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:

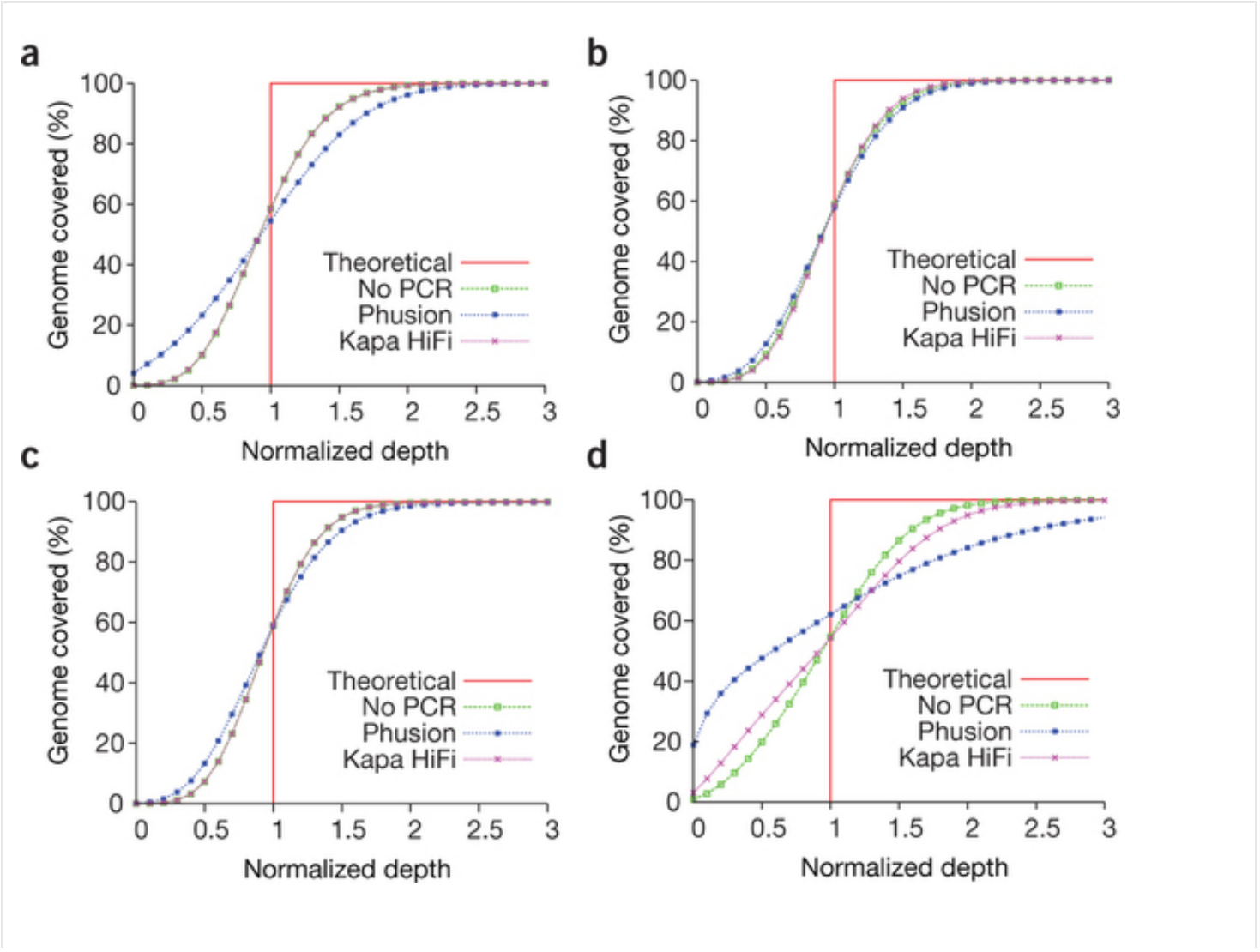
PCR amplification introduces bias into Illumina sequencing libraries[1]. Although amplification-free library preparation solves this, micrograms of starting material are usually required. Most researchers follow standard protocols using Phusion polymerase, which has processivity and fidelity advantages 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[2, 3, 4], 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[1] 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 quantitiative 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 Kappa 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

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

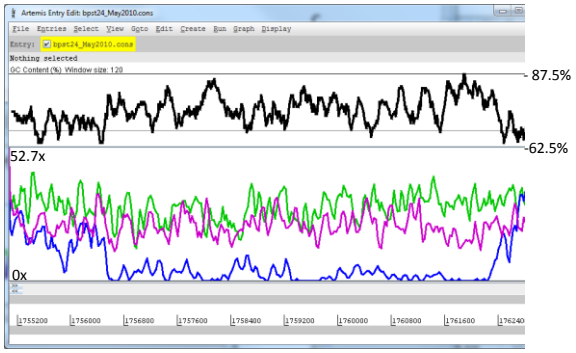
Supplementary Figure 1	Genome browser screenshots of selected regions in four genomes.
Supplementary Figure 2	Evenness of coverage based on different library amplification conditions across four genomes.
Supplementary Figure 3	Genome browser screenshot of an AT-rich region of the human X chromosome.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

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

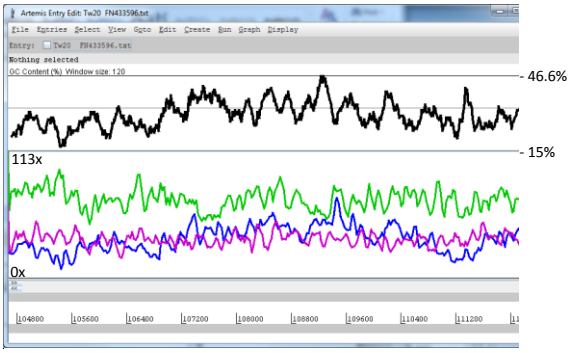
Nature Methods: doi:10.1038/nmeth.1814

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

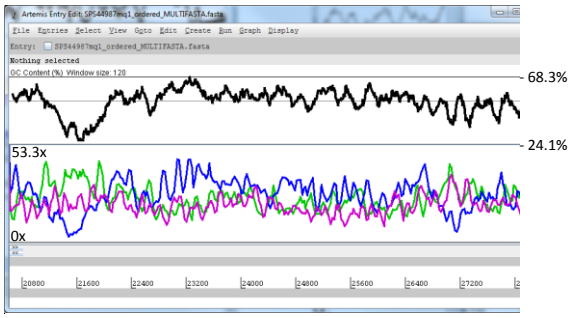
a.



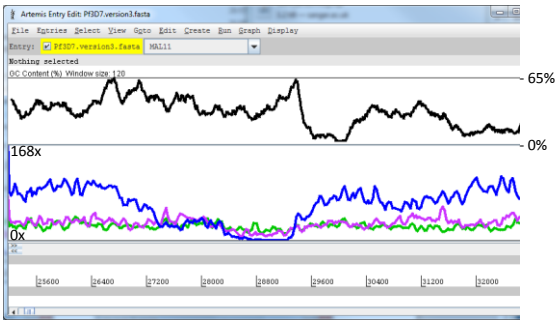
c.



b.



d.

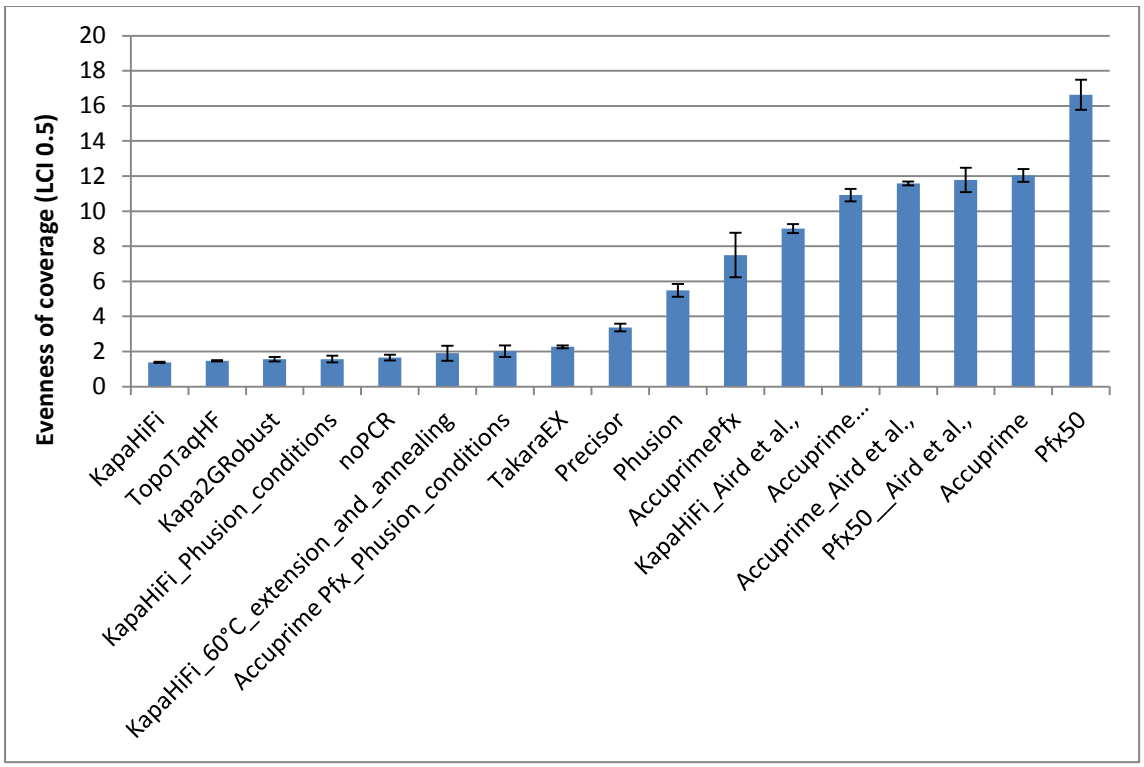


Genome browser screenshots of selected regions in the genomes of: **a.** *B. pertussis* (GC-rich region); **b.** *S. pullorum*; **c.** *S. aureus* and **d.** *P. falciparum* (AT-rich *var* gene region of chromosome 11). Libraries were prepared without PCR (green line), with 14 cycles of PCR using Phusion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

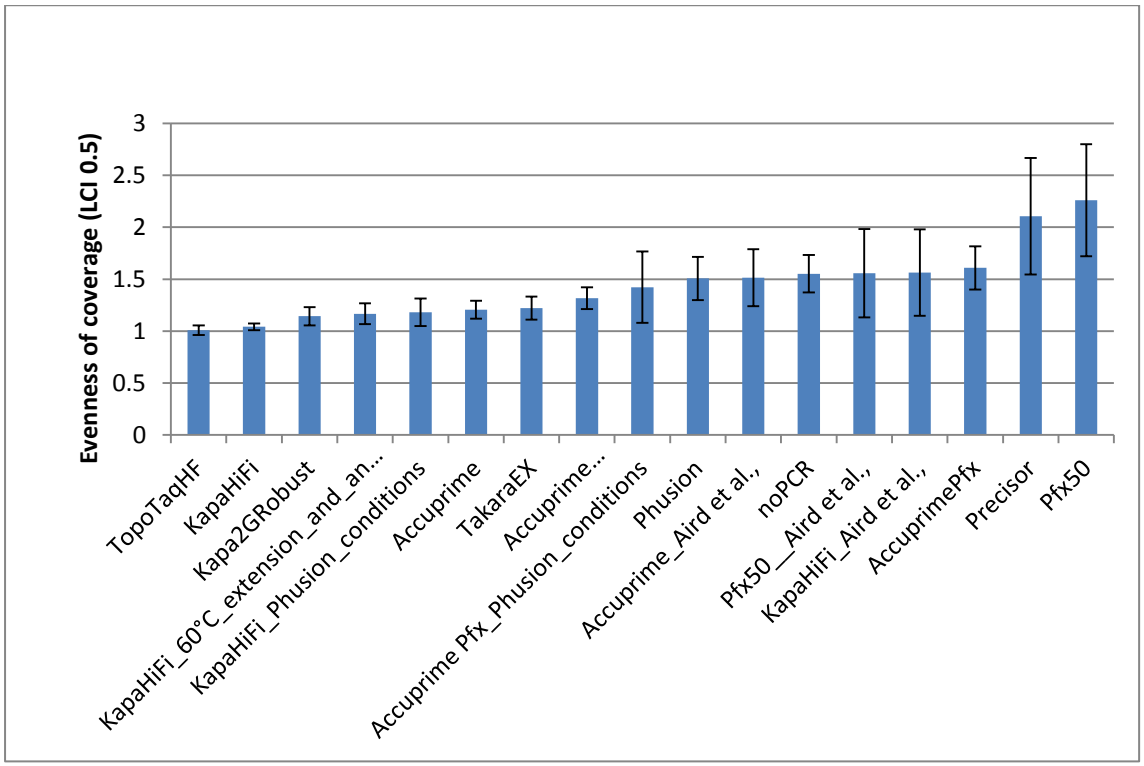
Nature Methods: doi:10.1038/nmeth.1814

Supplementary Figure 2. Evenness of coverage based on different library amplification conditions across four genomes.

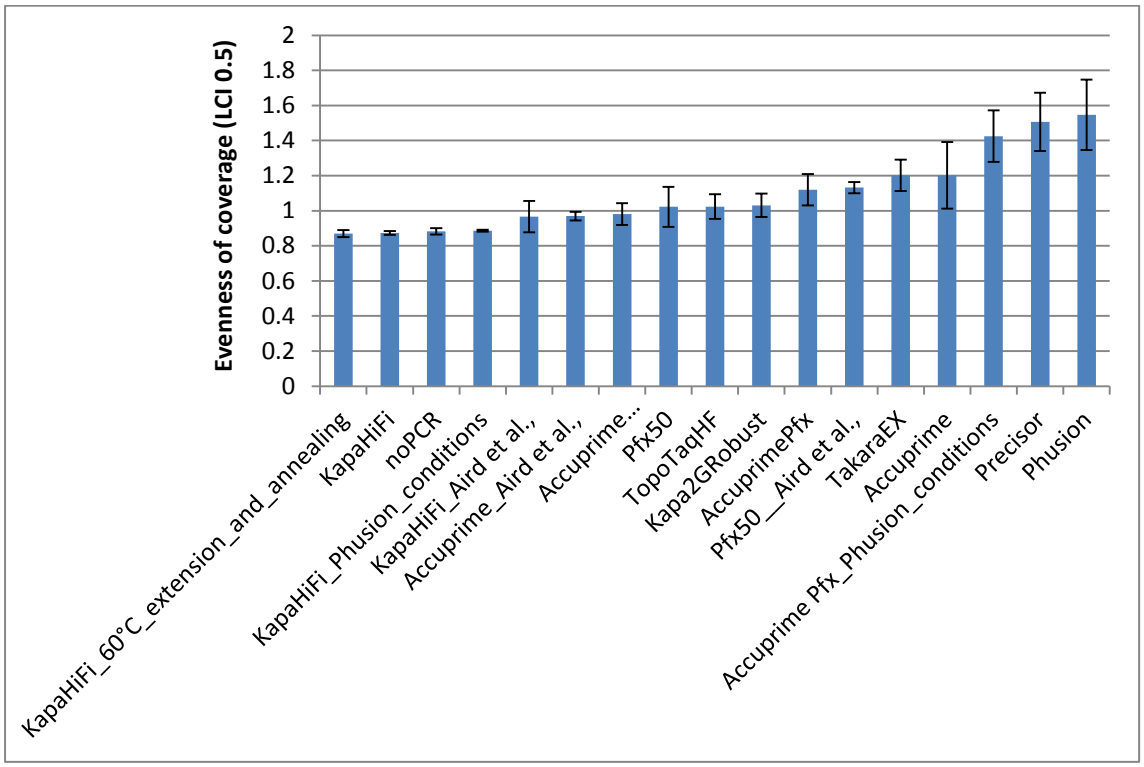
a. *B. pertussis*



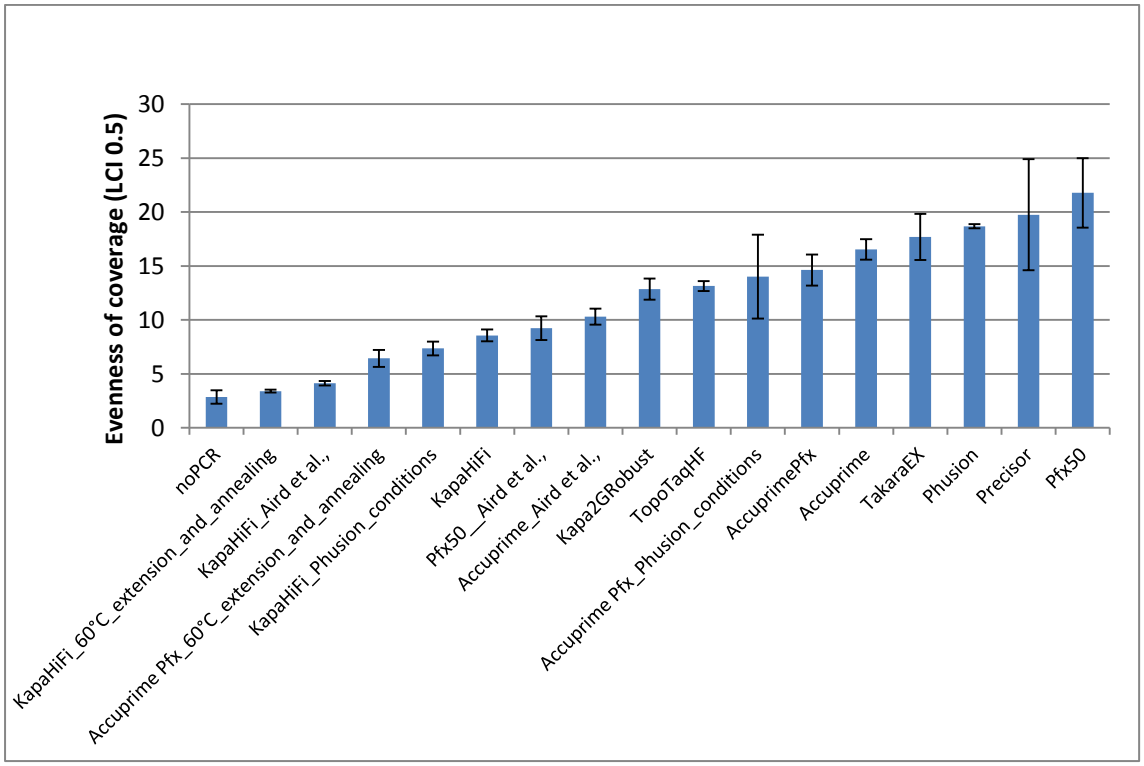
b. *S. pullorum*



c. *S. aureus*



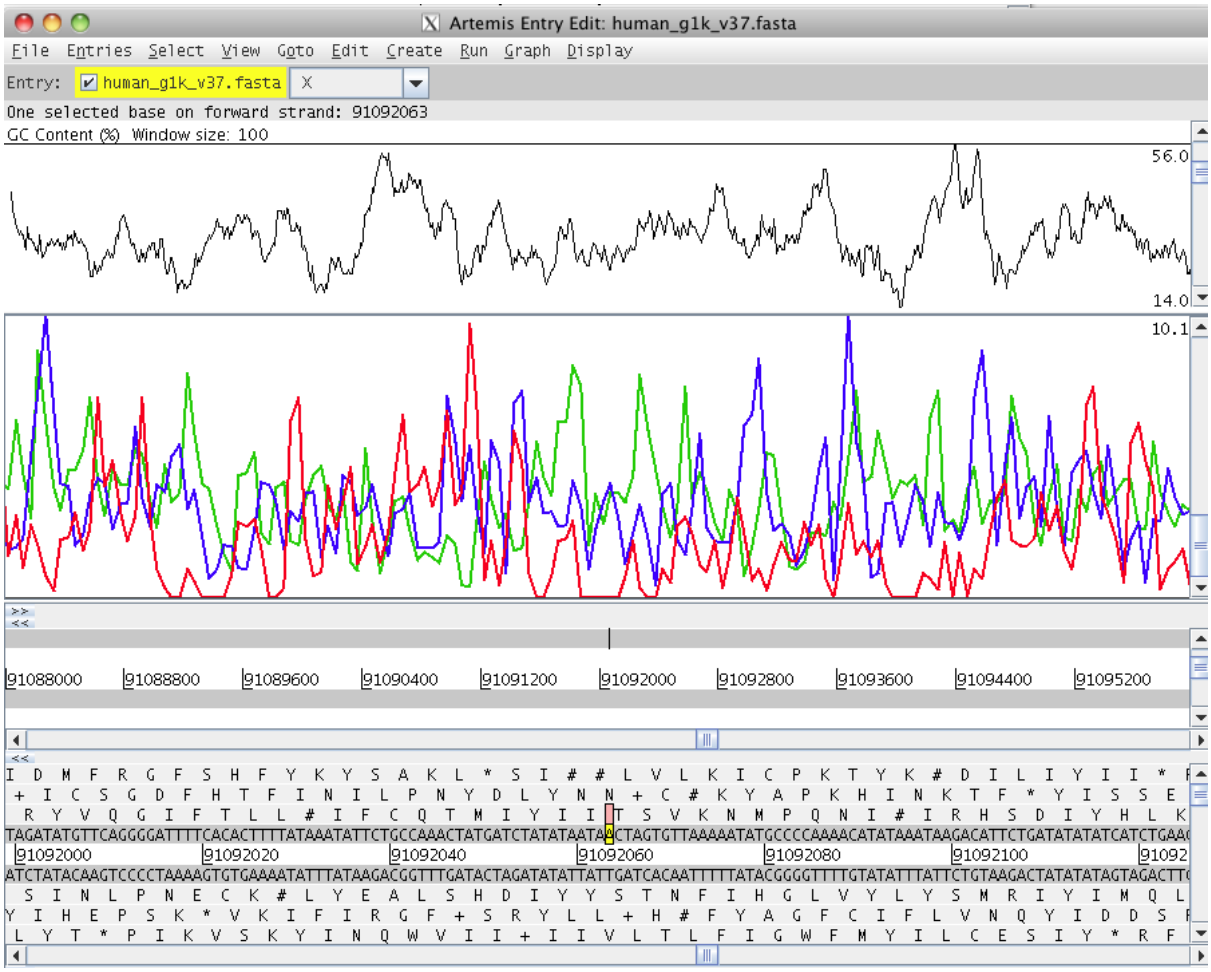
d. *P. falciparum*



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.

Nature Methods: doi:10.1038/nmeth.1814

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.

Nature Methods: doi:10.1038/nmeth.1814

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-GATCGGAAGAGCGGTTCAGCAGGAATGCCGA*G 3'

iPCR index read sequencing primer

5' AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC 3'

PE1.0

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T 3'

Modified multiplexing PE2.0 oligos

Oligo name	Single correcting, double & shift detecting octamers	Sequence obtained	PCR primers
iPCRtagT1	AACGTGAT	ATCACGTTAT	5' CAAGCAGAAGACGGCATACGAGATAACGTGATGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT2	AAACATCG	CGATGTTTAT	5' CAAGCAGAAGACGGCATACGAGATAAACATCGGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT3	ATGCCTAA	TTAGGCATAT	5' CAAGCAGAAGACGGCATACGAGATATGCCTAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT4	AGTGGTCA	TGACCACTAT	5' CAAGCAGAAGACGGCATACGAGATAGTGGTCAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT5	ACCACTGT	ACAGTGGTAT	5' CAAGCAGAAGACGGCATACGAGATACCACTGTGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT6	ACATTGGC	GCCAATGTAT	5' CAAGCAGAAGACGGCATACGAGATACATTGGCGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT7	CAGATCTG	CAGATCTGAT	5' CAAGCAGAAGACGGCATACGAGATCAGATCTGGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT8	CATCAAGT	ACTTGATGAT	5' CAAGCAGAAGACGGCATACGAGATCATCAAGTGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT9	CGCTGATC	GATCAGCGAT	5' CAAGCAGAAGACGGCATACGAGATCGCTGATCGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT10	ACAAGCTA	TAGCTTGTAT	5' CAAGCAGAAGACGGCATACGAGATACAAGCTAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT11	CTGTAGCC	GGCTACAGAT	5' CAAGCAGAAGACGGCATACGAGATCTGTAGCCGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT12	AGTACAAG	CTTGTACTAT	5' CAAGCAGAAGACGGCATACGAGATAGTACAAGGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT13	AACAACCA	TGGTTGTTAT	5' CAAGCAGAAGACGGCATACGAGATAACAACCAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT14	AACCGAGA	TCTCGGTTAT	5' CAAGCAGAAGACGGCATACGAGATAACCGAGAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT15	AACGCTTA	TAAGCGTTAT	5' CAAGCAGAAGACGGCATACGAGATAACGCTTAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT16	AAGACGGA	TCCGTCTTAT	5' CAAGCAGAAGACGGCATACGAGATAAGACGGAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT17	AAGGTACA	TGTACCTTAT	5' CAAGCAGAAGACGGCATACGAGATAAGGTACAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT18	ACACAGAA	TTCTGTGTAT	5' CAAGCAGAAGACGGCATACGAGATACACAGAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT19	ACAGCAGA	TCTGCTGTAT	5' CAAGCAGAAGACGGCATACGAGATACAGCAGAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'

iPCRtagT20	ACCTCCAA	TTGGAGGTAT	5' CAAGCAGAAGACGGCATACGAGATACCTCCAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT21	ACGCTCGA	TCGAGCGTAT	5' CAAGCAGAAGACGGCATACGAGATACGCTCGAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT22	ACGTATCA	TGATACGTAT	5' CAAGCAGAAGACGGCATACGAGATACGTATCAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT23	ACTATGCA	TGCATAGTAT	5' CAAGCAGAAGACGGCATACGAGATACTATGCAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'
iPCRtagT24	AGAGTCAA	TTGACTCTAT	5' CAAGCAGAAGACGGCATACGAGATAGAGTCAAGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3'

noPCR adapter

T_no_PCR

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T 3'

B_no_PCR

5' P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTT*G 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 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 5x read coverage (Coverage < 5x). 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 10x. 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:

$$lci(d) = \int_0^d C(x)dx$$

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.

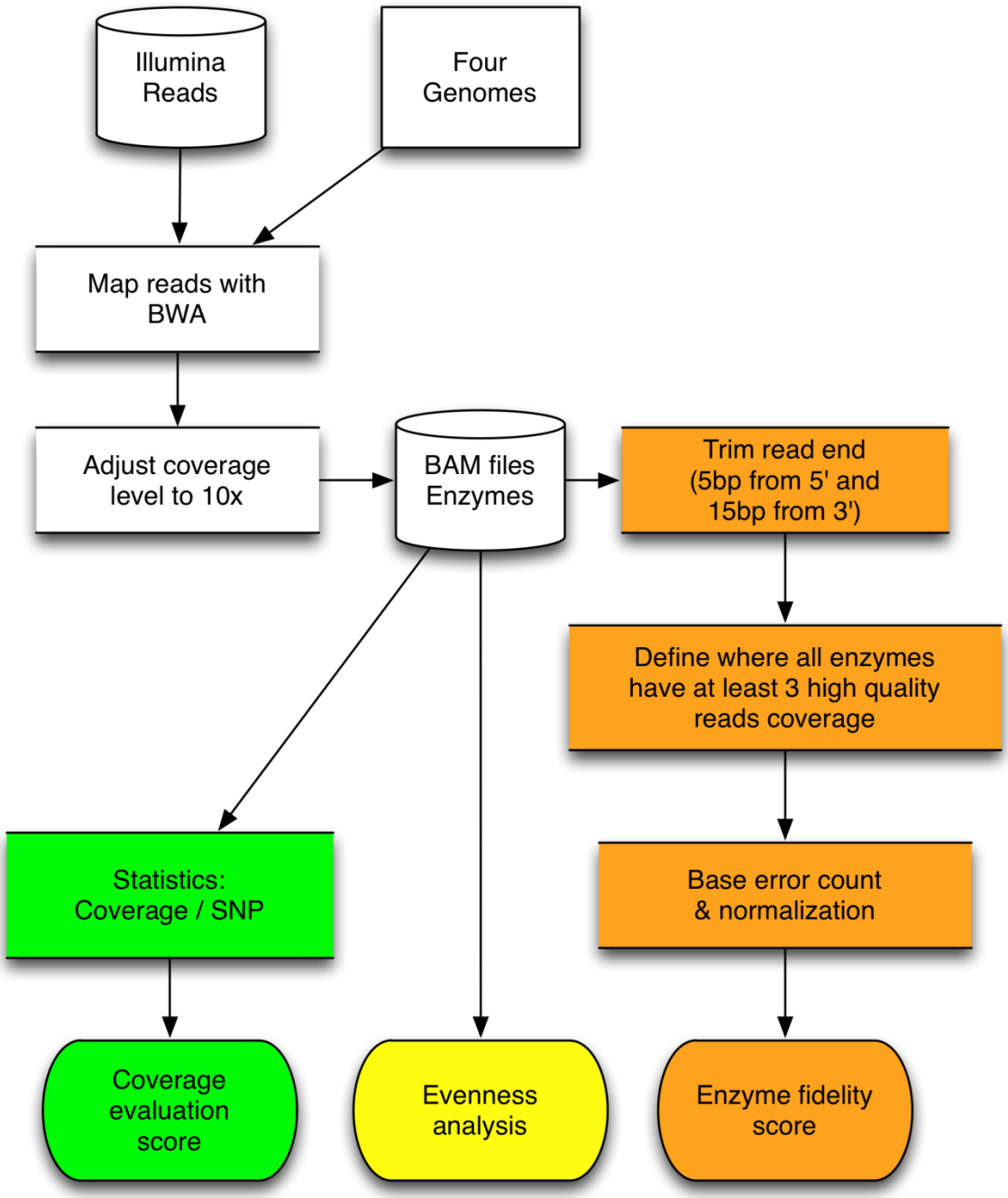
Enzyme-dependent fidelity

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.

1. Kozarewa, I. *et al*. *Nat Methods.* **6,** 291-295 (2009).
2. Li, H. and Durbin, R. *Bioinformatics.* **25,** 1754-1760 (2009).
3. Li, H. *et al*. *Bioinformatics.* **25,** 2078-2079 (2009).

Nature Methods: doi:10.1038/nmeth.1814