Optimal enzymes for amplifying sequencing libraries

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[2–4], 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

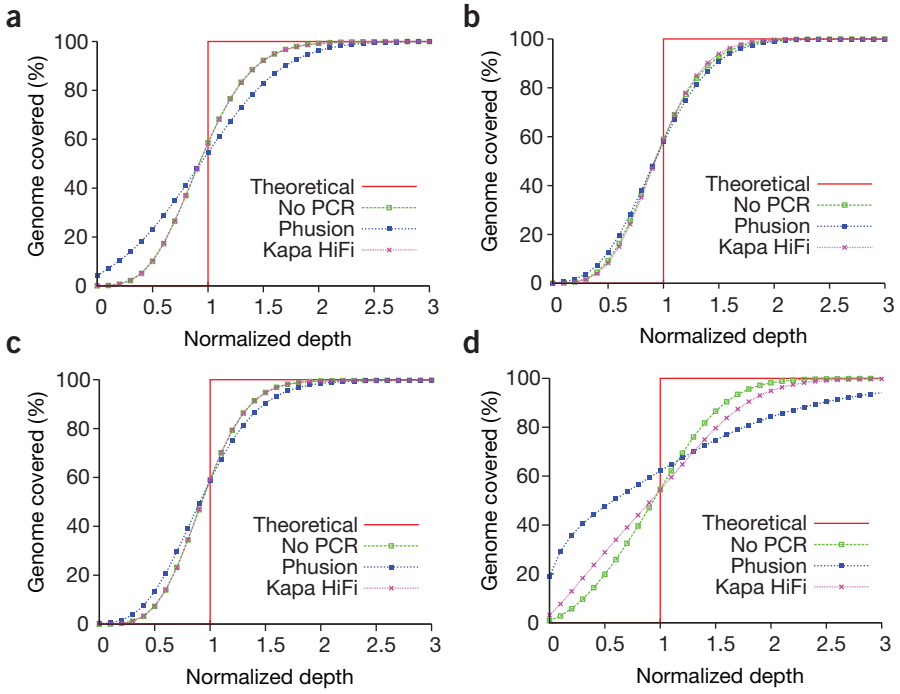

Figure 1 | Genome coverage uniformity plots for 10× Illumina sequence coverage. (**a**–**d**) Coverage of *B. pertussis* (**a**), *S. pullorum* (**b**), *S. aureus* (**c**) and *P. falciparum* (**d**) prepared without PCR (no PCR) or with 14 cycles of PCR using Phusion polymerase or Kapa HiFi polymerase. 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).

© 2012 Nature America, Inc. All rights reserved.

Tecan Ex. 2071, Page 1 of 2
IDT, Inc. v. Tecan Genomics - IPR2025-00015

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.

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.

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

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

Wellcome Trust Sanger Institute, Hinxton, UK.
e-mail: mq1@sanger.ac.uk

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).

© 2012 Nature America, Inc. All rights reserved.

Tecan Ex. 2071, Page 2 of 2
IDT, Inc. v. Tecan Genomics - IPR2025-00015