Tools for Enhancing Sequence Diversity and Reducing Bias in DNA-seq Library Preparation

Suchitra Ramani, Dawn Obermoeller and Masoud M. Toloue

Bioo Scientific, 3913 Todd Lane Suite 312 Austin, TX 78744
Correspondence should be addressed to M.T. (mtoloue@biooscientific.com)

The generation of high quality next generation sequencing data begins with libraries that have the desired insert size and proper adapter ligation. Artifacts during library preparation can result in PCR duplicates, uneven read coverage and poor adapter ligation efficiencies. We examined this by generating and analyzing DNA sequences from *Escherichia coli* genomes. Our improved protocol significantly reduces bias, increases ligation efficiency, improves indexing flexibility and increases high throughput functionality in DNA libraries prepared for sequencing.

INTRODUCTION

Next generation sequencing (NGS), introduced nearly 30 years after Sanger Sequencing, has revolutionized the process of sequencing (1). High throughput sequencing of multiplexed libraries in a short amount of time has significantly reduced the cost of sequencing. Due to efficient large scale sequencing, it has resulted in a plethora of options: from genome wide sequencing of complex targets to identifying how proteins interact with DNA (2).

After the completion of sequencing the human genome in 2003, it became obvious that novel methods would have to be developed to sequence multiple genomes of diverse species including our own. Since then, strategies for massive parallel sequencing have revolutionized research across diverse scientific disciplines. Despite these advances, DNA and RNA sample preparations, one of the most important aspects of next generation sequencing, continue to use outdated and cumbersome methods. Current fragmentation, ligation and amplification methods are susceptible to sequence bias, require significant preparatory time and are highly inefficient. Generating high quality NGS data begins with high quality libraries that have the desired insert size and proper adapter ligation (3).

NEXTflex™ DNA SEQUENCING KITS

Genomic DNA fragments can be transformed into libraries ready for sequencing with the following 4 steps: 1) End-Repair (blunt end formation), 2) Adenylation (adding an "A" base), 3) Ligation of adapters and 4) PCR amplification. The NEXTflex™ DNA Sequencing Kit is designed to prepare single, paired-end and multiplexed genomic DNA libraries for sequencing using Illumina®/Solexa® GAII and HiSeq 2000/1000 platforms. This kit has been designed to increase ligation efficiency, improve indexing flexibility and increase high throughput functionality (**Table 1**). The number of ligation events (adapters binding to genomic DNA) directly correlates with the degree of sequence diversity and number of sequencing reads. Using specialized modifications to the NEXTflex Ligation Mix, we have developed "Enhanced Adapter Ligation Technology" which yields 1.5 - 2 fold more reads per sequencing run (**Table 2**). By making improvements to the ligation enzymatic mix, the user will now have the ability to perform ligations with longer adapters and expect to see better binding efficiencies. This NEXTflex™ kit simplifies workflow by using master mixed reagents and magnetic bead based cleanup, reducing pipetting and eliminating time consuming steps in library preparation. In addition, the availability of up to 48 unique adapter barcodes makes this the most high-throughput kit available.

Protocol

Table 1 | Comparison of NEXTflex™ and Company X's Library Prep Kits.

	Competitor X's Protocol	NEXTflex™ DNA Sequencing Kit
Enhanced Adapter Ligation Technology	No	Yes
Compatible Multiplex Barcodes	12	96
Price/Reaction	Higher	Lower
Magnetic Bead Purification	Yes	Yes
Automation Friendly	Yes	Yes
Length of Protocol	Shortened	Shortened

Table 2 | Summary of Sequence Data Between NEXTflex™ and Company X's Libraries.

	Sample	Genome Size (Mb)	Read Length (bp)	Insert Size (bp)	Number of Reads	Read Coverage	Perfectly Mapped Reads	# of Unknown Bases	Contig N50
NEXTflex™	*E.coli*	5	36	200	29,718,673	115	96%	3,201	102,408
Company X's Kit	*E.coli*	5	36	200	19,311,940	75	89%	6,540	95,299

DEEP SEQUENCING OF *E. COLI* GENOME

As can be seen in **Table 2**, sequencing of the *E. coli* genome was performed using the NEXTflex™ DNA Sequencing kit and Barcodes and compared to Company X's protocol. The NEXTflex™ enhanced ligation technology resulted in significantly improved reads, coverage and assembly. The *E. coli* genome was assembled with less scaffolds (result of connecting contigs by linking data), had an increased average maximum scaffold size and N50 (measure of contig length containing a typical nucleotide, maximum length such that 50% of a nucleotides lie in a contig).

MULTIPLEXING OPTIONS

The NEXTflex™ DNA Barcodes are adapters containing indexed sequences and offer an improved multiplexing workflow and flexible setup. This new automation-friendly format enables multiplexing of up to 48 samples for a total of 384 reactions. The NEXTflex DNA Barcodes are available in sets of 6, 12, 24 and 48 unique adapters. The ability to pool samples in an efficient way significantly decreases hands on time while providing robust data quality.

The NEXTflex™ DNA Barcodes simplify high-throughput sequencing with up to 48 available adapter combinations. Using the NEXTflex™ DNA Sequencing Kit, 48 libraries were prepared and tagged with a different barcoded adapter. The percentage of index reads that could be used is represented by the height of the bars in **Figure 1**. Most barcoded adapters were read perfectly. The NEXTflex™ index utilizes double error correction ensuring that single base changes during phasing do not unrecognizably alter the barcode sequence. The NEXTflex™ adapters contain the full complement of

Figure 1 | NEXTflex DNA Barcodes Adapter Reads. The percentage of usable index reads is illustrated for all 48 of the NEXTflex DNA Barcode adapters.

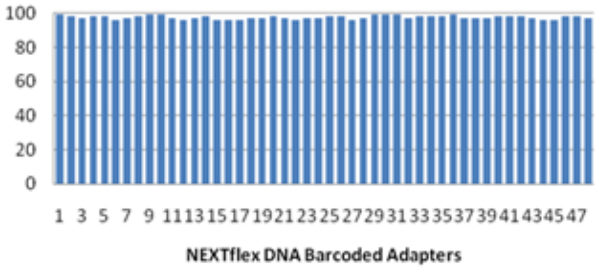
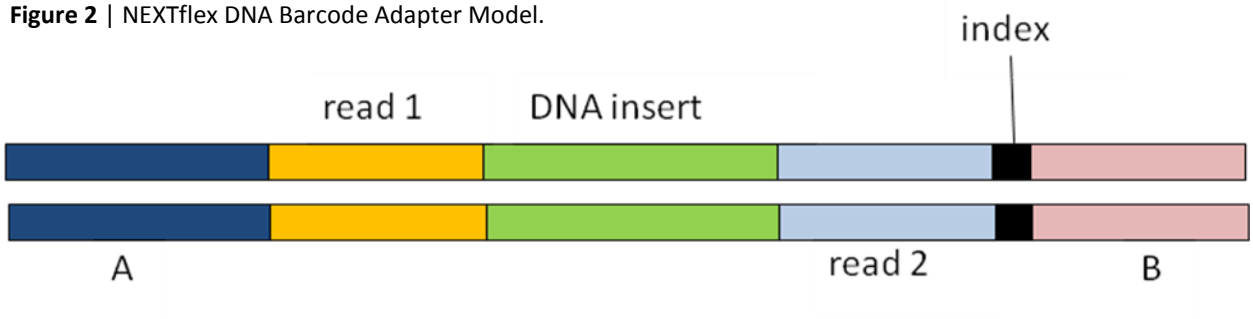


Figure 2 | NEXTflex DNA Barcode Adapter Model.



sequencing flow cell binding regions (**Figure 2. A, B**), which eliminates the need to perform PCR to add the barcode tag.

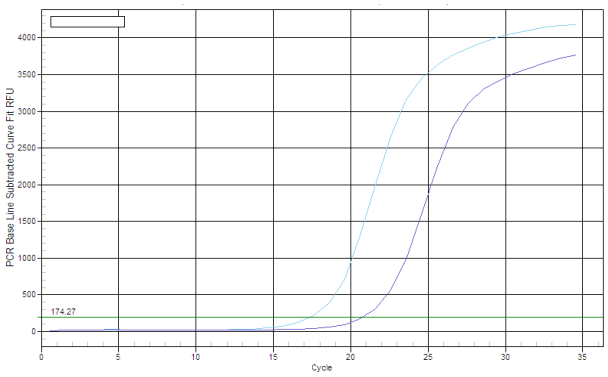
AMPLIFICATION-FREE TECHNOLOGIES

Despite the advances made recently in NGS library preparation technologies, there still lies a great inefficiency as most commercial library preparation approaches today require the amplification of DNA by PCR, an inherently biased procedure. Amplifying AT or GC rich genomic regions often leads to sequence biased nucleotide compositions and poses a serious challenge during analysis. These amplification artifacts are introduced during PCR and can increase the proportion of these sequences that are duplicated and cause an uneven distribution of read coverage across targeted sequencing regions. This can result in problems with genome assembly and variation analysis from the short reads. In assessing bias, the presence of low quality and high GC reads often cannot be aligned against a reference genome. High GC containing profiles tend to shift toward higher GC content, an indication of poor base representation (4,5).

It is essential to obtain as true of a representation of the genome as possible. Amplification biases significantly hinder this by reducing complexity. Further sequencing of the same library is not sufficient to improve diversity and often masks low concentration fragments. This necessitates preparing new libraries, optimizing input DNA and re-sequencing until desired depth is achieved. The main aims for library enhancement should be to overcome these limitations by developing a transforming library preparation method with a highly efficient ligation step and a method that completely eliminates the need for PCR.

Bioo Scientific's NEXTflex PCR-Free DNA Sequencing Kit offers an amplification-free method of library preparation, reducing the incidence of duplicate sequences, improving read mapping and SNP calling and aiding de novo assembly. Using specially designed master-mixed enzymes, the NEXTflex™ PCR-Free DNA Sequencing kit completely eliminates the need for amplification, enabling better read mapping and a reduction in duplicate sequences leading to reduced sequence cost and bias for more representative base identities and better de novo assembly. Using the NEXTflex™ PCR-Free Kit, the user will reduce the number of duplicate sequences ensuring a more representative matched number of reads. While the quantity of template generated using the NEXTflex™ PCR-Free DNA Sequencing Kit is lower than the NEXTflex™ DNA Sequencing Kit, library, quantification by qPCR demonstrates that from as a little as 3 µg of DNA a sufficient amount of 300-500 bp PCR-free library can be obtained for greater than 600 high density HiSeq lanes (Figures 3, 4). The NEXTflex™ PCR-Free

Figure 3 | Example of qPCR template amplification for 3 µg and 1 µg of input DNA. The Ct value for 3 µg (light blue) is at least 3 Ct lower than the 1 µg (purple) sample.



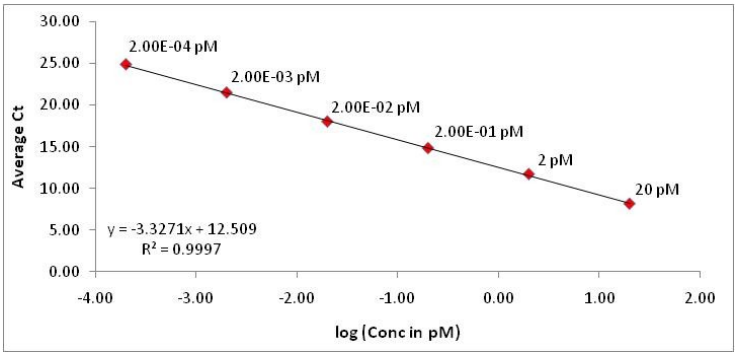


Figure 4 | Standard curve generated by plotting the log concentration of library DNA (pM) against average Ct values. Used to determine cluster density before loading sample on a flow cell.

Barcodes, available in sets of up to 48 barcodes, can be used to multiplex these reactions.

The optimized NEXTflex library preparation steps are significantly better and more robust than standard protocols in that they improve reads, coverage and assembly. In the PCR-Free version of the protocol amplification biases are eliminated by completely removing the PCR-amplification step. The PCR-Free version does require using qPCR an inexpensive and quick assay for measuring representation bias and cluster density in libraries (**Figures 3, 4**). Finally with the advent of higher sequencing coverages, multiple samples can be sequenced in a single flow cell lane. Simultaneous sequencing of large numbers of samples is possible by including sample-specific short identifying nucleotide sequences on adapter sequences. Indexing has the obvious benefit of multiplexing samples within a single flow cell lane, reduces costs significantly, has the advantage of measuring base error rate, allows the user to perform cross genomic studies, time courses, drug induced cellular experiments and monitor day to day expression variability between samples. The combination of high ligation efficiency, the ability to eliminate the PCR step completely and large multiplexing capability make NEXTflex the superior choice for library preparation.

ACKNOWLEDGEMENTS:

This work was supported by a grant to M.T. (1R43HG006221-01) from the National Human Genome Research Institute at the National Institutes of Health and a grant (1047285) from the National Science Foundation.

REFERENCES:

1) Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. Trends Genet. 24(3):133-41.

2) Metzker, M.L. (Jan, 2010) Sequencing technologies - the next generation. Nat Rev Genet. 11(1):31-46.

3) Quail, M.A. et al. (Dec 2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods. 5(12):1005-10.

4) Aird, D. et al. (Feb 2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12(2):R18.

5) Kozarewa, I et al. (Apr 2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. (4):291 -5.

6) Cronn, R. et al. (Nov 2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36(19):e122.