Optimal enzymes for amplifying sequencing libraries

Michael A Quail, Thomas D Otto, Yong Gu, Simon R Harris, Thomas F Skelly, Jacqueline A McQuillan, Harold P Swerdlow & Samuel O Oyola

<table>
<thead>
<tr>
<th>Supplementary Figure 1</th>
<th>Genome browser screenshots of selected regions in four genomes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Figure 2</td>
<td>Evenness of coverage based on different library amplification conditions across four genomes.</td>
</tr>
<tr>
<td>Supplementary Figure 3</td>
<td>Genome browser screenshot of an AT-rich region of the human X chromosome.</td>
</tr>
<tr>
<td>Supplementary Table 1</td>
<td>Oligos used for Illumina library construction.</td>
</tr>
<tr>
<td>Supplementary Methods</td>
<td>Methods.</td>
</tr>
</tbody>
</table>

Note: Supplementary Tables 2–4 are available on the Nature Methods website.
Supplementary Figure 1. Genome browser screenshots of selected regions in four genomes.

Genome browser screenshots of selected regions in the genomes of: a. *B. pertussis* (GC-rich region); b. *S. pullorum*; c. *S. aureus* and d. *P. falciparum* (AT-rich var gene region of chromosome 11).

Libraries were prepared without PCR (green line), with 14 cycles of PCR using Phusion polymerase (blue line) and with 14 cycles of PCR using Kapa HiFi polymerase (purple line). In each window the top graph shows the percentage GC content at each position, with the numbers on the right denoting the minimum and maximum values. The middle graph in each window (purple, green and blue traces) is a coverage plot showing depth of reads (unnormalised) mapped at each position and below that are the coordinates of the selected region in the given genome.
Supplementary Figure 2. Evenness of coverage based on different library amplification conditions across four genomes.

a. *B. pertussis*

b. *S. pullorum*
c. *S. aureus*

![Graph showing evenness of coverage for S. aureus](image)

d. *P. falciparum*

![Graph showing evenness of coverage for P. falciparum](image)
Evenness of coverage Low Coverage Index (LCI) observed from different library amplification conditions across; a. *B. pertussis*; b. *S. pullorum*; c. *S. aureus* and d. *P. falciparum*, genomes. After initially testing a wide range of enzymes and conditions (Supplementary table 3) a subset of libraries, that included the best performing enzymes and conditions, were repeated and run on both Illumina GAIIx and HiSeq platforms. All data sets were randomly normalised to 10x coverage by taking the first number of reads representing that coverage from the output fastq file. Here the average evenness of coverage metric (LCI 0.5) across all 3 runs is plotted. We use LCI as standard deviation measurements can be heavily biased by the coverage situation close to the average depth such that problematic gaps and low-covered regions are not truly reflected in the standard deviation value. The Low Coverage Index (LCI) best reflects the situation of low coverage of sequencing reads across the genome. Mathematically the value of LCI can be viewed as a weighted average of proportions of bases at different levels of low coverage (see Supplementary Note 1). It gives more weight to lower coverage levels.

Conditions are ranked with the library giving the lowest LCI (0.5) value is on the left and the conditions giving the highest value and hence the most uneven coverage on the right. Error bars show the observed variation across the three replicate datasets. All libraries were multiplexed. 16-20 libraries were run per flowcell lane and all four genome libraries for a particular enzyme/condition were kept together. In the second GAIIx and Hiseq runs all samples were run on one flowcell and barcodes used to identify particular genomes/enzymes were changed from those used during the first run to eliminate any bias that might be introduced in the multiplexing process.
Supplementary Figure 3. Genome browser screenshot of an AT-rich region of the human X chromosome.

Genome browser screenshot of an AT-rich region of the human X chromosome. Libraries were prepared without PCR (green line), with 14 cycles of PCR using Phusion polymerase (red line) and with 14 cycles of PCR using Kapa HiFi polymerase (blue line). Each library was run in a single Illumina GAIIx lane and yielded 2 to 3 x average coverage. Data was mapped against build 37 of the human genome. The top graph shows the percentage GC content at each position, with the numbers on the right denoting the minimum and maximum values. The middle graph in each window (red, green and blue traces) is a coverage plot showing depth of reads (unnormalised) mapped at each position and below that are the coordinates of the selected region in the given genome. Coverage with the phusion polymerase amplified library repeated falls to zero in regions close to AT-rich sequences whereas coverage from libraries prepared without PCR and with Kapa HiFi does not.
**Supplementary Table 1**

Oligos used for Illumina library construction.

Note: * indicates phosphorothioate. All oligos were PAGE purified.

**PE adapter**

PEad_top

5' ACACTCTTTCCCTACACGACGCTCTTCCGATC*T

3'

PEad_bottom

5' P-GATCGGAAGACGCGGTTCAGCAGGAATGCCGA*A

3'

iPCR index read sequencing primer

7' AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC

**PE1.0**

7' AATGATACGGCGACACGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T

**Modified multiplexing PE2.0 oligos**

<table>
<thead>
<tr>
<th>Oligo name</th>
<th>Single correcting, double &amp; shift detecting octamers</th>
<th>Sequence obtained</th>
<th>PCR primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPCRtagT1</td>
<td>AACGTGAT</td>
<td>ATACACGTAT</td>
<td>5' CAAGCAGAAAGCGCATACAGGATACAGTGAATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT2</td>
<td>AAACATCG</td>
<td>CGATGTTTAT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT3</td>
<td>ATGCTTAA</td>
<td>TTAGGCAT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT4</td>
<td>AGTGTCGA</td>
<td>GACCACACT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT5</td>
<td>ACCACTGT</td>
<td>ACAGTGAT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT6</td>
<td>ACATTGGGC</td>
<td>GCCAATGT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT7</td>
<td>CAGATCG</td>
<td>CAGACTCT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT8</td>
<td>CATCAAGT</td>
<td>ACTTGAT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT9</td>
<td>CGCTGATC</td>
<td>GATCGAGG</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT10</td>
<td>ACAAGCTA</td>
<td>TAGCTTG</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT11</td>
<td>CTGTAGCC</td>
<td>GGCTACAG</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT12</td>
<td>AGTACAGG</td>
<td>CTTTGACT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT13</td>
<td>ACAACACAA</td>
<td>TGTTGTTT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT14</td>
<td>ACACAGAA</td>
<td>TCTCGGTGT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT15</td>
<td>ACCGCTTA</td>
<td>TAAAGGTG</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT16</td>
<td>AAGGACGGGA</td>
<td>TCCGCTTTT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT17</td>
<td>AAGGTACA</td>
<td>TGCTCTTTT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT18</td>
<td>ACACAGAA</td>
<td>TTCTGCTGT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
<tr>
<td>iPCRtagT19</td>
<td>ACAGCAGA</td>
<td>TCTGCTGT</td>
<td>5' CAAGCAGAAAGCGCATACGAGATAAACACGAGATCTCGCAGTAACGCCTCTCCGATC*T 3'</td>
</tr>
</tbody>
</table>
noPCR adapter

T_no_PCR
5’ AATGATACGGCGACACGGGTCTACACTCTTTCCCTACAGCTCTTGATGTTCGCCTCGATC'T 3’

B_no_PCR
5’ P-GATCGGAAGAGCGGTTCGACAGCTCTTCCGAGCTTGCCTTACTTCACACGGACGTATCGTATGACC

iPCRtagT20  ACCTCAA  TTGGAGGTAT  5’ CAAGCAGAAGACGGCATACGAGATACCTCCAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC’T 3’
iPCRtagT21  ACGTCTGA  TCGACGGAT  5’ CAAGCAGAAGACGGCATACGAGATACCTCCAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC’T 3’
iPCRtagT22  ACGTATCA  TGATACTGAT  5’ CAAGCAGAAGACGGCATACGAGATACCTCCAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC’T 3’
iPCRtagT23  ACTATGCA  TGATAGTAT  5’ CAAGCAGAAGACGGCATACGAGATACCTCCAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC’T 3’
iPCRtagT24  AGAGTCAA  TTGACTCTAT  5’ CAAGCAGAAGACGGCATACGAGATACCTCCAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC’T 3’
Supplementary Methods

Library construction

DNA (5 µg in 120 µl of 10 mM Tris.HCl pH8.5) from each test genome (*Bordetella pertussis* ST24, *Salmonella pullorum* S449/87, *Staphylococcus aureus* TW20 and *Plasmodium falciparum* 3D7) was sheared in an AFA microtube using a Covaris S2 device (Covaris Inc.) with the following settings: Duty cycle 20, Intensity 5, cycles/burst 200, 45 seconds.

Sheared DNA was purified by binding to an equal volume of Ampure beads (Beckman Coulter Inc.) that had been reconstituted in 16 % PEG 6000, 1.8 M NaCl, and eluted in 32 µl of 10 mM Tris-HCl, pH8.5. End-repair, A-tailing and paired end adapter ligation were performed (as per the protocols supplied by Illumina, Inc. using reagents from New England Biolabs- NEB,) with purification using a 1.5:1 ratio of standard Ampure to sample between each enzymatic reaction. PCR free libraries were constructed according to Kozarewa et al.\(^1\).

After ligation, excess adapters and adapter dimers were removed using two Ampure clean-ups, first with a 1.5:1 ratio of standard Ampure to sample, followed by a 1.2:1 ratio of Ampure beads reconstituted in 15.6 % PEG 6000, 1.8 M NaCl. PCR free libraries were then used as is. Other libraries ligated with standard Illumina paired-end adapters were diluted to 2 ng/µl and 1 µl was used as template for PCR amplification using a variety of test conditions and alternative enzymes as listed in Supplementary Table 2. Generally each enzyme was used with its supplied buffer and the manufacturer’s recommendations for denaturation, annealing and extension times and temperatures were followed. All PCR reactions were performed in 0.2 µl thin wall microtubes on an MJ tetrad thermal cycler with 1x buffer and 200 nM final concentration of standard PE1.0 and modified multiplexing PE2.0 primers (Supplementary Table 1). After PCR, excess primers and any primer dimer were removed using two Ampure clean-ups, first with a 1.5:1 ratio of standard Ampure to sample then with a 1.2:1 ratio of Ampure beads reconstituted in 15.6 % PEG 6000, 1.8 M NaCl.

All libraries were quantified by real time PCR using the SYBR fast Illumina library quantification kit (Kapa Biosystems) and pooled so as to give equal genome coverage from each library.

Typically libraries were multiplexed in sets of 16-24 per lane with the four test genomes amplified under the same conditions always being kept together in a single lane. Each multiplexed library pool was sequenced on an Illumina GAIIx instrument for 76 cycles from each end plus an 8 bp-index sequence read.

Design of multiplexing oligos

Unique sequence tags allowing library multiplexing via PCR were introduced into the central portion of the adapter between the R2 sequencing primer and P7 sequences and sequenced using a short third sequencing read using a primer that is the reverse complement of the read 2 primer. The index sequence of 8bp was designed such that deconvolution would still be possible if two errors were introduced during sequencing or if the sequence slipped one base in either direction due to an insertion or deletion. The oligo sequences used are presented in Supplementary Table 2.
Data processing (see Figure A)

After sequencing, reads were mapped to each genome reference sequence using BWA\(^2\). SAMtools\(^3\) was used to generate coverage data from the pileup mapping output. Each genome dataset was normalized to 10 \(x\) coverage.

In addition to the inbuilt Illumina pipeline quality metric procedures, we developed analysis metrics to compare the quality of sequence data generated under each set of conditions. Our analysis metrics assessed three aspects of data quality (Figure A):

1) Genome coverage - to assess representation of extreme base composition loci focusing on selected genomic regions; 2) evenness of coverage metrics - comparing the overall representation and depth across the entire genome; and 3) fidelity metrics - assessing enzyme dependent errors.

All datasets have been deposited in the ENA read archive under accession number ERP000804.

Genome coverage

We counted the number of bases in the genome that were not covered at all by any reads (Coverage=0) and those with less than 5 \(x\) read coverage (Coverage < 5\(x\)). SAMtools was used to generate coverage plots and bash/awk scripts were used for coverage counting.

Evenness of coverage metrics

We extracted genome coverage information from the pileup data derived by SAMtools from mapped reads after normalizing to a uniform depth of 10\(x\). Evaluation of evenness of coverage was based on cumulative distributions over the normalized overall average depth. A measurement of low-coverage index lci \((d)\) is defined as the integration of the cumulative coverage distribution \(C(x)\) from 0 to \(d\) to give an overall assessment of the coverage at the low end of distribution:

\[
\text{lci}(d) = \int_{0}^{d} C(x)dx
\]

The value lci (0.5) that gives a measurement of the coverage below one half of the average depth in the distribution was used to compare evenness of coverage for each data set.

Enzyme-dependent fidelity

Enzyme-dependent fidelity metrics assess the possibility of errors caused by an amplification enzyme. These errors are differentiated from sequencing errors as the base
quality at the selected region must be high. We excluded the first 5 and the last 15 bases of each read as the Illumina technology tends to produce fewer correct base calls in these regions. We call a fidelity error in a read if; 1) a base and its four neighboring bases on each side have a quality higher than or equal to 30 (≥Q30), 2) it is not a known variant and 3) 94% of all the enzymes tested have at least one high-quality read for this base. The total number of errors per enzyme was counted and a normalized error score was generated (Figure A).

**Figure A:** Overview of Analysis Pipeline.

The different colors indicate the specific analyses performed.
Rank

We ranked the results for each dataset. Supplementary table 3 shows the results for the Low Coverage Index (LCI 0.5) evenness coverage score, defined above. Supplementary Table 4 shows the results for genome coverage at 0x and < 5x as well as the fidelity score by assigning the best a score of 1 and the worst a score of zero. Intermediate results were ranked on a pro rata basis.
