Optimization of library amplification for next-generation sequencing

Katja Heitz, Ioanna Andreou, Peter Hahn, Annika Piotrowski, Frank Reinecke, Dirk Löffert, and Nan Fang

QIAGEN GmbH, QIAGEN Strasse 1, 40724 Hilden, Germany

Abstract

Uniform coverage of all genomic regions during next-generation sequencing (NGS) is critical for efficiently utilizing sequencing capacity. It is also crucial in preventing loss of important sequence information due to dropout or under-representation of certain regions. Coverage uniformity is especially important in applications such as microbiome-sequencing, where different microbial strains may have significantly different GC contents. GC content-related sequencing bias may potentially lead to under-representation or even complete loss of the genomic regions or microbial strains with a very low or very high percentage of GC bases.

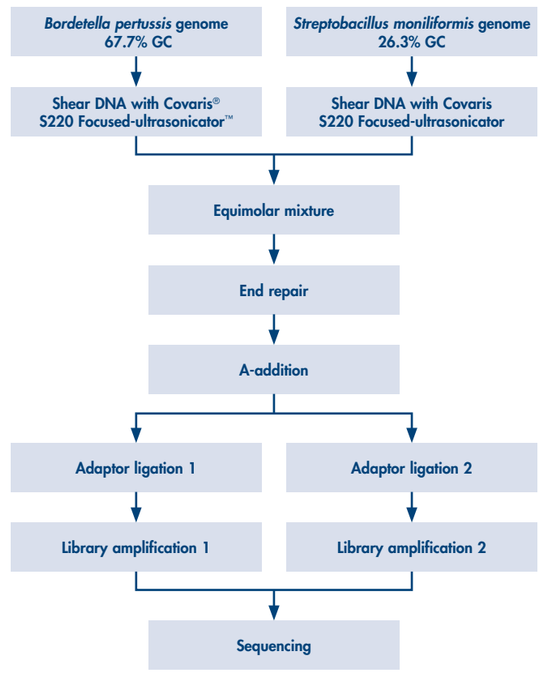
The PCR step of the NGS library construction procedure has been shown to be the major source of GC bias in the NGS workflow.[1,2] To solve this common problem in the NGS field, we established a test system where a mixture of high-GC and low-GC bacteria genomes was used to optimize library amplification conditions. This system has been subsequently used to develop a novel NGS library amplification mix that amplifies genomic regions with widely different GC contents with minimal bias and high fidelity.

1. Quail MA, Otto TD, Gu Y, et al. (2012) Optimal enzymes for amplifying sequencing libraries. Nat. Methods **9**, 10.
2. Aird D, Ross MG, Chen WS, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. **12**, R18.

A model system to optimize NGS library amplification

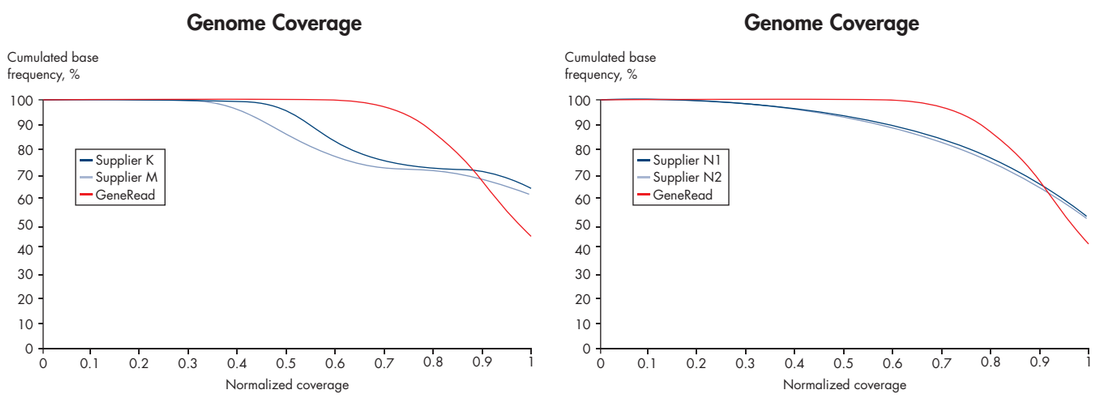
We designed a test system with mixed bacterial genomes and diverse GC contents to optimize library amplification conditions. The genomic DNA from *Bordetella pertussis*, which has 67.7% GC bases, and genomic DNA from *Streptobacillus moniliformis*, which contains 26.3% GC bases, are mixed in equal molarity. The DNA mixture is then subjected to the standard Illumina® sequencing library construction procedure: end-repair, A-addition, and adaptor ligation. Adaptors with different indices are used to test different amplification conditions, with the assumption that the 6-nt indices will not affect amplification efficiency. Adaptor-ligated sequencing libraries are subjected to 14 cycles of PCR amplification to test potential amplification bias extensively.

An ideal library amplification solution should amplify the two genomes with dramatically different GC contents in this gDNA mixture with equal efficiency and high fidelity. The experimental workflow is shown in the flow-chart.



Comparison of genome coverage evenness

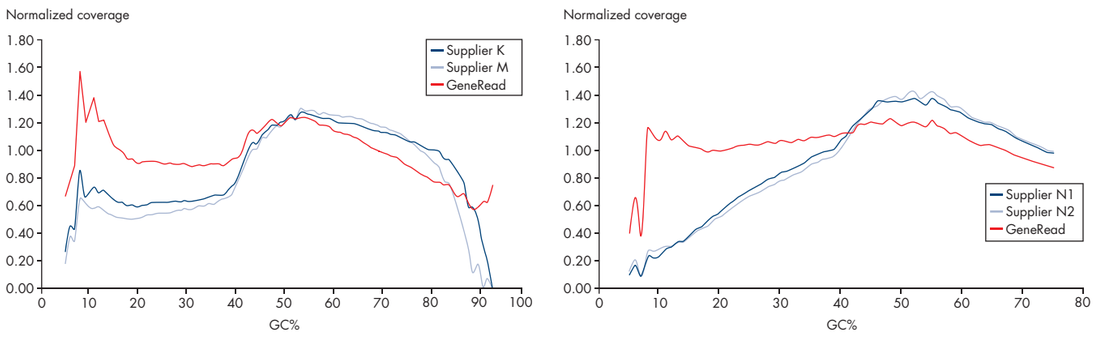
With the mixed high-GC and low-GC genomes as model system, we optimized the composition of the GeneRead™ Library Amplification Mix, as well as the PCR cycling conditions. The optimized GeneRead Library Amplification Mix shows significantly better genome coverage compared with other commercially available solutions. Two independent experiments are shown, where GeneRead Library Amplification Mix was compared with commercially available library amplification mixes. With GeneRead over 90% of all bases from the mixture of the two genomes are covered with 0.8 or more of normalized coverage. In contrast, the libraries generated by other amplification mixes showed only about 70% of all based covered with 0.8 or more of normalized coverage.



Comparison of the coverage evenness of GeneRead Library Amplification Mix to that of library amplification mixes from other suppliers. Normalized coverage: The ratio of "coverage" in specific bases vs. the mean coverage of all bases. A number of 1 represents mean coverage, a number <1 represents lower than mean coverage and a number >1 represents higher than mean coverage. **Cumulated base frequency:** Fraction of all bases with this coverage or higher in percent.

Comparison of coverage evenness on genomic regions with differing GC contents

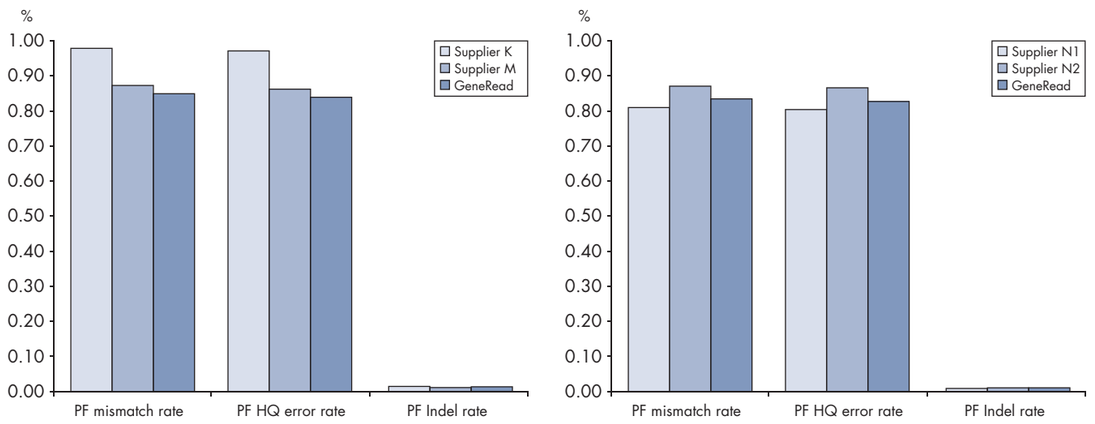
Using the model system, we were also able to optimize the performance of the library amplification mix so that it amplifies genomic regions with different GC contents with similar efficiency. Two independent experiments are shown, where GeneRead Library Amplification Mix was compared with commercially available library amplification mixes. The optimized GeneRead Library Amplification Mix delivers the most even region coverage, even with vastly different GC contents. The most significant advantage is observed with good sequencing coverage in AT-rich regions.



Comparison of the coverage evenness of GeneRead Library Amplification Mix to that of library amplification mixes from other suppliers. Normalized coverage: The ratio of "coverage" in a specific GC bin vs. the mean coverage of all GC bins. A number of 1 represents mean coverage, a number <1 represents lower than mean coverage and a number >1 represents higher than mean coverage. **GC:** The G+C content of the reference sequence. Values are 0–100%.

Comparison of sequencing error rate in libraries generated using different amplification mixes

The amplification fidelity of different PCR mixes was also evaluated using the model system. Two independent experiments are shown, where GeneRead Library Amplification Mix was compared with commercially available library amplification mixes. The sequencing library generated with GeneRead Library Amplification Mix shows comparable or lower errors in the sequencing. These errors are detectable above the systematic sequencing errors arising from the Illumina miSeq® sequencing system.



Error rate comparison. Sequencing libraries were generated and amplified with either GeneRead Library Amplification Mix or the high-fidelity PCR mixes from other suppliers. **PF mismatch rate**: Rate of bases mismatching the reference for all bases aligned to the reference sequence. **PF HQ error rate**: Percentage of bases that mismatch the reference in PF HQ-aligned reads. **PF indel rate**: Number of insertion and deletion events per 100 aligned bases. Uses the number of events as the numerator, not the number of inserted or deleted bases.

Summary

We established a model system where two bacterial genomes with significantly different GC contents are mixed and used to optimize PCR-based library amplification conditions. Based on this test system, we developed the GeneRead Library Amplification Mix, which shows significantly less bias in amplifying genomic regions with differing GC contents.

Such unbiased NGS library amplification solution is highly suited for NGS applications including:

- Metagenomic sequencing — where diverse microbiomes in a sample may have genomes with highly variable GC contents
- Whole genome sequencing — where GC- and AT-rich isochores in the genome often have functional implications such as variable gene expression levels
- Methylation analysis by bisulfite sequencing — where bisulfite treatment of DNA dramatically increases AT content
- Library construction from low input materials — where a high number of PCR cycles is needed and unbiased amplification is essential

The applications presented here are for molecular biology applications. They are not intended for the diagnosis, prevention or treatment of a disease.

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at www.qiagen.com or can be requested from QIAGEN Technical Services or your local distributor.

1085861 08/2014