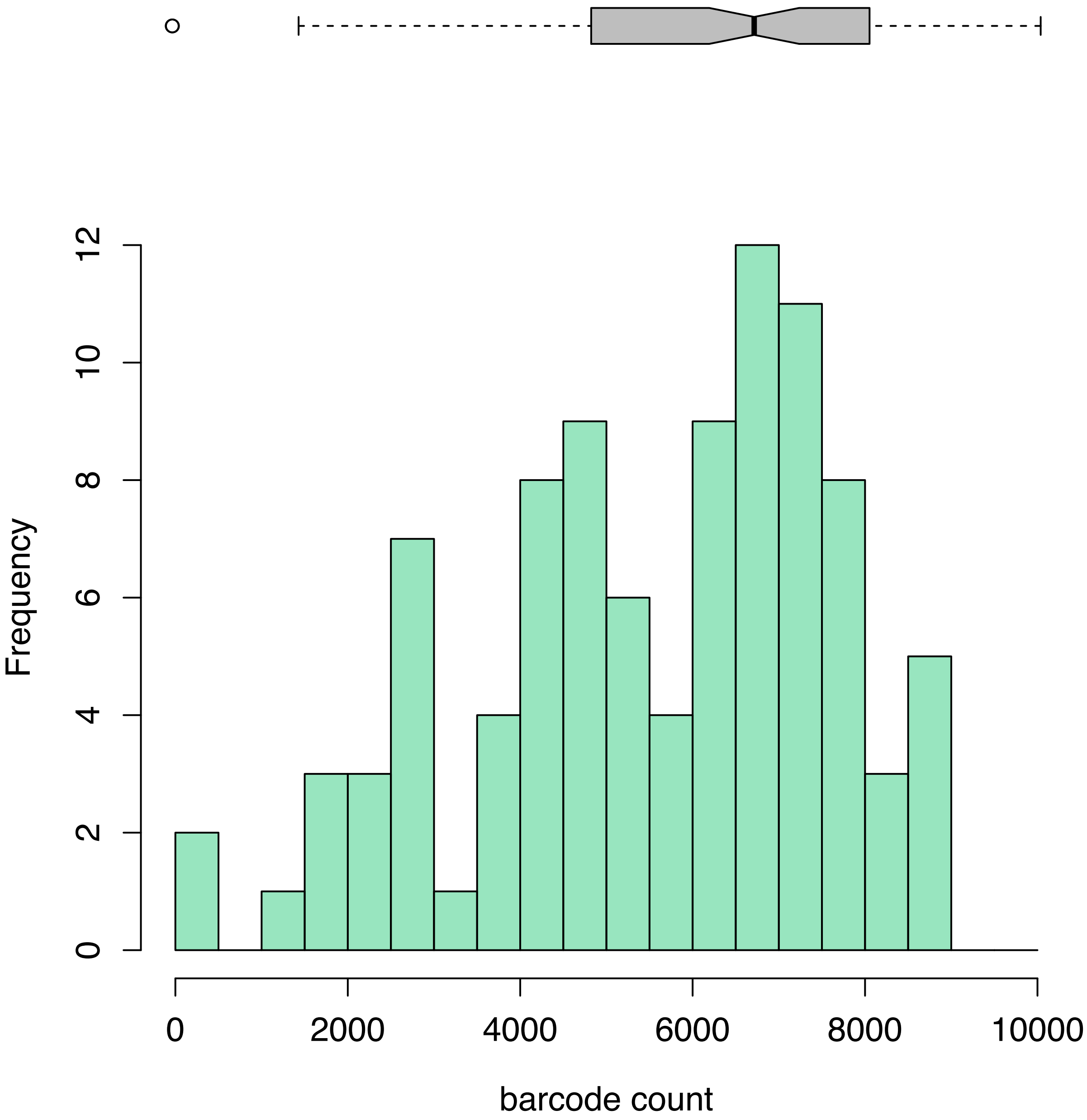


Figure 2.
Unique barcode distribution across the 96 Hackflex libraries (1st quartile: 4292; median: 5954; mean: 5517; 3rd quartile: 7128; standard deviation: 2074.66).

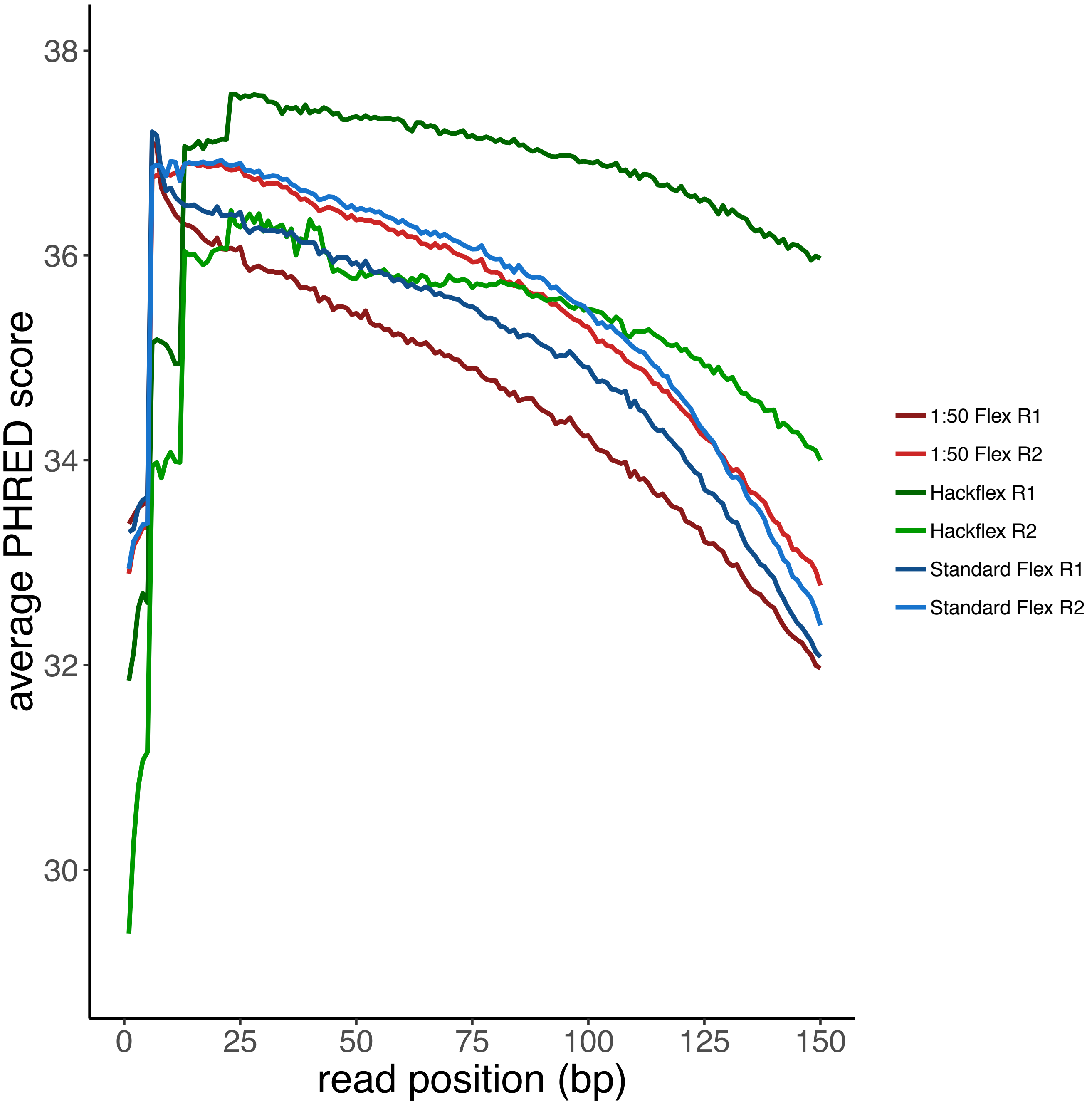


Figure 3.
Quality (PHRED) scores for reads from Standard Flex, 1:50 Flex and Hackflex library before quality filtering and trimming of the first 150 bp of each read.

Table 2.
Coverage obtained from mapping libraries against all contigs of the *E. coli* MG1655 nanopore assembly (4,645,181 bp). Total number of reads after cleaning indicating the number of paired end reads obtained from each library and coverage observed.

Library	Protocol	Total number of reads after cleaning	Mapped reads	Unmapped reads	Mismatch rate	Average coverage depth	Coverage coefficient of variation
Standard Flex	Nextera Flex kit	776,728	684,234 (99.9%)	60	0.0008	16.75X	0.319588814
1:50 Flex	Nextera Flex kit (1:50 BLT dilution)	690,930	690,855 (99.9%)	75	0.0009	18.04X	0.323339352
Hackflex	Adapted protocol	818,478	669,679 (99.8%)	1307	0.0006	18.01X	0.316022228

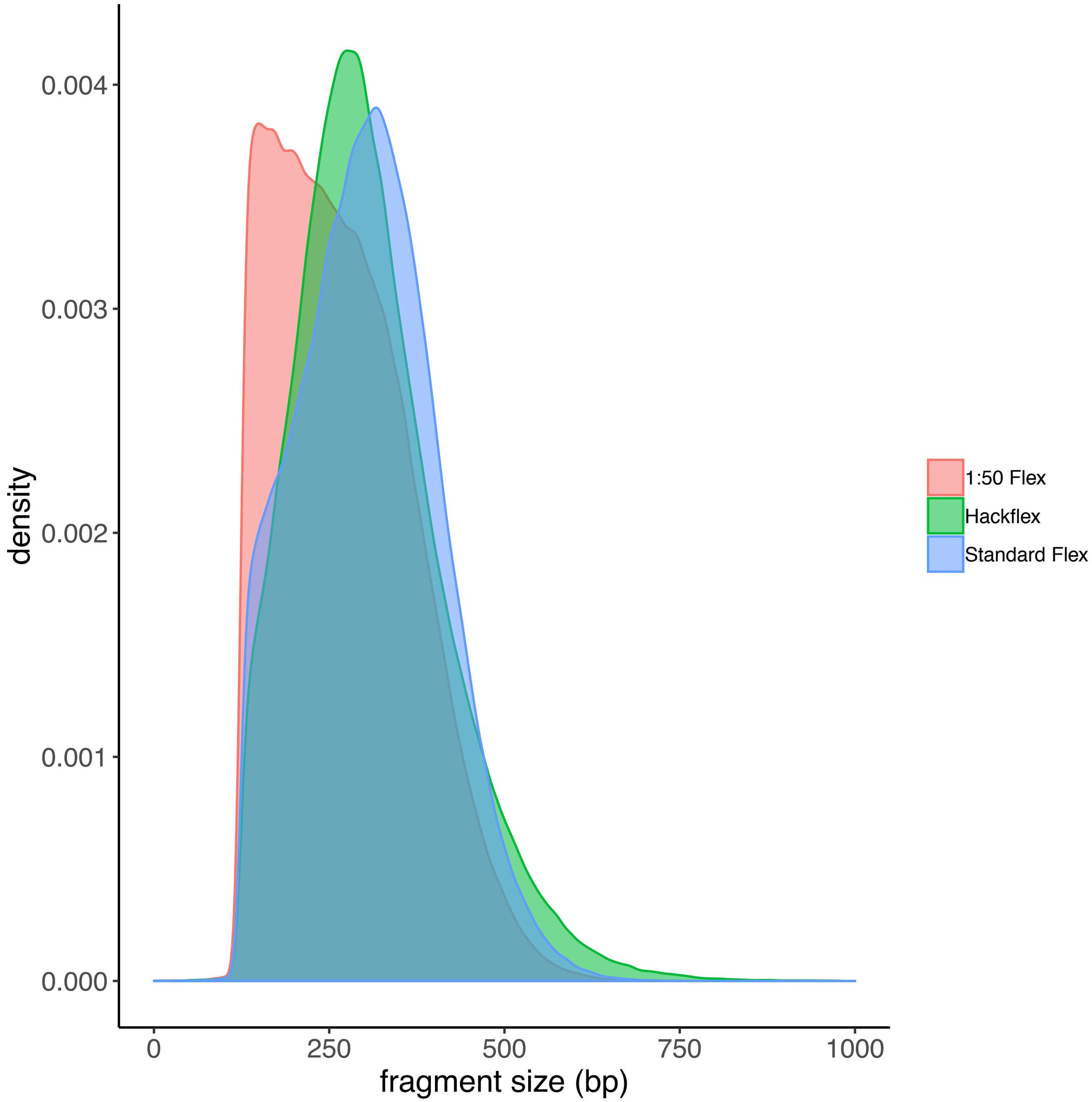


Figure 4.
Fragment size (genomic distance between reads in a read pair) distribution of Standard Flex, 1:50 Flex and Hackflex after quality filtering and trimming.

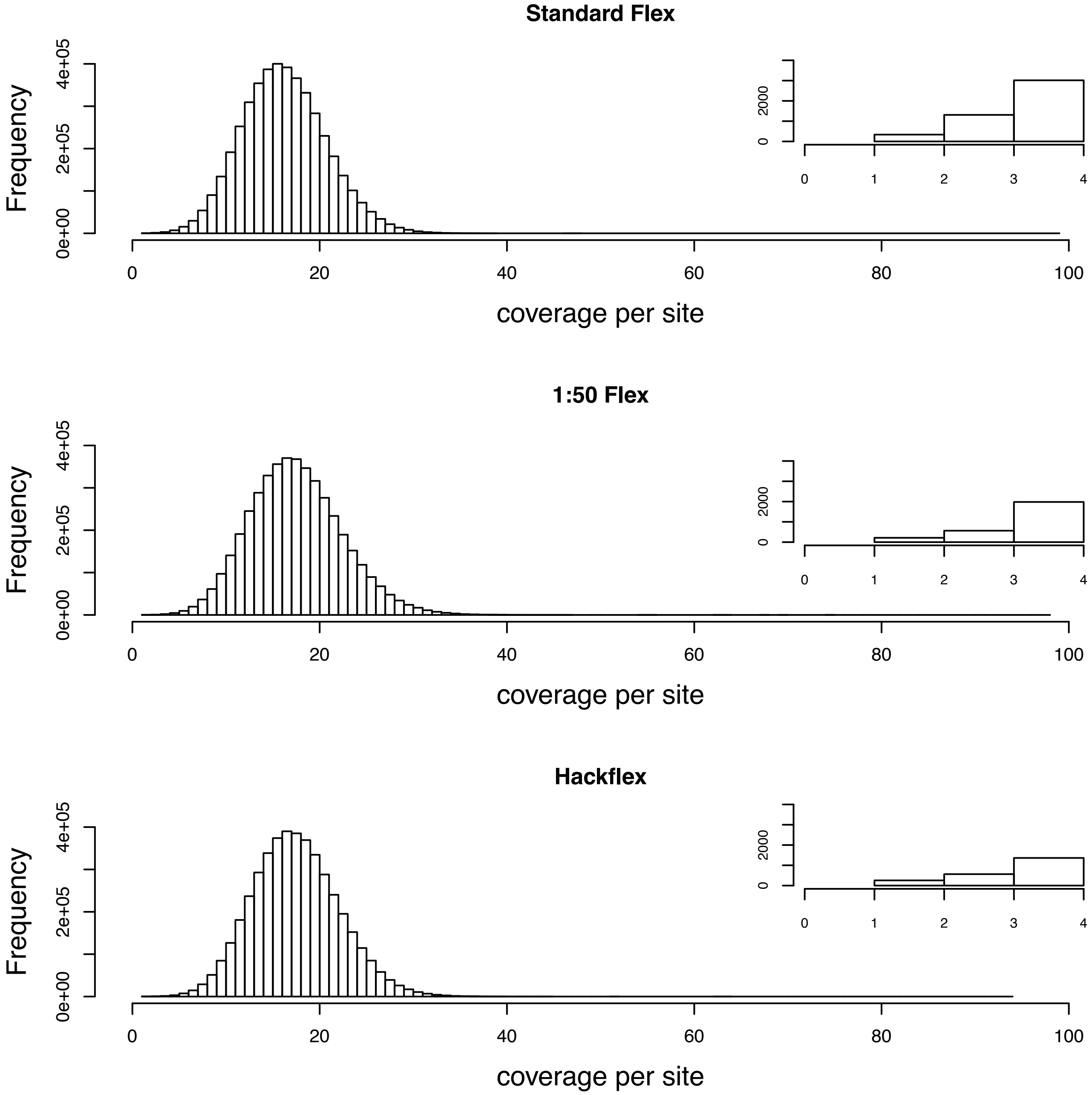


Figure 5.

Coverage distribution of unique reads across the *E. coli* MG1655 genome obtained with Standard Flex (top), 1:50 Flex (center) and Hackflex (bottom). Histograms on the top right of each plot show the lower-coverage end for each library. Raw reads were trimmed, quality filtered, down-sampled and PCR duplicates were removed, then aligned to the reference genome using samtools [8] and bwa-mem [9] .

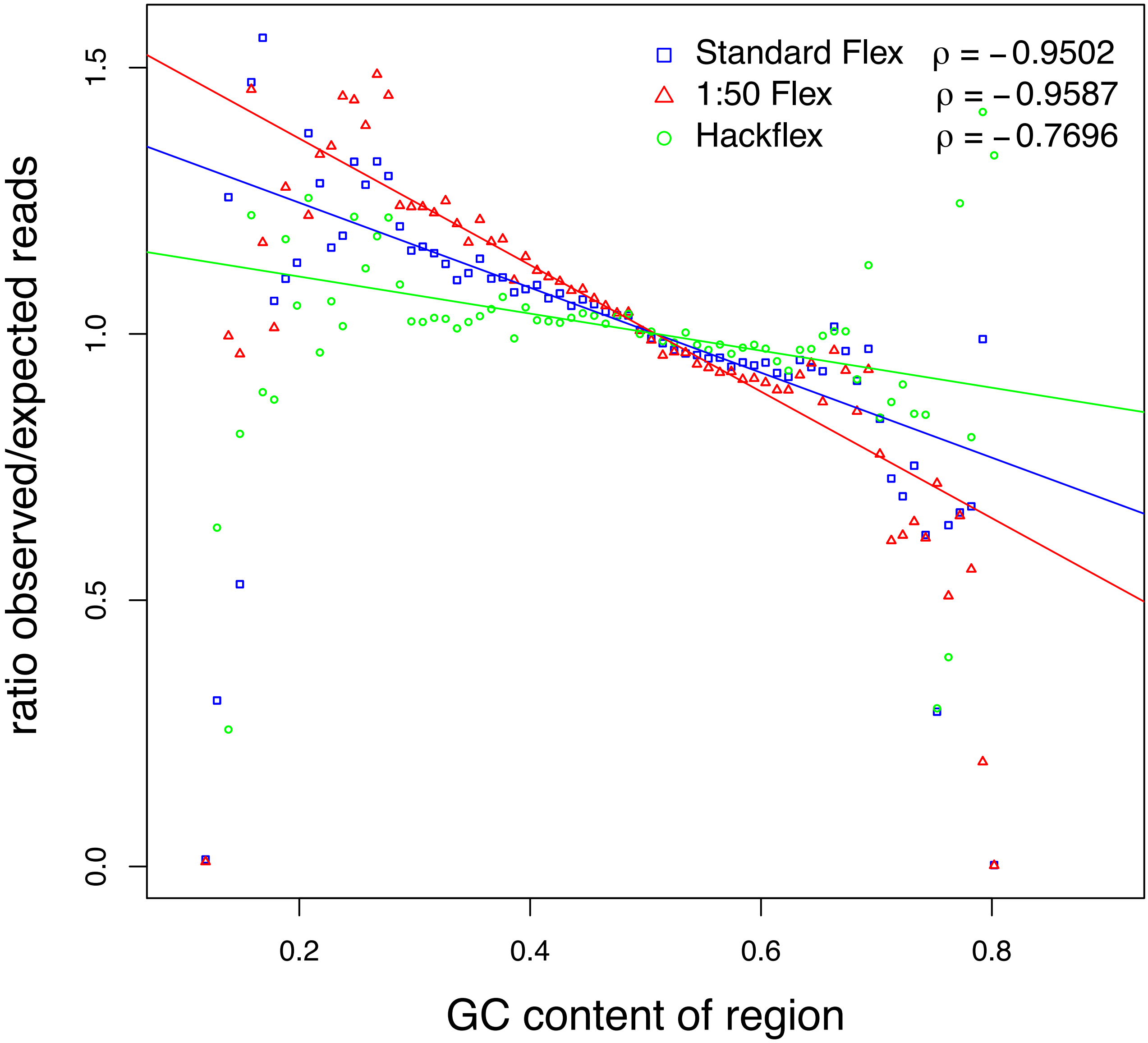


Figure 6.
GC content bias assessed by calculating the GC content of the *E. coli* MG1655 reference genome (102 bins) and plotting the coverage obtained with Standard flex (blue square), 1:50 Flex (red triangle) and Hackflex (green circle). The GC fraction of our *E. coli* MG1655 largest contig is 0.508402 (2270820/4466582 bp). Regression lines and weighted Pearson's correlation coefficients between GC content and coverage as displayed (p-values: Standard Flex: 3.05922e-104; 1:50_Flex: 3.112934e-112; Hackflex: 3.261771e-41).

Supplementary Table 1 (S1). Quantities and costs of Hackflex reagents as per retail price.

Nextera Flex	Hackflex reagent	Quantity per reaction	Batch cost (A$)	Quantity per 4800 reactions	Price per 4800 reactions (A$)	Price per 96 reactions (A$)
BLT	BLT 1:50	0.2 ul	$6844/4800 reactions	960 ul	6844	136.880
	nuclease free water	9.8 ul	$32.38/500 ml	47.040 ml	3.108	0.062170
TB1	20mM Tris	0.0000606 g	$615/5 kg	0.291 g	0.0358	0.000715
	20 mM MgCl	0.0000477 g	$237/5 kg	0.229 g	0.0108	0.000217
	50% DMF	12.5 ul	$41.90/250 ml	60 ml	10.056	0.201120
	nuclease free water	12.5 ul	$32.38/500 ml	60 ml	3.886	0.077712
TSB	0.2% SDS	0.0003 ul	$54.70/25 g	1.25 g	2.735	0.054700
	nuclease free water	10 ul	$32.38/500 ml	48 ml	3.108	0.062170
TWB	10% PEG 8000	0.01 g	$172/1 kg	48 g	8.256	0.165120
	0.25 M NaCl	0.0014609 g	$82/5 kg	7.013 g	0.115	0.002300
	TE	100 ul	$113/500 ml	480 ml	108.48	2.169600
EPM	polymerase	2 ul	1,582/1000 units/800 ul	9,600 ul	*20890.23	417.805
	dNTPs	4 ul	NA	NA	NA	NA
	5x GXL buffer	10 ul	NA	NA	NA	NA
	i5	5 ul	**	**	**	**
	i7	5 ul	**	**	**	**
	nucl.free water	19 ul	$32.38/500 ml	91.2 ml	5.906	0.118122
SPB	SPRI	45 ul	$1900/60 ml	216 ml	6840	136.800
RSB	nuclease free water	45 ul	$32.38/500 ml	216 ml	13.988	0.279763
Total	-	-	-	-	-	694.68

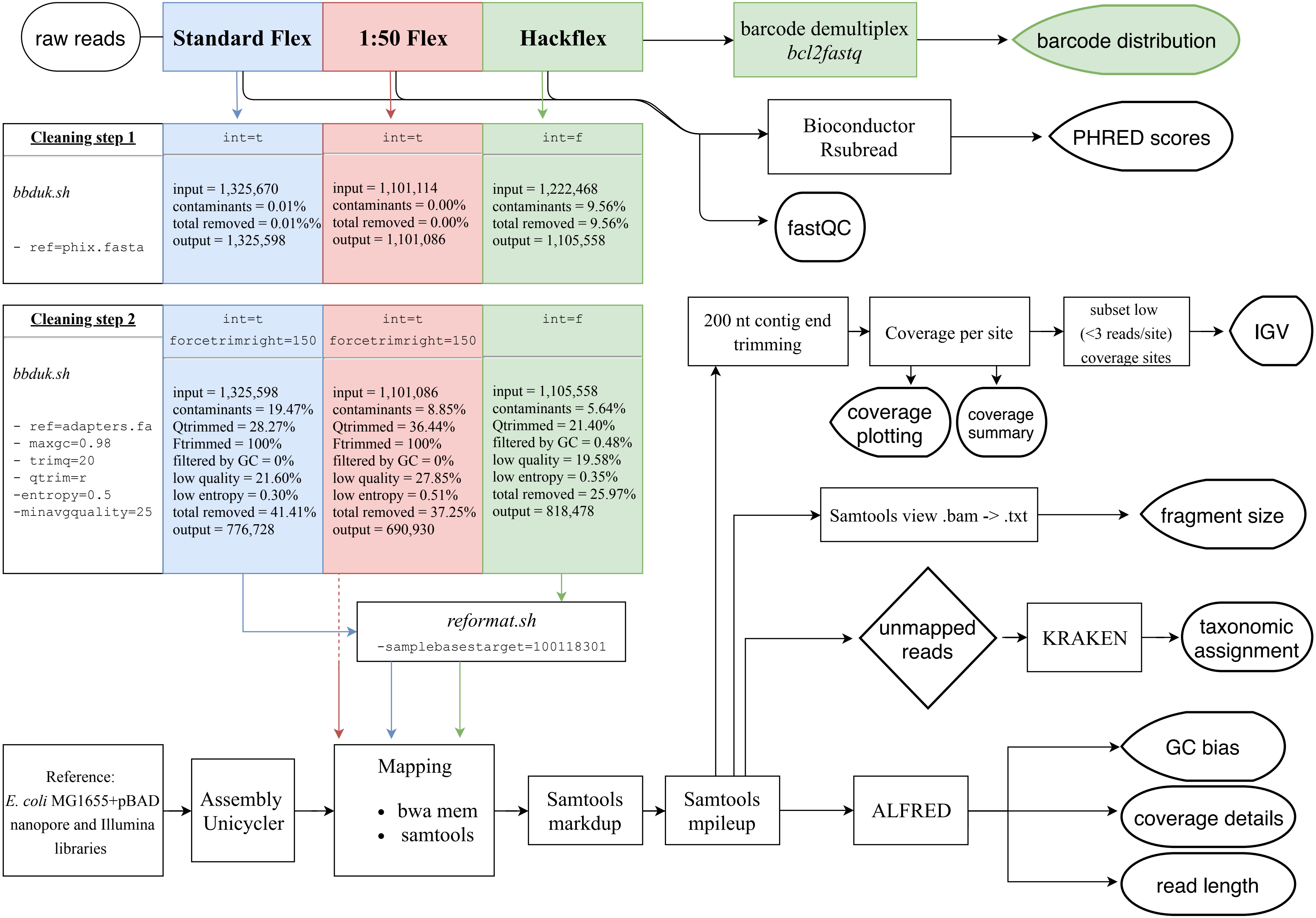
*corrected accounting for pipette error, 727 ul usable instead of 800 ul

**i5 and i7 oligos prices not included as sold separately from library prep kit, in both Hackflex as Nextera Flex.

NA: missing values for dNTPs and 5X GXL buffer as reagents costs are included in the polymerase row (together sold as kit: PrimeSTAR GXL DNA Polymerase kit (Takara)).

Supplementary Table 3 (S3).
Summary of protocol details for the construction of the additional libraries and obtained yields.

Library ID	Protocol	Polymerase	Species	BLT dilution	PCR cycles	Input DNA (ng)	Yield (nM)
Sa_SF_1:50	Standard Flex	EPM	*S. aureus*	1:50	12	10	20.6
Sa_SF_1	Standard Flex	EPM	*S. aureus*	1	5	200	18.4975
Pa_SF_1:50	Standard Flex	EPM	*P.aeruginosa*	1:50	12	10	18.6725
Pa_SF_1	Standard Flex	EPM	*P.aeruginosa*	1	5	200	23.0535
Sa_KAPA_1:50	Hackflex	2x KAPA	*S. aureus*	1:50	12	10	25.3415
Sa_KAPA_1:20	Hackflex	2x KAPA	*S. aureus*	1:20	12	10	34.558
Pa_KAPA_1:50	Hackflex	2x KAPA	*P.aeruginosa*	1:50	12	10	23.2525
Pa_KAPA_1:20	Hackflex	2x KAPA	*P.aeruginosa*	1:20	12	10	33.1865



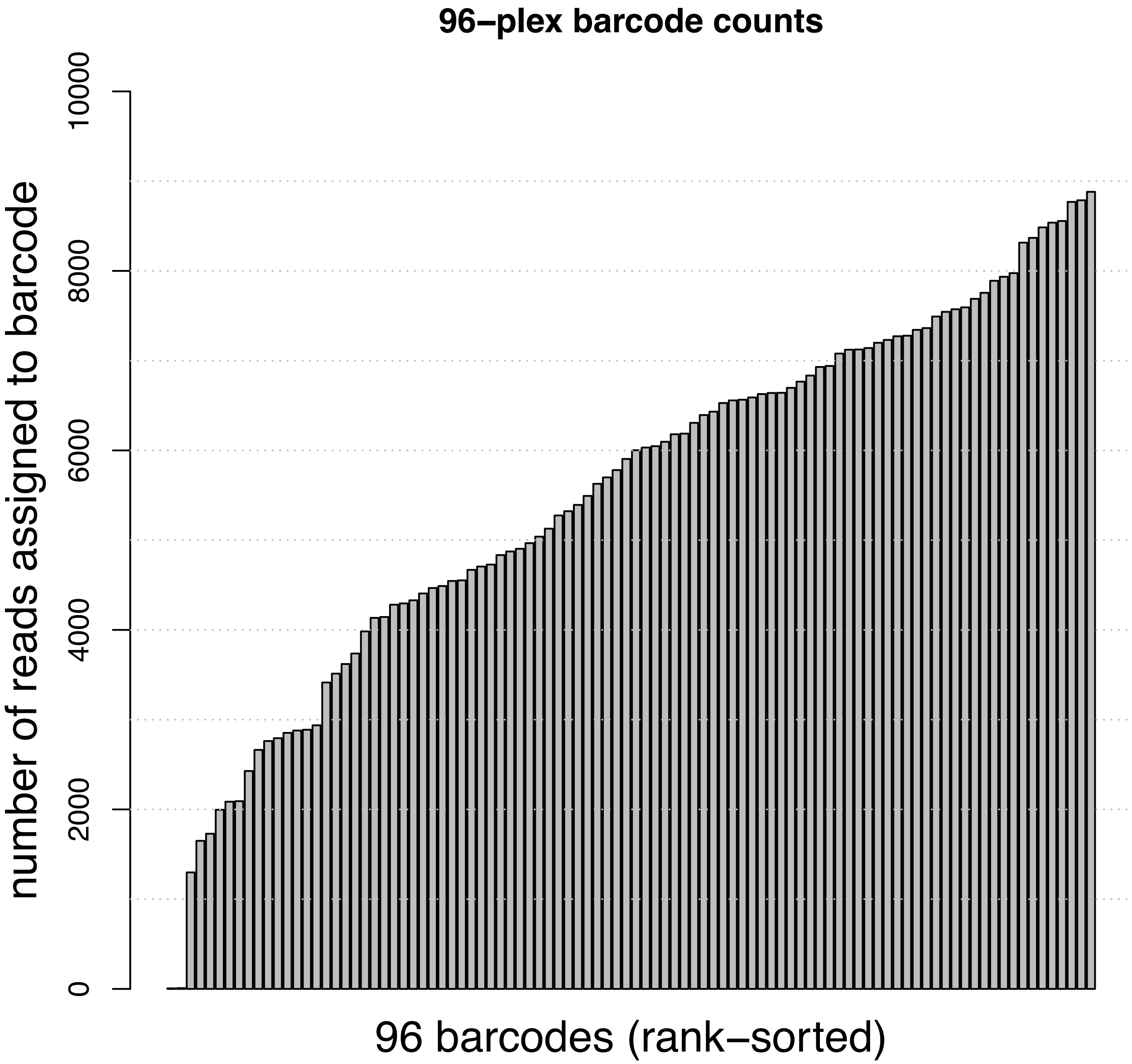
Cleaning step 1	int=t	int=t	int=f

bbduk.sh
- ref=phix.fasta
	input = 1,325,670
	contaminants = 0.01%
	total removed = 0.01%%
	output = 1,325,598
		input = 1,101,114
		contaminants = 0.00%
		total removed = 0.00%
		output = 1,101,086
			input = 1,222,468
			contaminants = 9.56%
			total removed = 9.56%
			output = 1,105,558

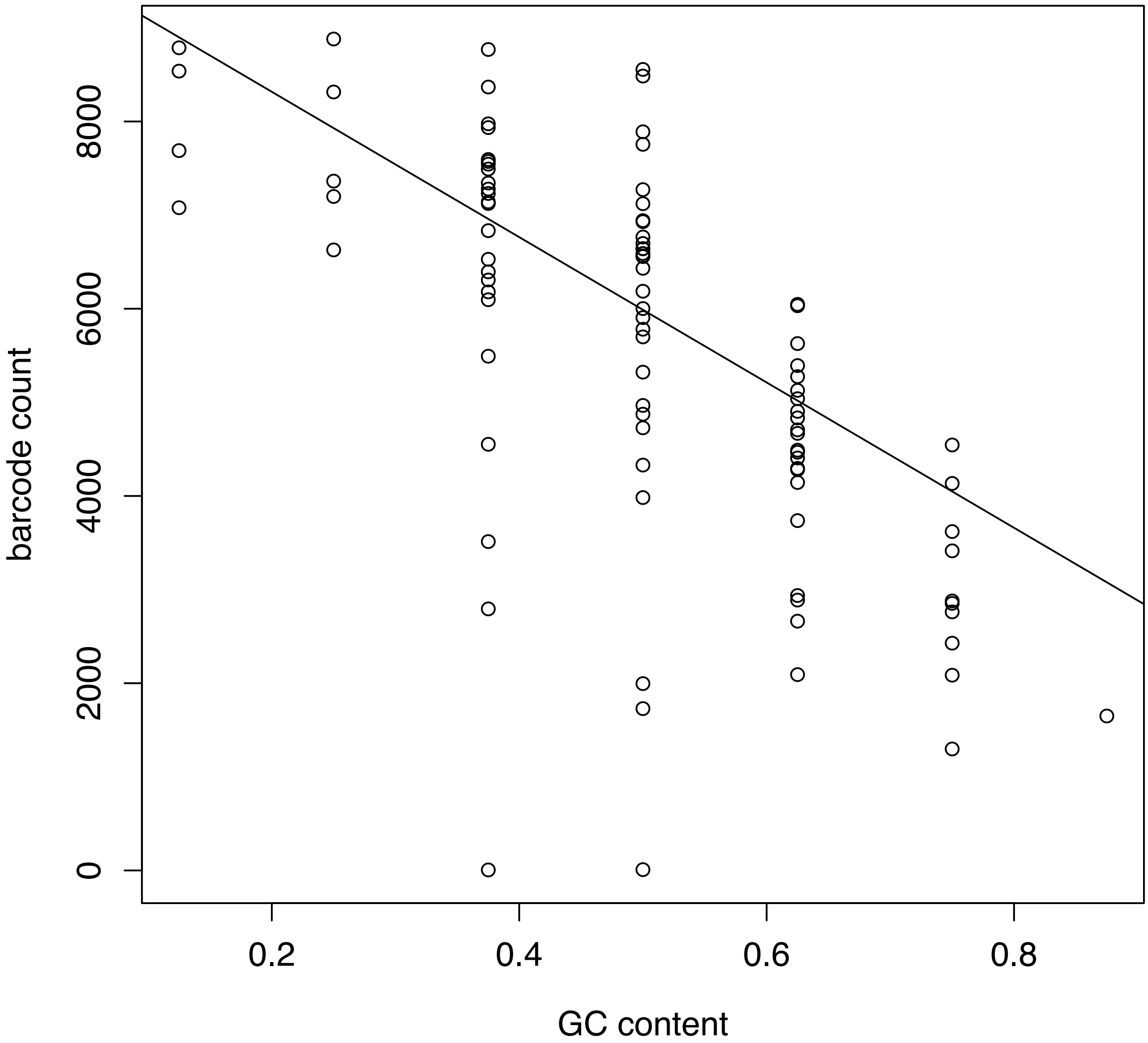
Cleaning step 2	int=t
	forcetrimright=150
		int=t
		forcetrimright=150
			int=f

bbduk.sh
- ref=adapters.fa
- maxgc=0.98
- trimq=20
- qtrim=r
-entropy=0.5
-minavgquality=25
	input = 1,325,598
	contaminants = 19.47%
	Qtrimmed = 28.27%
	Ftrimmed = 100%
	filtered by GC = 0%
	low quality = 21.60%
	low entropy = 0.30%
	total removed = 41.41%
	output = 776,728
		input = 1,101,086
		contaminants = 8.85%
		Qtrimmed = 36.44%
		Ftrimmed = 100%
		filtered by GC = 0%
		low quality = 27.85%
		low entropy = 0.51%
		total removed = 37.25%
		output = 690,930
			input = 1,105,558
			contaminants = 5.64%
			Qtrimmed = 21.40%
			filtered by GC = 0.48%
			low quality = 19.58%
			low entropy = 0.35%
			total removed = 25.97%
			output = 818,478

Supplementary Figure 4 (S4). Schematic overview of data analysis methods used in this study.



Supplementary Figure 5 (S5). Barcode indexes performance.
Number of barcodes assigned to each 8-bp index pair obtained with Hackflex; 97.9% (94 of 96) falling within a 6.8-fold range of relative abundance. Coefficient of variation 0.38 and 0.34, including and excluding the two outliers, respectively.

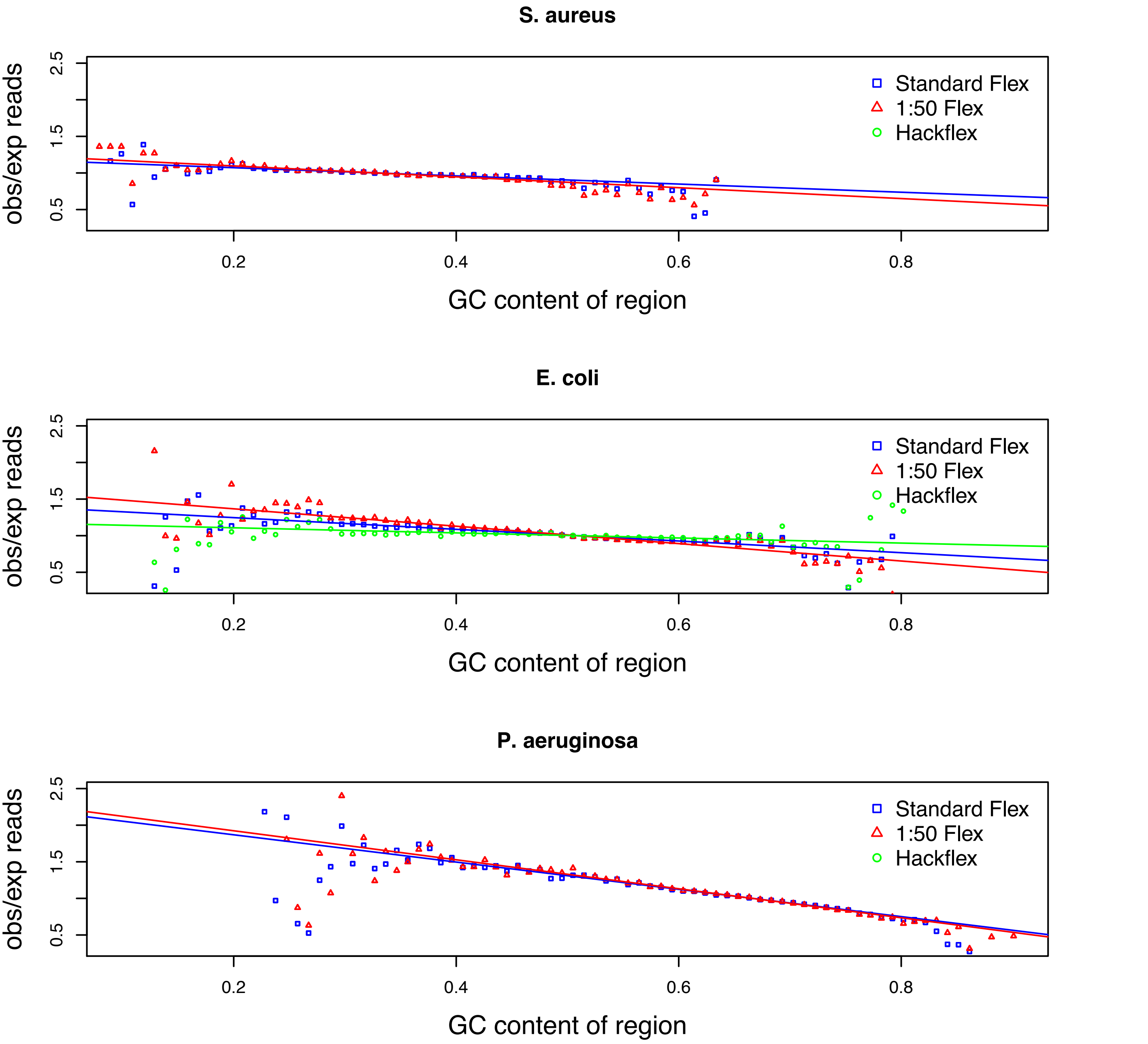


Supplementary Figure 6 (S6). Barcode GC content.
Weighted Pearson's correlation between GC content of 8-bp barcodes used in Hackflex and number of barcodes obtained per well (ρ=-0.704; p-value=1.3043e-16).

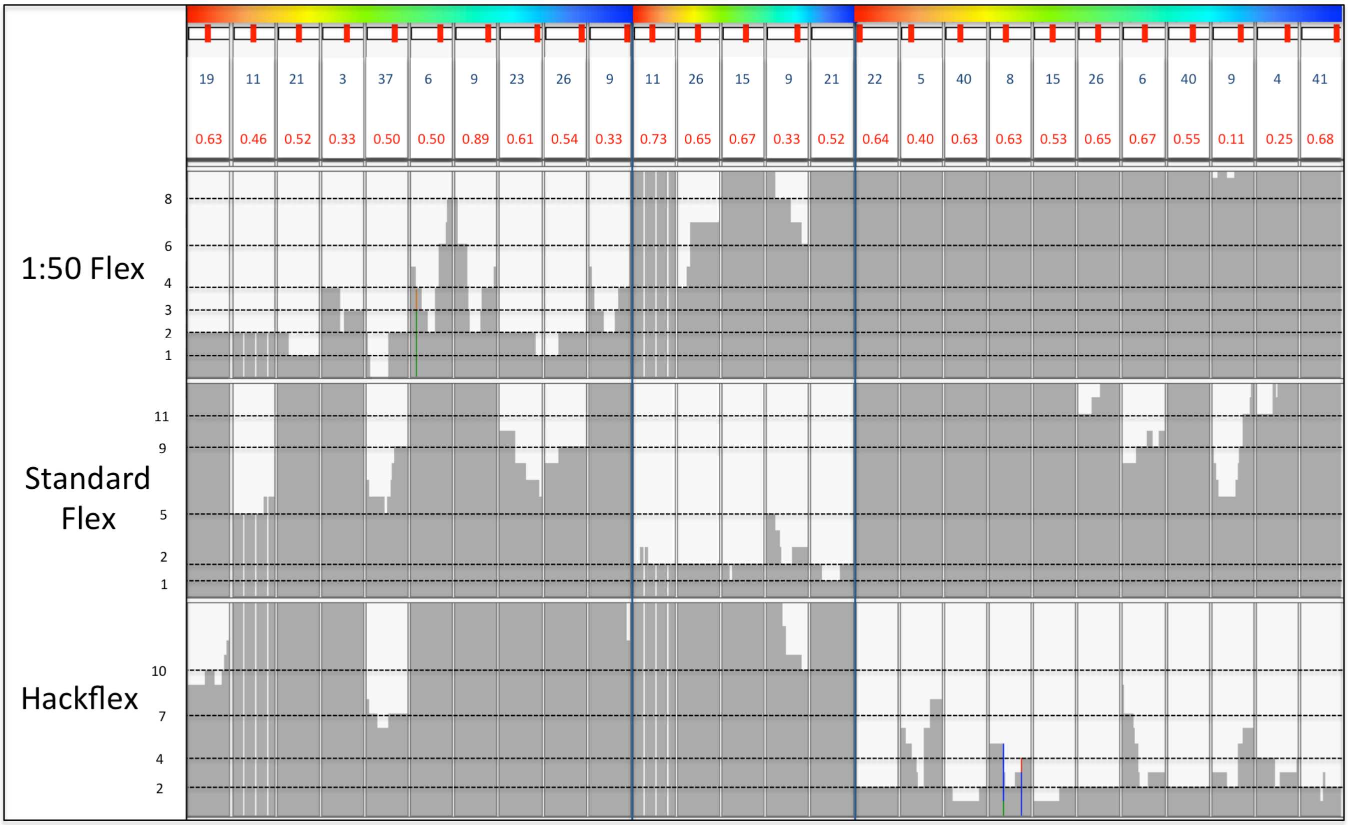
Supplementary Table 7 (S7).

Detailed coverage summary (output generated with ALFRED[13]).

	Standard Flex	1:50 Flex	Hackflex
#QCFail	0	0	0
QCFailFraction	0	0	0
#DuplicateMarked	0	0	0
DuplicateFraction	0	0	0
#Unmapped	60	75	1307
UnmappedFraction	8.76816e-05	1.08549e-04	1.94788e-03
#Mapped	684234	690855	669679
MappedFraction	0.999912	0.999891	0.998052
#MappedRead1	342111	345413	334652
#MappedRead2	342123	345442	335027
RatioMapped2vsMapped1	1.00004	1.00008	1.00112
#MappedForward	342218	345518	334913
MappedForwardFraction	0.500148	0.500131	0.500110
#MappedReverse	342016	345337	334766
MappedReverseFraction	0.499852	0.499869	0.499890
#SecondaryAlignments	0	0	0
SecondaryAlignmentFraction	0	0	0
#SupplementaryAlignments	874	870	889
SupplementaryAlignmentFraction	0.00127734	0.00125931	0.00132750
#SplicedAlignments	0	0	0
SplicedAlignmentFraction	0	0	0
#Pairs	342584	345900	335937
#MappedPairs	342542	345843	335083
MappedPairsFraction	0.999877	0.999835	0.997458
#MappedSameChr	341726	344772	333980
MappedSameChrFraction	0.997496	0.996739	0.994175
#MappedProperPair	341540	344579	333438
MappedProperFraction	0.996953	0.996181	0.992561
#ReferenceBp	4645181	4645181	4645181
#ReferenceNs	0	0	0
#AlignedBases	99193711	99631430	99421785
#MatchedBases	99111466	99544371	99360760
MatchRate	0.999171	0.999126	0.999386
#MismatchedBases	82245	87059	61025
MismatchRate	0.000829135	0.000873811	0.000613799
#DeletionsCigarD	1436	1259	1321
DeletionRate	1.44767e-05	1.26366e-05	1.32868e-05
HomopolymerContextDel	0.385794	0.337569	0.381529
#InsertionsCigarI	418	351	457
InsertionRate	4.21398e-06	3.52298e-06	4.59658e-06
HomopolymerContextIns	0.222488	0.173789	0.201313
#SoftClippedBases	63474	32332	28115
SoftClipRate	0.000639899	0.000324516	0.000282785
#HardClippedBases	0	0	0
HardClipRate	0	0	0
ErrorRate	0.001487730	0.001214490	0.000914468
MedianReadLength	151:151	151:151	151:151
DefaultLibraryLayout	2	2	2
MedianInsertSize	274	311	299
MedianCoverage	20	20	20
SDCoverage	23.6252	25.8771	27.6204
CoveredBp	4645181	4645140	4645181
FractionCovered	1.000000	0.999991	1.000000
BpCov1ToCovNRatio	8.39580e-06	8.39587e-06	1.24861e-05
BpCov1ToCov2Ratio	0.314516	0.351351	0.202091
MedianMAPQ	60	60	60



Supplementary Figure 9 (S9). GC coverage bias for additional libraries.
GC content bias assessed by calculating the GC content of the reference genome (102 bins) for *S. aureus* (top), *E. coli* MG1655 (center) and *P. aeruginosa* (bottom), and plotting the coverage obtained with Standard flex (blue square), 1:50 Flex (red triangle) and Hackflex (green circle). Weighted Pearson's correlation coefficients between GC content and coverage: (top) Standard Flex: ρ = -0.92, p-value: 6.20e-86; 1:50 Flex: ρ = -0.94; p-value: 3.85e-96; (center): Standard Flex: ρ = -0.95, p-value: 3.05e-104; 1:50 Flex: ρ = -0.96; p-value: 3.11e-112; Hackflex: ρ = -0.77; p-value: 3.43e-41; (bottom): Standard Flex: ρ = -0.99, p-value: 1.54e-181; 1:50 Flex: ρ = -0.99; p-value: 3.39e-175.



Supplementary Figure 10 (S10). Lowest coverage sites for Standard Flex, 1:50 Flex and Hackflex.

Numbers on the y-axis of each plot report the number of reads from a specific library, mapping against a specific genomic site. Coverage is shown at respective positions for the other libraries. Numbers on the top (blue) indicate the size of the low coverage site, whereas numbers on the bottom (red) indicate the GC fraction of the respective site. The color gradient line on the top indicates the start (red) and the end (blue) of the reference genome. (IGV-adapted figure)

Supplementary Table 11 (S11). Genomic positions of sites with lowest coverage for Standard Flex, 1:50 Flex and Hackflex.

Start and end positions of lowest coverage sites are reported for each library, together with the respective site length (nt) and GC content (%).

Library	Start position	End position	Low coverage site size (nt)	GC content (%)
1:50 Flex	1992408	1992427	19	63.2
	2003562	2003573	11	45.5
	2164696	2164717	21	52.4
	2489408	2489411	3	33.3
	2790154	2790191	37	45.9
	3011010	3011016	6	50.0
	3337143	3337152	9	88.9
	3663577	3663600	23	60.9
	3663627	3663653	26	53.8
	3768032	3768041	9	33.3
Standard Flex	1852510	1852521	11	72.7
	2003633	2003659	26	65.4
	2294000	2294015	15	66.7
	3104777	3104786	9	33.3
	4386000	4386021	21	52.4
Hackflex	269457	269479	22	63.6
	1021126	1021131	5	40.0
	1458255	1458295	40	62.5
	1646571	1646579	8	62.5
	1875073	1875088	15	53.3
	1956680	1956706	26	65.4
	2142707	2142713	6	66.7
	2498083	2498123	40	55.0
	2790274	2790283	9	11.1
	3118064	3118068	4	25.0
	3509476	3509517	41	68.3