Cross-Site Comparison of Enzymatic Illumina Library Construction Kits


Abstract

The ability to produce genomic DNA libraries for Illumina sequencing depends on fragmentation of the DNA. Fragmentation of high molecular weight DNA has typically been accomplished through physical means such as nebulization or sonication. These techniques frequently involve substantial dilution of limiting samples and are not readily scalable for high throughput production. In response, various manufacturers have developed strategies for enzymatic fragmentation of the DNA. Open questions include how well these techniques provide unbiased representation of the samples under study, whether nucleotide composition of the target material plays a role, and how consistent are the outputs across multiple sites. To address these questions the DNA sequencing Research Group (DSRG) compared five different methods for enzymatic fragmentation of DNA prior to next-gen library generation. Two of the kits, from Agilent (A) and Illumina (NXT), are transposase based, whereas the three others from Qiagen (Q), Kapa Biosystems (K), and New England Biolabs (NEB) utilize other nuclease. Both mouse genomic DNA and bacterial pooled DNAs representing a spectrum of GC compositions were used as starting material. We will discuss the data generated from the different library kits focusing on library complexity, context effects and site-to-site consistency as well as the implications for the selection of kits for specific experimental requirements.

Materials and Methods

Five commercially available fragmentation and corresponding library prep kits were chosen for study: 1) Illumina Nextera XT; 2) NEBNext Fragmentase + NEBNext Ultra II; 3) Kapa HyperPlus; 4) Agilent SureSelect (QXT and 5) Qiagen QIAseq FX. See below for experimental design. DNA samples were quantified by Qubit, and assessed for quality via agarose gel at a single site before distribution to the DSRG member sites. 50 ng of DNA was used as input for each library except Nextera XT which used the recommended 1 ng input. Aliquots of each library were sequenced before the final PCR in order to assess the fraction of potentially productive molecules in each prep. All libraries were pooled and sequenced on an Illumina NextSeq, using both a Mid- and High Output kit to generate 2x75 bp reads. For all samples, quality trimming and filtering were done using FASTX toolkit (v0.11.2) (quality score Q20 and minimum length 50nt as cutoffs). The reads were then mapped to reference genomes (mm10 for mouse and NC_009313.1 Ecoli, NC_004641.1 S. epidermidis, NC_012860.1 P. fluorescens and NC_012803.1 M. luteus) using Bowtie2 (v2.2.6) with default parameters. For mouse samples, uniquely mapped reads were used for profiling TSS and genebody coverage using NGSplot (v.2.47) and GC content was estimated using gcStats (v2.47) and GC content and insert size were estimated using kmaip (v2.1.3) and nucleotide distribution of first 20 bases was determined using Homer (v4.7b).

Overall Results

Bacterial Pool Library Performance

Sequenced Insert sizes of bacterial libraries

Genomic areas of over and under representation Bins are considered underrepresented if they contain less than 20% of the median for that sample (lighter color) and overrepresented in they contain 500% of the median (darker color). Each column is a different replicate.

Murine Library Performance

GC content distribution of mapped reads

Insert size histogram of mapped reads

Acknowledgments

The DSRG would like to thank the ABRF EB for study support; Agilent Technologies, and New England Biolabs for donation of library prep reagents; Qiagen and Kapa Biosystems for donation of both library preparation and sequencing reagents; and Scott Tighe for donation of purified, characterized bacterial DNA samples.