Achieving High Coverage and Yield from GC and AT Rich Genomes

Introduction

Standard Next Generation Sequencing (NGS) library preparation methods include a polymerase chain reaction (PCR) amplification step prior to cluster generation and sequencing, to meet minimum molarity cutoffs required by different sequencing platforms. Biases associated with PCR amplification include uneven coverage of regions with extreme base composition, increased numbers of duplicate fragments, decreased mapping quality and poor variant calling. Particularly intractable regions of extreme base composition include GC-rich regions, which remain difficult to uniformly PCR amplify. While the human genome is not considered GC-rich (3.2 Gb; 41% GC), certain regions of the genome can be extremely GC rich, e.g. retinoblastoma tumor suppressor gene *RB1* (104 of the first 136 coding bases are either G or C). Unbiased amplification of highly complex whole genome sequencing (WGS) libraries, whole exome sequencing (WES) libraries and targeted re-sequencing libraries is required to obtain high quality data for *de novo* genome assembly and accurate calling of single nucleotide polymorphisms (SNPs). Here, we show that the NEXTflex™ PCR polymerase is a leader in providing high quality amplification. The NEXTflex PCR polymerase, used in all of the NEXTflex DNA-seq and RNA-seq NGS library preparation kits, contains a highly optimized and robust enzyme that exhibits minimal GC bias and produces uniform coverage of difficult to sequence genomes.

Methods

Genomic DNA was isolated from *Deinococcus radiodurans* (3.2 Mb; 67% GC) and *Saccharomyces cerevisiae* (12.1 Mb; 38% GC). Genomic DNA was Covaris sheared to an average fragment size of 200 bp. 20 ng of *D. radiodurans* (GC-rich) and *S. cerevisiae* (AT-rich) DNA was used as starting material for library preparation using the NEXTflex™ ChIP-Seq Kit, following the gel-free protocol. Following barcoded-adapter ligation, GC-rich and AT-rich DNA samples were divided equally and independently PCR amplified for 14 cycles according to manufacturer's instructions. Each library was prepared in duplicate. Average library size was assessed by Bioanalyzer and concentrations determined by qPCR. Normalized libraries were clustered on-board, and 150 bp paired-end sequencing was performed on the HiSeq 2500 across 4 lanes of 2 flowcells using rapid run mode.

11 different PCR enzymes used in library preparation kits for next generation sequencing were tested in this sequencing experiment. The performance of the three best are highlighted here: NEXTflex PCR polymerase, Supplier D and Supplier I. This subset of enzymes produced the most consistent metrics, including library yield, GC bias and raw coverage, of the 11 PCR kits tested across two genomes. Quality control sequencing metrics against the AT-rich genome show more paired-end reads for the NEXTflex PCR polymerase, and more total alignments and non-duplicated fragments than Suppliers D and I (Table 1).

Results

Library Yield and Complexity

Several metrics, including library yield, were utilized to compare polymerase performance. Achieving sufficient library yields is important in order to meet minimum molarity cutoffs required by different sequencing platforms. Each polymerase produces distinct library quantities (Figure 1). Library yield comparison across the AT-rich genome shows a similar trend: libraries PCR amplified using NEXTflex PCR polymerase or Supplier D's enzyme produce almost two times more library product than Supplier I. Conversely, across the GC-rich genome, NEXTflex PCR polymerase and Supplier I are equal; however, Supplier D library yields are reduced. This result shows the robust ability of the NEXTflex PCR polymerase to equally amplify both GC-rich and AT-rich genomes, and the inconsistent PCR efficiency of Suppliers D and I when using the same template for PCR. Library yields are important, but can give rise to misleading interpretations. Library yield inconsistencies produced by Supplier D suggest an unequal amount of AT-rich over GC-rich clones. Although library yields for Supplier D over the AT-rich genome are roughly equal to library yields of NEXTflex PCR polymerase (Figure 1), sequence complexity and richness (Table 1) are lower, while GC bias is higher and coverage is uneven (Figures 2, 3 and 4). The same trend is evident for Supplier I (Figure 2, 3 and 4).

GC Bias

To obtain a global representation of each PCR enzyme's effect on sequence integrity, we examined PCR-introduced GC bias across both *S. cerevisiae* and *D. radiodurans* genomes using Picard GC bias plots (Figure 2). Low GC bias would produce a normalized coverage slope of one, or a horizontal line over windows of GC-rich content. Alternatively, high GC bias would produce a normalized coverage slope greater than one, over windows of GC-rich content (Figure 2, *S. cerevisiae*, supplier I). Picard analysis shows NEXTflex PCR polymerase to be the most consistently unbiased polymerase tested, regardless of genomic complexity and extreme base composition. Across the *S. cerevisiae* genome, NEXTflex PCR polymerase showed the lowest level of GC bias, with close to mean coverage across the entire genome, including GC-rich and AT-rich stretches. Unlike the NEXTflex PCR polymerase, normalized coverage of Supplier I across the *S. cerevisiae* genome is noticeably biased and uneven, producing uneven coverage of the genome, with fewer reads mapping to GC-rich regions and higher levels of PCR duplication in AT-rich regions. Conversely, the most average normalized coverage across the *D. radiodurans* genome was produced by Supplier I. NEXTflex PCR polymerase and Supplier D had slightly higher GC bias over GC-rich windows of the genome. Overall, however the NEXTflex PCR polymerase provides a balanced representation of both AT-rich and GC-rich regions, which neither Polymerase D or I attains.

Genome Coverage

In addition to GC bias, coverage across average GC-rich and AT-rich regions was considered for each polymerase. Across all four chromosomes of the *D. radiodurans* genome, NEXTflex PCR polymerase consistently outperforms Suppliers D and I, with higher levels of unbiased coverage (Figure 3). The NEXTflex PCR polymerase produced 110-160x coverage of the entire *D. radiodurans* genome compared to 75-110x coverage and 95-130x coverage for Suppliers D and I, respectively. Similar to coverage data of the GC-rich genome, the NEXTflex PCR polymerase shows higher coverage of the first eight chromosomes of the AT-rich genome than both Suppliers D and I (Figure 4; coverage of chromosomes 1-8 shown, 9-16 showed similar results). The NEXTflex PCR polymerase produced 30-40x coverage of the first eight chromosomes of the *S. cerevisiae* genome, compared to 25-35x coverage and 20-25x coverage for Suppliers D and I, respectively.

Conclusions

Our comparison between the NEXTflex PCR polymerase and those of Suppliers D and I found the NEXTflex PCR polymerase, which is used in all of the NEXTflex DNA-seq and RNA-seq NGS library preparation kits, to be the most unbiased enzyme for NGS library prep across two highly diverse genomes, while consistently delivering even and high read coverage. Consistency of library yield between these two genomes demonstrates that this enzyme is extremely robust and can be applied to numerous NGS applications including, but not limited to, WGS, WES, metagenomics, target sequencing and other uses.

Tables and Figures

	NEXTflex PCR Polymerase		Supplier D		Supplier I	
	AT-rich	GC-rich	AT-rich	GC-rich	AT-rich	GC-rich
Total Paired-End Reads	2363781	1725421	2094505	1824620	1575501	1698010
Total Alignments	1964864	1599803	1755982	1688314	1340286	1587546
Non-Duplicated Fragments	1926568	1528556	1727176	1590569	1321802	1541873

Table 1. Sequencing QC metrics represent an average of replicates for NEXTflex™ PCR polymerase and a subset of competing enzymes, Suppliers D and I, across two distinctly different genomes.

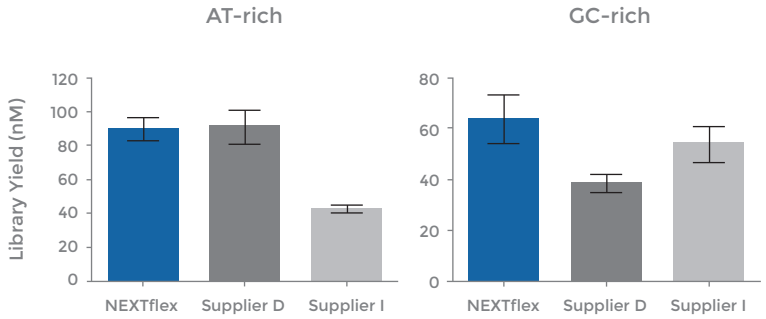

Figure 1. S. cerevisiae (AT-rich) and D. radiodurans (GC-rich) sequencing library yields determined by qPCR.

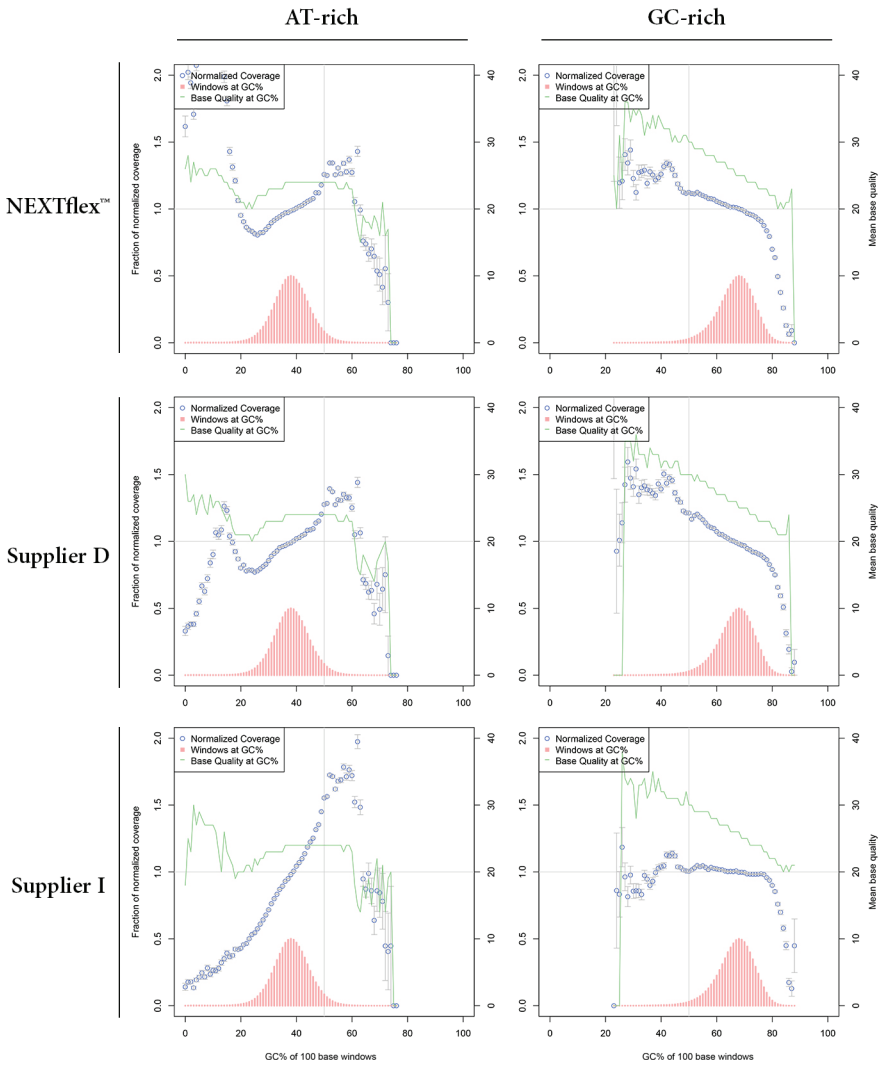

Figure 2. GC bias across AT-rich genome, S. cerevisiae; and GC bias across GC-rich genome, D. radiodurans (left and right panels, respectively) for NEXTflex PCR polymerase, Supplier D and Supplier I. Windows represent GC content of the genome. Mean base quality is determined by error rate of all bases of all reads assigned to windows of GC content. Normalized coverage is the ratio of coverage in this GC bin vs. the mean coverage of all GC bins.

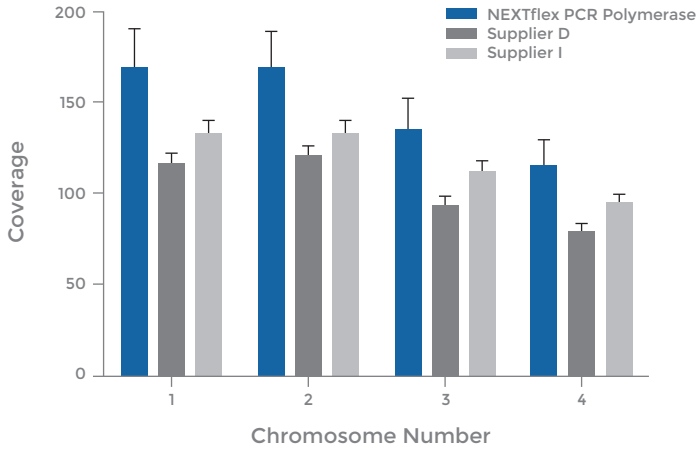

Figure 3. Raw coverage of the four chromosomes of the D. radiodurans genome (GC-rich). Coverage is defined by the number of times each base of the genome was read. Error bars represent standard deviation (n=2).

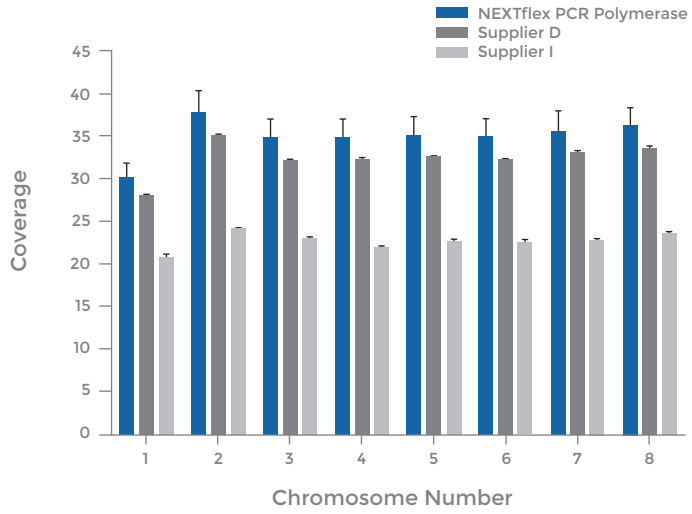

Figure 4. Raw coverage of the first eight chromosomes of the S. cerevisiae genome (AT-rich). Coverage is defined by the number of times each base of the genome was read. Error bars represent standard deviation (n=2).