Cross-Site Comparison of Enzymatic Illumina Library Construction Kits

Marie Adams[1], Charles M. Nicolet[8] , Yuriy Alekseyev[2], Ketaki P. Bhide[11], Jun Fan[3], Jeanne M. Geskes[4], Joanna S. Kerley-Hamilton[5], Jamie P. Kershner[6], Jessica W. Podnar[9], Savita Shanker[10], Vidya Subramanian[7], Jyothi Thimmapuram[11,] Stuart S. Levine[7],

1. Univ. of Wisconsin 2. Boston Univ. 3. Marshall Univ. 4. Univ. of Michigan 5. Dartmouth Univ. (currently at Kapa Biosystems) 6. Univ. of Colorado Boulder 7. Massachusetts Inst. of Technology 8. Univ. of Southern California 9. Univ of Texas 10. Univ. of Florida 11. Purdue Univ.

Abstract

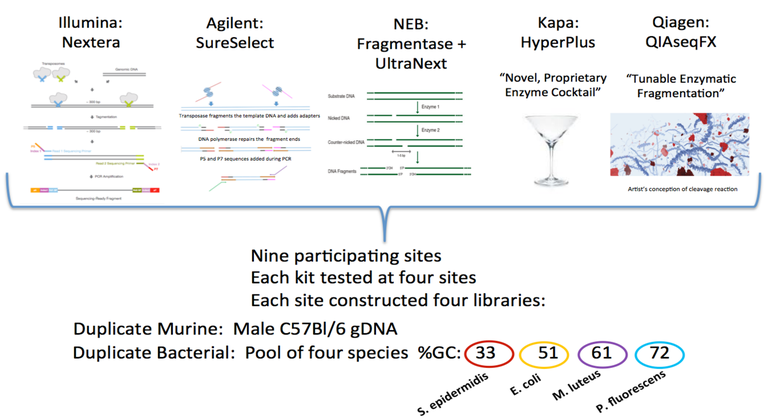
The ability to produce genomic DNA libraries for Illumina sequencing depends on fragmentation of the DNA. Fragmentation of high molecular weight DNA has typically been accomplished through physical means such as nebulization or sonication. These techniques frequently involve substantial dilution of limiting 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 nucleases. 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

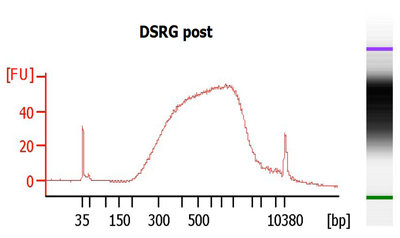
Five commercially available fragmentation and corresponding library prep kits were chosen for study: 1) Illumina Nextera XT, 2) NEBNext Fragmentase +NEBNext Ultra II, 3) Kapa HyperPlus, 4) Agilent SureSelect QXT and 5) Qiagen QIAseq FX. See below for experimental design. DNA samples were quantified by 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 sequestered 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_000913.3 Ecoli, NC_004461.1 S. epidermidis, NC_012660.1 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 and insert size were estimated using qualimap (v2.1.3) and nucleotide distribution of first 20 bases was determined using HOMER (v4.7b).

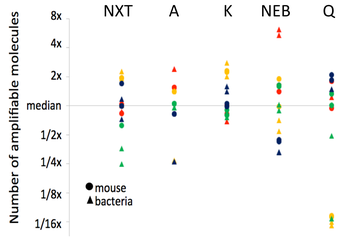
Experimental Design



Overall Results



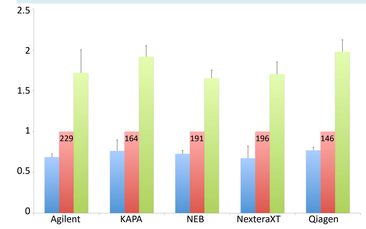
Bioanalyzer trace of 72-library pool



Library product yields show inter- and intra-site kit consistency

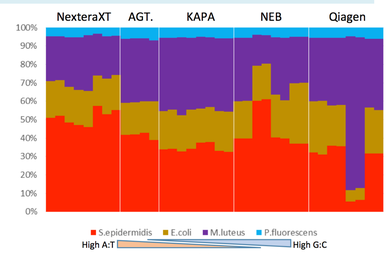
Ct of the unamplified libraries was compared to known standards to determine relative quantification, and normalized to the median value to measure variation in efficiency between sites (variable color).

Bacterial Pool Library Performance



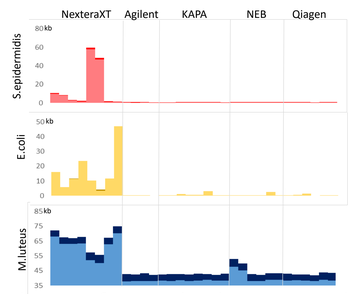
Sequenced Insert sizes of bacterial libraries

Median insert size varied among sites and kits and was set as 1.0 in the above graph (red). The average 90th percentile (green) and 10th percentile (blue) percent variances and standard deviations are indicated. The median insert sizes are indicated in the red bar.



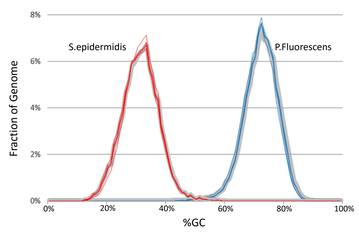
Representation of bacterial sequences

Despite equimolar pooling the final recovery of sequences showed bias against the highest GC samples with all kits. Kit differences in AT recovery, as well as individual variances, are also evident



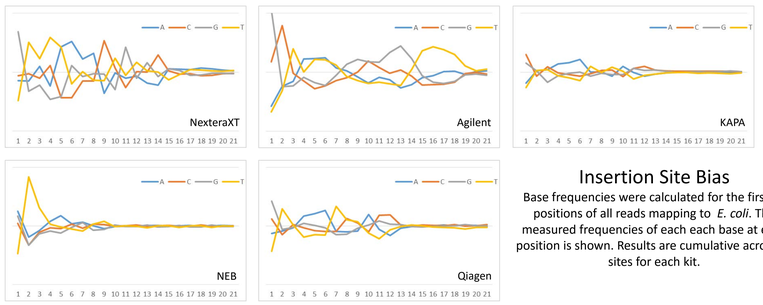
Genomic areas of over and under representation

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.



GC bias in bacterial strains

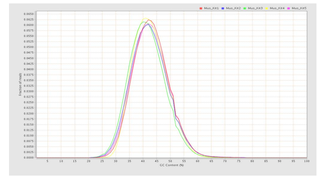
Cleavage or insertion sites were measured on each of the indicated strains, and the GC% for each binned site was calculated. Background distributions are shown in gray. For each sample, the %GC of the cut sites was measured for each read and plotted. No bias was detected among any of the five kits. Red/Blue lines – histograms of %GC for representative samples of each of the 5 tested kits.



Insertion Site Bias

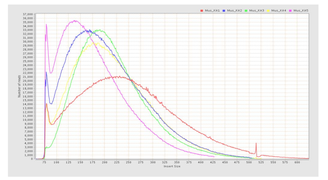
Base frequencies were calculated for the first 21 positions of all reads mapping to *E. coli*. The measured frequencies of each each base at each position is shown. Results are cumulative across all sites for each kit.

Murine Library Performance



GC content distribution of mapped reads

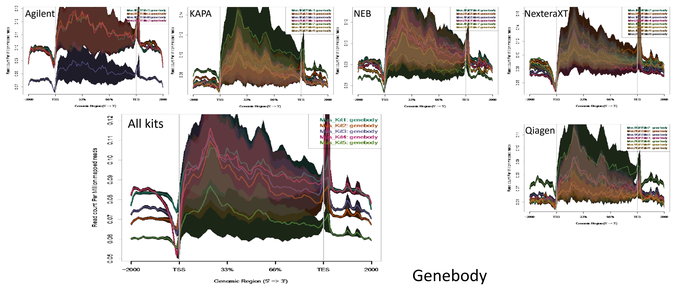
Data are averaged for each of the kits across all sites

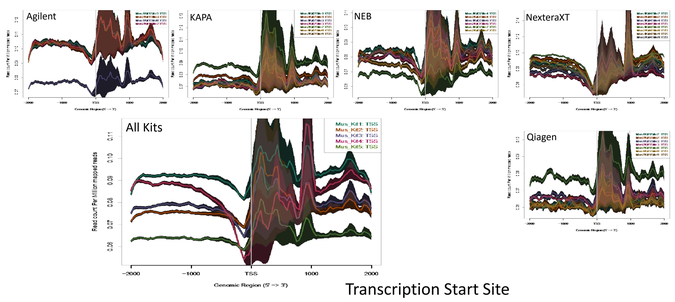


Insert size histogram of mapped reads

Data are averaged for each of the kits across all sites

Red is Agilent, Blue is KAPA, Green is NEB, Yellow is NexteraXT, Pink is Qiagen





Coverages across specific murine genomic regions

Kit specific coverages acrross the indicated genome regions are plotted for all aligning reads. Data from individual sites as well as the averaged values across all sites are shown.

Conclusions

Intra- and Inter-site variability was low with all kits for most of the measured metrics

Some differences between kits are apparent, but overall the kits behaved similarly to each other for most of the coverage metrics measured

Choice of enzymatic-based fragmentation methods may be made without regard to bias concerns

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.