

## NATURE METHODS | CORRESPONDENCE

# 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

*Nature Methods* **9**, 10–11 (2012) doi:10.1038/nmeth.1814

Published online 28 December 2011

**Subject terms:** Genomics Sequencing Bioinformatics Molecular Biology

To the Editor:

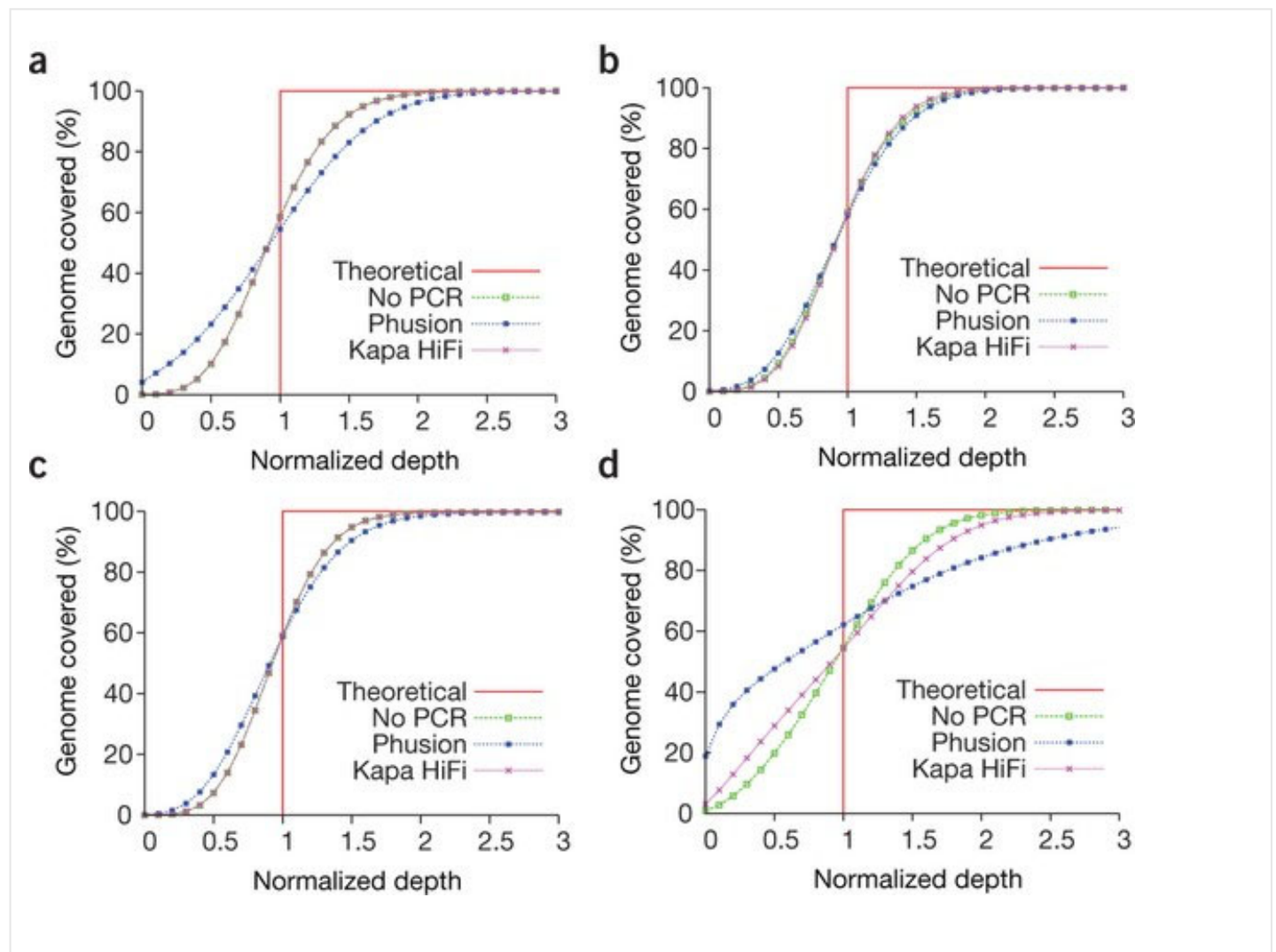
PCR amplification introduces bias into Illumina sequencing libraries<sup>1</sup>. Although amplification-free library preparation solves this, micrograms of starting material are usually required. Most researchers follow standard protocols using Phusion polymerase, which has processivity and fidelity advantages over most polymerases. Yet for genomics applications, our demands on DNA amplification systems often surpass their specification. Thermostable DNA polymerases such as Phusion are used to amplify mixtures of fragments, albeit with variable efficiency. Typically, (G+C)-neutral fragments are amplified with higher efficiency than extremely (G+C)-rich or (A+T)-rich fragments. The accumulation of these slight differences in amplification over multiple cycles often results in profound bias. There have been reports of using alternative DNA polymerases for Illumina library construction<sup>2, 3, 4</sup>, but these are infrequent, and comprehensive analyses are lacking. To reduce bias, we investigated many thermostable DNA polymerases and alternate reaction conditions for amplification of adapter-ligated fragments for Illumina sequencing. We expect this comparison to be relevant to other applications that involve simultaneous amplification of complex fragment mixtures.

To assess amplification efficiency across a comprehensive range of sequence contexts we made four libraries from microbial genomes with differing G+C content: 67.7% in *Bordetella pertussis*, 52% in *Salmonella pullorum*, 33% in *Staphylococcus aureus* and 19.3% in *Plasmodium falciparum*. For each enzyme and condition we used 2 nanograms of unamplified genomic fragments and 14 cycles of PCR (Supplementary Methods and Supplementary Tables 1 and 2). We indexed and ran the libraries on an Illumina Genome Analyzer IIx to give >10× coverage of each genome. For a fair comparison, we randomly trimmed datasets to contain reads representing 10× genome-wide coverage. We tabulated the depth of coverage observed at each position of the genome and calculated the fraction of each genome that was covered to a depth of less than 5× (Supplementary Table 3), ranking each dataset according to its performance and calculating a combined rank for each enzyme across all four genomes. To ensure reliable performance, we repeated the experiment using a subset of the top-ranking enzymes and conditions on both Illumina GAIIx and HiSeq2000. Finally, we reanalyzed data from all these runs and ranked each dataset according to its performance with respect to genome coverage and fidelity (Supplementary Table 4).

Libraries prepared without PCR amplification<sup>1</sup> performed best. Among the amplified libraries there were big differences, especially for the (G+C)-rich *B. pertussis* and (A+T)-rich *P. falciparum* genomes. The best enzyme overall for Illumina library preparation was Kapa HiFi (Kapa Biosystems), which performed well using either standard amplification, a quantitative PCR premix formulation or with annealing and extension at 60 °C. Genome coverage using Kapa HiFi was

far more uniform than that with Phusion, with the former performing remarkably close to results achieved without PCR (Fig. 1). Whereas the fidelity (accuracy) of Kapa HiFi was similar to that of Phusion in the regions amplified by both enzymes, Kapa HiFi had a higher overall error rate. This is because Kapa HiFi makes mistakes in regions that are very difficult or impossible for Phusion and other enzymes to amplify. We detected a small number of short insertions and deletions (indels; approximately three per million base pairs) in regions of the *P. falciparum* genome rich in TA repeats that only Kapa HiFi can amplify. We observed no increase in indels or substitutions for Kapa HiFi in the other genomes. Although notable, this does not present an appreciable problem because the indels are confined to single reads. Particularly in (A+T)- and (G+C)-rich regions, the coverage observed with Phusion-prepared libraries (and many of the other enzymes tested) fell to zero (Supplementary Fig. 1). In the same regions, coverage in libraries prepared without PCR amplification or using Kapa HiFi was often improved. In the more (G+C)-neutral genomes of *S. pullorum* and *S. aureus*, differences between one enzyme and the next best were small (though libraries prepared using either Kapa HiFi or without PCR exhibited more even coverage). Although Kapa HiFi performed the best overall, some of the other enzymes and conditions tested performed slightly better in individual situations (Supplementary Fig. 2; for example, TopoTaq HF for (G+C)-neutral genomes). To investigate whether our results had a beneficial effect on sequencing the human genome, we constructed libraries without PCR and with Kapa HiFi or Phusion polymerase amplification (Supplementary Fig. 3) and noted improved sequence coverage using Kapa HiFi, particularly over (A+T)-rich loci.

**Figure 1: Genome coverage uniformity plots for 10× Illumina sequence coverage.**



(a–d) Coverage of *B. pertussis* (a), *S. pullorum* (b), *S. aureus* (c) and *P. falciparum* (d) prepared without PCR (no PCR) or with 14 cycles of PCR using Phusion polymerase or Kapa HiFi polymerase. The percentage of the genome covered is plotted against the normalized cumulative depth of genome covered. Ideal coverage behavior (theoretical) is when all of the genome is equally covered at or above the average coverage depth. The closer observed coverage is to ideal coverage, the more uniform the coverage is in that dataset. The Kapa HiFi data in a–c are hidden behind the 'no PCR' data (same coverage uniformity).

In summary, we identified optimal enzymes for amplifying high complexity mixtures of DNA fragments. We expect that improvements from these high-performance enzymes will facilitate more complete analyses of a wide range of genomes using Illumina sequencing platforms and should apply to any other sequencing technology that relies on amplification.

## References

1. Kozarewa, I. *et al. Nat. Methods* **6**, 291–295 (2009).
2. Quail, M. *et al. Nat. Methods* **5**, 1005–1010 (2008).
3. Aird, D. *et al. Genome Biol.* **12**, R18 (2011).
4. Fisher, S. *et al. Genome Biol.* **12**, R1 (2011).

[Download references](#)

## Author information

### Affiliations

**Wellcome Trust Sanger Institute, Hinxton, UK.**

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

### Competing financial interests

The authors declare no competing financial interests.

### Corresponding author

Correspondence to: Michael A Quail

## Supplementary information

### PDF files

1. Supplementary Text and Figures (2.2M)  
Supplementary Figures 1–3, Supplementary Table 1, Supplementary Methods

### Excel files

1. Supplementary Table 2 (25K)  
Enzymes and conditions used for amplification step in Illumina library construction.
2. Supplementary Table 3 (29K)  
Four-genome alternative enzyme study. Rank order.
3. Supplementary Table 4 (119K)  
Four-genome alternative enzyme study. Performance ranking based on coverage and fidelity.

**Nature Methods** ISSN 1548-7091 EISSN 1548-7105

© 2012 Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.  
partner of AGORA, HINARI, OARE, INASP, ORCID, CrossRef and COUNTER

Nature Methods

## **Optimal enzymes for amplifying sequencing libraries**

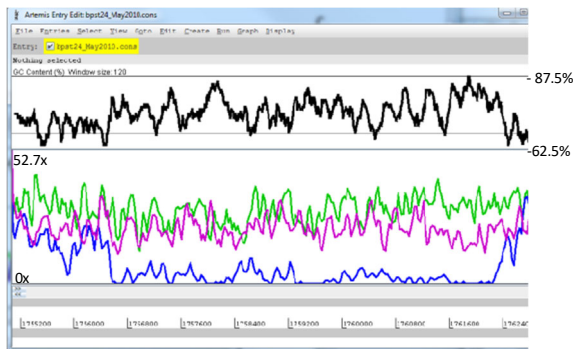
Michael A Quail, Thomas D Otto, Yong Gu, Simon R Harris, Thomas F Skelly,  
Jacqueline A McQuillan, Harold P Swerdlow & Samuel O Oyola

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

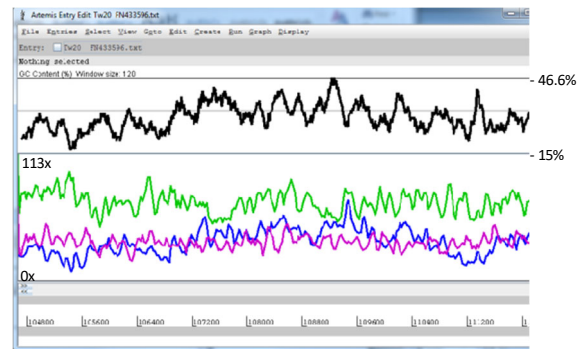
*Note: Supplementary Tables 2–4 are available on the Nature Methods website.*

**Supplementary Figure 1.** Genome browser screenshots of selected regions in four genomes.

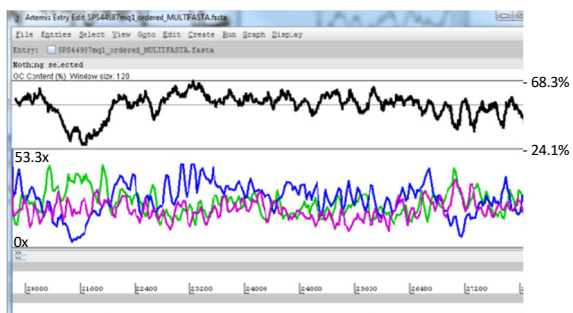
a.



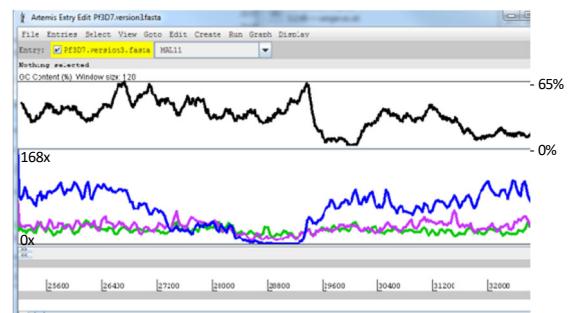
c.



b.



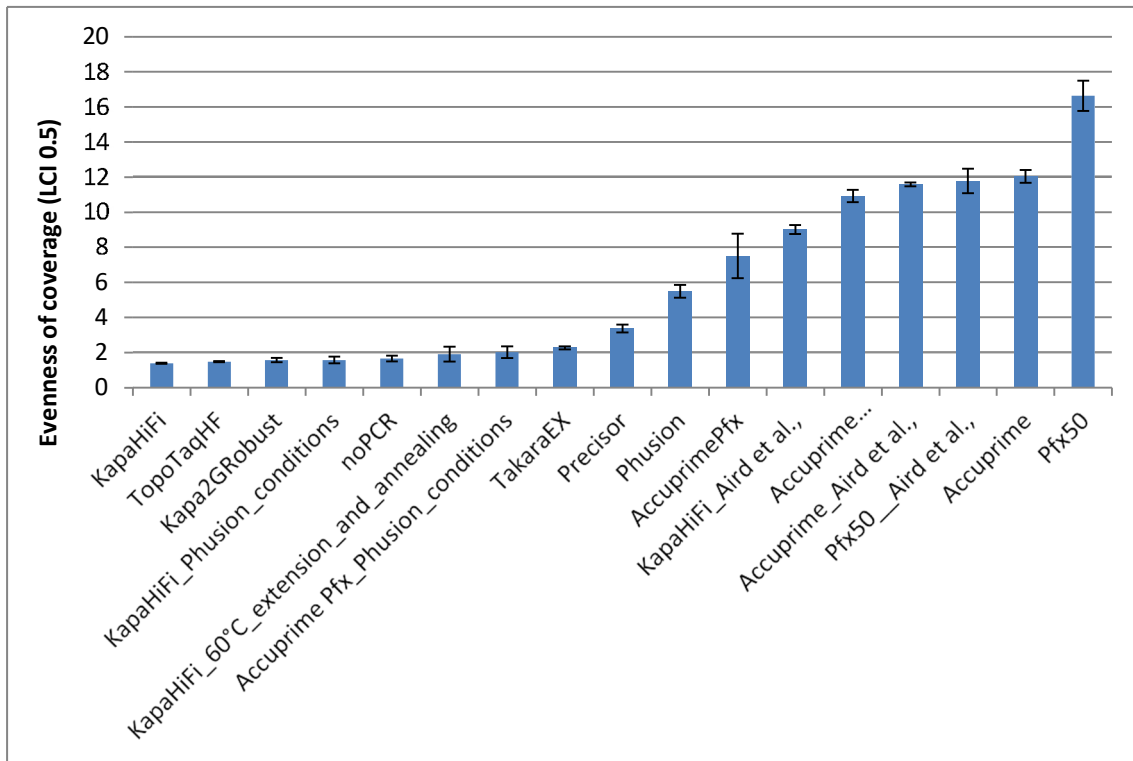
d.



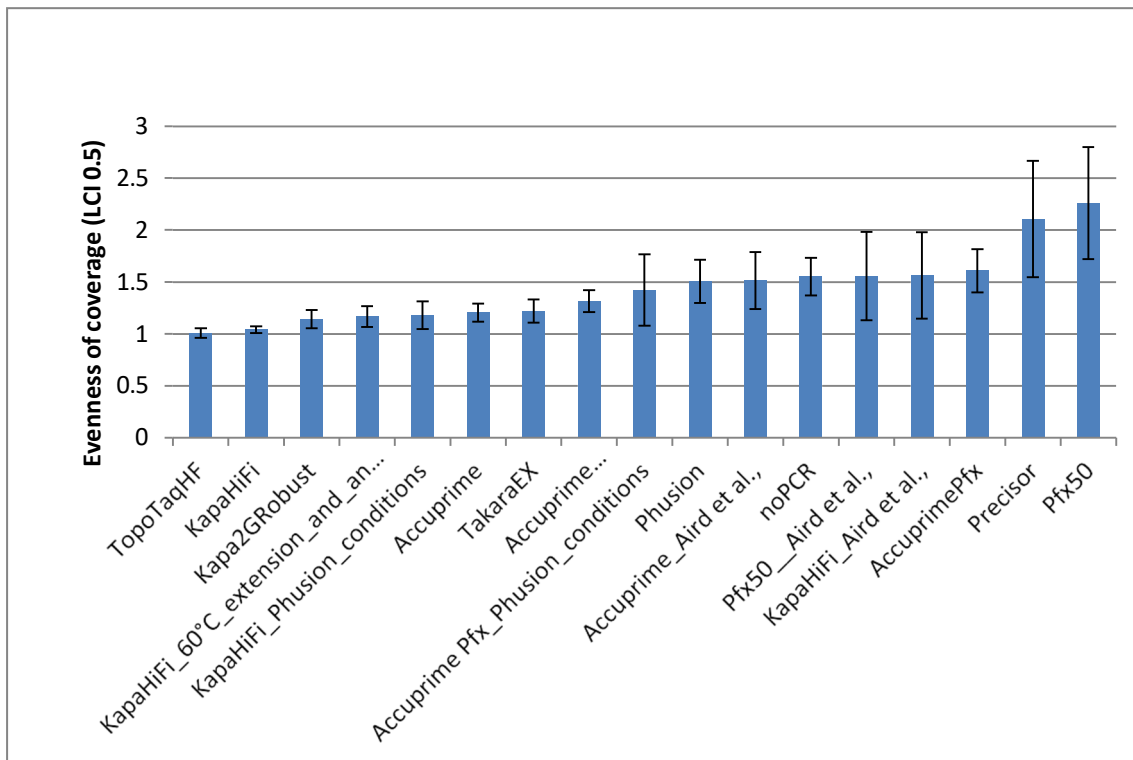
Genome browser screenshots of selected regions in the genomes of: **a.** *B. pertussis* (GC-rich region); **b.** *S. pullorum*; **c.** *S. aureus* and **d.** *P. falciparum* (AT-rich *var* gene region of chromosome 11). Libraries were prepared without PCR (green line), with 14 cycles of PCR using Phusion 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.

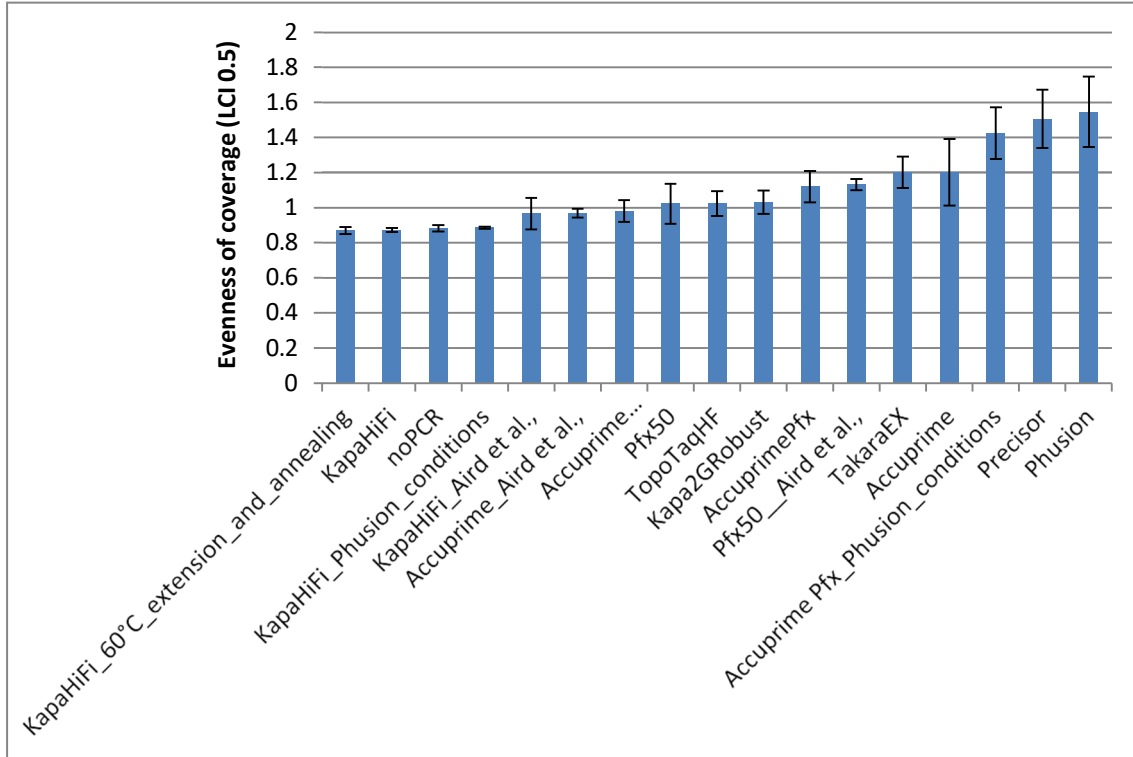
**a. *B. pertussis***



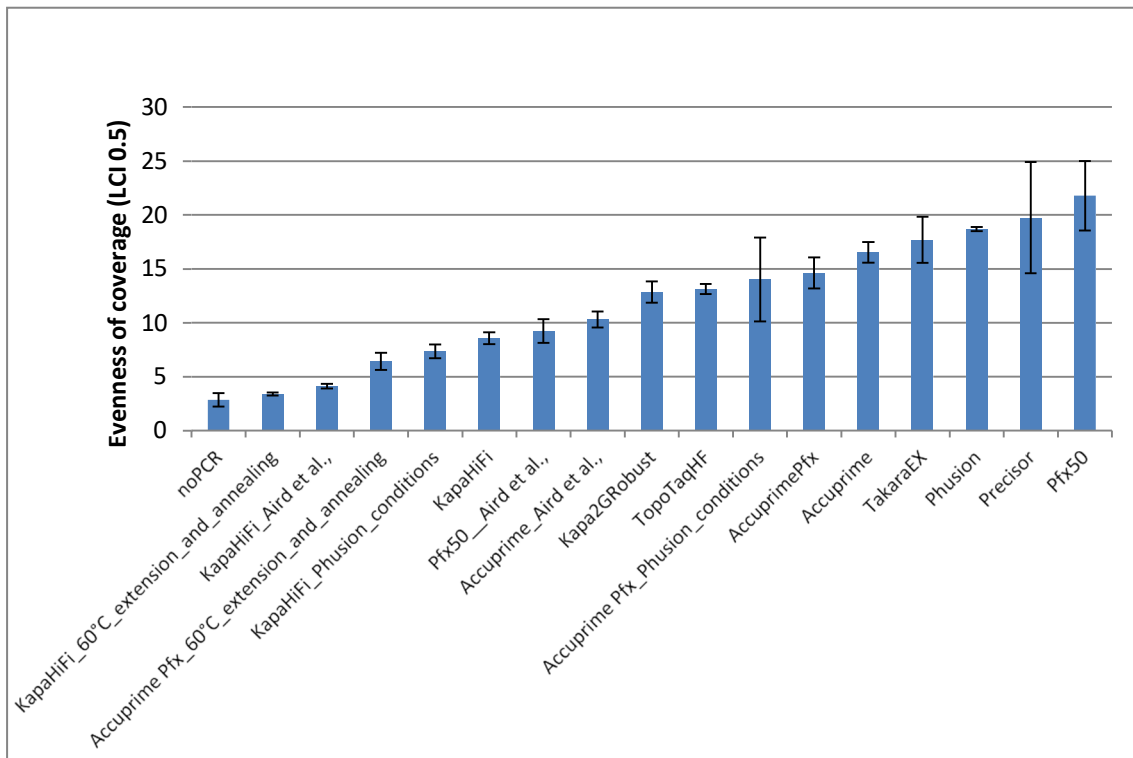
**b. *S. pullorum***



c. *S. aureus*



d. *P. falciparum*

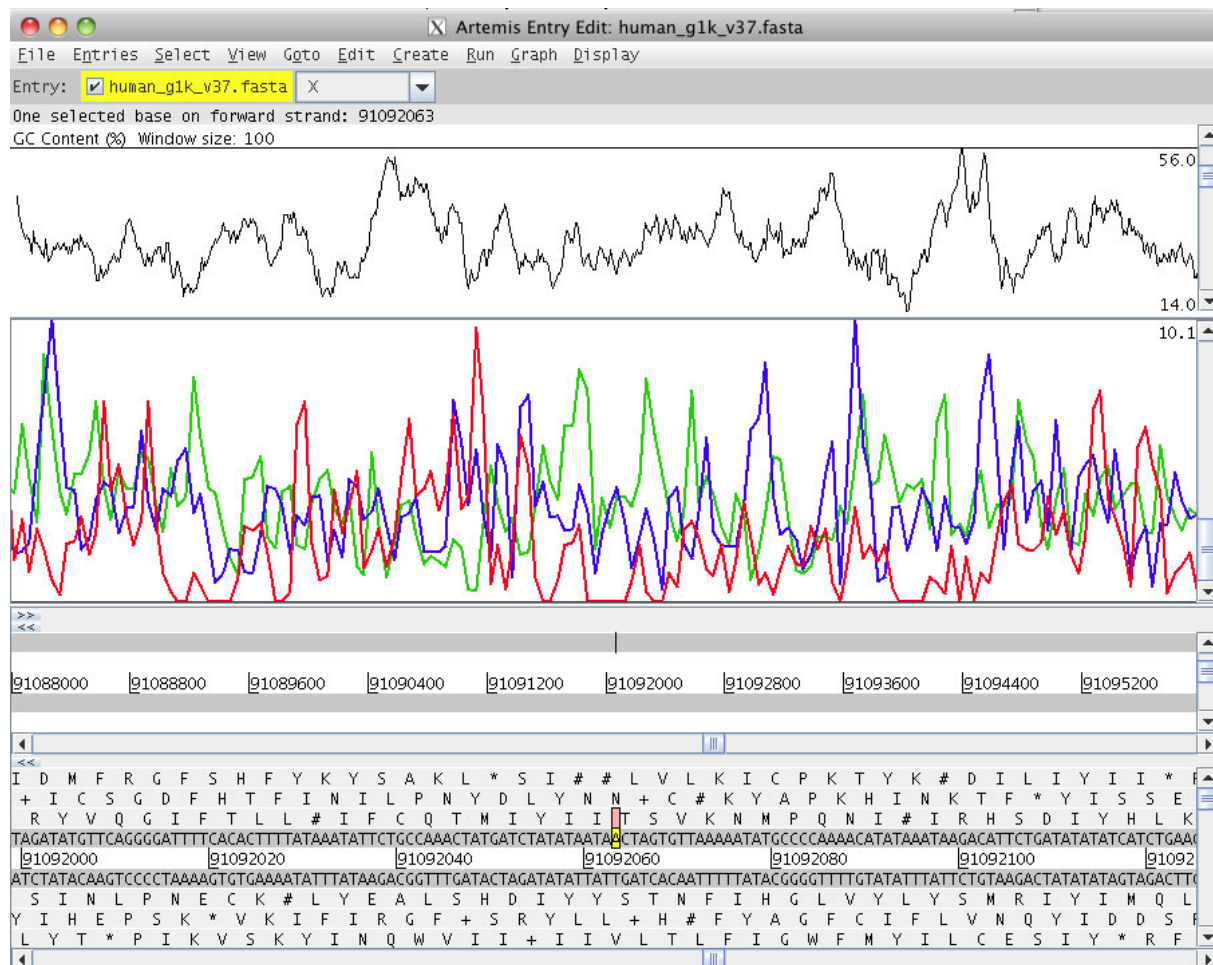




Evenness of coverage Low Coverage Index (LCI) observed from different library amplification conditions across; a. *B. pertussis*; b. *S. pullorum*; c. *S. aureus* and d. *P. falciparum*, genomes. After initially testing a wide range of enzymes and conditions (Supplementary table 3) a subset of libraries, 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' AACTCTTTCCCTACACGACGCTCTTCCGATC\*T 3'

##### PEad\_bottom

5' P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGA\*G 3'

#### iPCR index read sequencing primer

5' AAGAGCGGTTACGACGGAATGCCGAGACCGATCTC 3'

#### PE1.0

5' AATGATA<sup>\*</sup>CGGCGACCACCGAGATCTACTCTTTCCCTACACGACGCTCTTCCGATC\*T 3'

#### Modified multiplexing PE2.0 oligos

Oligo name	Single correcting, double & shift detecting octamers	Sequence obtained	PCR primers
iPCRtagT1	AACGTGAT	ATCACGTTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAACGTGATGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT2	AAACATCG	CGATGTTTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAAACATCGGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT3	ATGCCTAA	TTAGGCATAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATATGCCTAAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT4	AGTGGTCA	TGACCACTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAGTGGTCAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT5	ACCACTGT	ACAGTGGTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATA <sup>*</sup> CCACTGTGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT6	ACATTGGC	GCCAATGTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATA <sup>*</sup> ATTGGCAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT7	CAGATCTG	CAGATCTGAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATCAGATCTGGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT8	CATCAAGT	ACTTGATGAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATCATCAAGTGAAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT9	CGCTGATC	GATCAGCGAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATCGCTGATCGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT10	ACAAGCTA	TAGCTTGTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATA <sup>*</sup> CAAGCTAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT11	CTGTAGCC	GGCTACAGAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATCTGTAGCCGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT12	AGTACAAG	CTTGACTACT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAGTACAAGGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT13	AACAACCA	TGGTTGTTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAAACAACAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT14	AACCGAGA	TCTCGGTTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAACCGAGAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT15	AACGCTTA	TAAGCGTTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAACGCTTAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT16	AAGACGGA	TCCGTCTTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAAAGCGGAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT17	AAGGTACA	TGTACCTTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATAAGGTACAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT18	ACACAGAA	TTCTGTGTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATA <sup>*</sup> ACACAGAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT19	ACAGCAGA	TCTGCTGTAT	5' CAAGCAGAAGACGGCATA <sup>*</sup> CGAGATA <sup>*</sup> ACAGCAGAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T 3'

iPCRtagT20	ACCTCCAA	TTGGAGGTAT	5' CAAGCAGAAGACGGCATAACGAGATACTCCAAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTCCGATC*T 3'
iPCRtagT21	ACGCTCGA	TCGAGCGTAT	5' CAAGCAGAAGACGGCATAACGAGATAACGCTCGAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTCCGATC*T 3'
iPCRtagT22	ACGTATCA	TGATACGTAT	5' CAAGCAGAAGACGGCATAACGAGATAAGTATCAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTCCGATC*T 3'
iPCRtagT23	ACTATGCA	TGCATAGTAT	5' CAAGCAGAAGACGGCATAACGAGATACTATGCAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTCCGATC*T 3'
iPCRtagT24	AGAGTCAA	TTGACTCTAT	5' CAAGCAGAAGACGGCATAACGAGATAGAGTCAAGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTCCGATC*T 3'

### noPCR adapter

#### T\_no\_PCR

5' AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTCCGATC\*T 3'

#### B\_no\_PCR

5' P-GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTT\*G 3'

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/657161063103006025>