APPROACHING SINGLE-CELL SEQUENCING BY UNDERSTANDING NGS LIBRARY COMPLEXITY AND BIAS

ABSTRACT

Demands are growing on genomics to deliver higher quality sequencing data from samples with less input quantity. As the number of genomic equivalents decreases at lower input amounts, library bias and library complexity increasingly affect data quality. Less biased, more complex libraries result in more even and complete coverage, resulting in better sequencing efficiency and reduced costs. This application note describes the sequence coverage performance and preservation of molecular complexity of next generation sequencing (NGS) libraries generated from human and microbial genomic DNA using Accel-NGS® 2S DNA Library Kits for whole-genome sequencing (WGS) on the Illumina® platform. Comparisons to the leading commercially available methods are also presented. From the lowest input DNA quantity supported for PCR-free libraries (100 ng for Accel-NGS 2S PCR-free vs.1 µg for the leading kit), this new technology demonstrates more than 50% higher library complexity. Coverage of extreme GC-rich regions is characterized using the 1,000 bad promoters (~79% GC) defined by the Broad Institute. The subsequent sequencing results demonstrate superior coverage performance of these high GC-content promoters, where 88% are covered better with Accel-NGS 2S PCR-free than the market leading kit. Similarly, sequencing analysis of libraries prepared from AT-rich and GC-rich bacterial genomes show excellent coverage distribution.

INTRODUCTION

The degree to which NGS provides accurate and complete genomic information has commonly been determined by assessing the bias associated with sequencing instrumentation and the read chemistries specific to available platforms. As these platforms have improved to reduce sequencing read errors and other sequencing artifacts, library preparation techniques are now playing an increasingly critical role in determining the fraction of the genome captured and presented as sequenceable information. Current library protocols, based on both commercially available kits and home-brewed methods, produce libraries with varying degrees of bias due to the effects of base composition on adapter ligation chemistries[1] and with varying degrees of complexities that are always significantly below the predicted theoretical number of uniquely fragmented inserts[2]. Addressing these sources of data loss specifically for WGS provides a more direct pathway for better genetic analysis by providing more complete coverage of the entire genome. In comparison, targeted sequencing methods, such as exome sequencing, inherently demonstrate less even coverage and require significantly greater depth of coverage for reliable variant calling compared to WGS[3].

Regardless of the enzymology employed, no method is capable of lossless conversion of all starting material into library molecules. This shortcoming is reflected in both the molar yield and complexity yield converted from the input DNA: a molar conversion of 10-20% and a complexity conversion of less than 1% of the starting material are standard for NGS libraries[2]. Methods that maximize the unbiased conversion of unique inserts into functionally adapted library molecules result in higher yields and more complexity, creating a diverse set of library molecules leading to more uniquely mapped reads, better representation of diverse regions, and consequently more uniform sequence coverage. In turn, more uniform coverage reduces or eliminates missing sections of the genome from the sequencing data set while simultaneously leading to lower sequencing costs.

Library complexity depends on factors introduced before the preparation begins, including input quantity of starting material and fragmentation size, as well as factors directly related to the preparation itself, such as efficiency of adapter ligation and the amount of library molecule duplication as a result of PCR amplification, if performed. Complexity of the starting material is calculated based on the Avogadro constant and the fragment size from a given input quantity:

$$\text{Unique Molecules} = \text{Input (g)} \times \frac{\text{mol} \cdot \text{bp}}{660\ \text{g}} \times \frac{1}{\text{fragment size (bp)}} \times \frac{6.022 \times 10^{23}\ \text{molecules}}{\text{mol}}$$

For most microbial genomes and the human genome, the number of duplicates arising due to random fragmentation coincidence is negligible, so this formula estimates starting complexity with sufficient accuracy for these comparisons[2]. Starting complexity is vast in magnitude, with just 100 ng of genomic DNA fragmented to 200 bp producing over 450 billion unique molecules. When sequencing, it is important to maintain a high ratio of the number of unique library molecules to the number of reads in order to avoid representing the same molecule more than once. Therefore, selecting a library preparation method that captures the highest possible complexity is imperative to maximizing data output. **Table 1** illustrates the impact of library complexity on meaningful sequencing coverage using the conversion efficiencies of current leading kits, which is less than 1%. For inputs below 100 ng, fewer than 30 genome equivalents are captured by these preps, so the field standard of 30X sequencing coverage is unattainable without a portion of that fold coverage being comprised of duplicates that do not contribute to uniquely mapped reads.

When limiting input quantity is a factor and PCR-based library amplification is required, a much larger fraction of duplicate library molecules arises, reducing the number of useful sequencing reads and effective coverage, which decreases confidence in variant calling. Additionally, PCR increases the AT and GC bias, reducing the representation of regions that contain high amounts of either of these base combinations. Using an adapter ligation technology that captures greater complexity prior to amplification and maximizes molar yield reduces the number of PCR cycles required and allows sequencing more deeply without saturating read data with duplicates. This preserves more complete coverage of the genome. Recently, the improved performance and availability of high-efficiency, PCR-free library preparation methods has enabled the analysis of library complexity as a function of adapter ligation chemistry at lower input amounts. Here, we demonstrate the excellent PCR-free library complexity preserved by Accel-NGS 2S kits and the impact this preservation of complexity has on coverage of challenging sequences. We also present coverage data for low input samples where library amplification is required.

Table 1: Impact of Library Complexity on Meaningful Depth of Sequencing Coverage

INPUT (ng)	MOLECULAR COMPLEXITY OF INPUT (200 bp FRAGMENTS)	HUMAN GENOME EQUIVALENTS OF INPUT	LIBRARY COMPLEXITY YIELD		LIBRARY GENOME EQUIVALENTS YIELD	
			LEADING KITS (0.25%)	ACCEL-NGS 2S (2.5%)	LEADING KITS (0.25%)	ACCEL-NGS 2S (2.5%)
1000	4.56E+12	166667	1.14E+10	**1.14E+11**	417	**4167**
100	4.56E+11	16667	1.14E+09	**1.14E+10**	42	**417**
10	4.56E+10	1667	1.14E+08	**1.14E+09**	4	**42**
1	4.56E+09	167	1.14E+07	**1.14E+08**	0	**4**

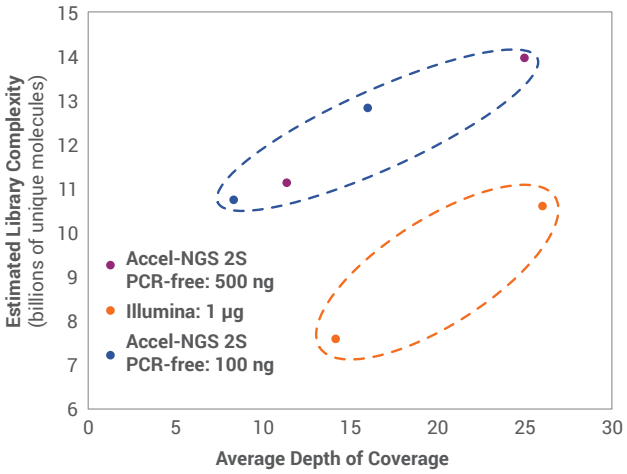
Libraries prepared with the leading commercial kits do not capture enough complexity for meaningful 30X sequencing when starting with less than 100 ng of input material because the number of genome equivalents captured is less than 30. In contrast to leading commercial kits, Accel-NGS 2S kits support one order of magnitude less input for meaningful 30X sequencing because it captures up to a 10-fold greater percentage of the input complexity. Note: complexity conversion rates for the leading kits are based off the 0.25% conversion rate observed for Illumina at 1000 ng and the 2.5% conversion rate observed for the Accel-NGS 2S kits at 100 ng, 10 ng, and 1 ng.

RESULTS

Accel-NGS 2S DNA Library Kits utilize a proprietary adapter attachment enzymology that first maximizes the number of available ends for ligation with two dedicated repair steps that repair both ends of both strands of each DNA fragment. Then, two ligation steps sequentially add adapter sequences to the 5' and 3' repaired ends. This approach leads to highly efficient conversion of insert fragments into library molecules, allowing PCR-free libraries to be generated from 100 ng of starting material, or when ten samples can be pooled, it is possible to use 10 ng of starting material. In addition, the template-independent adapter attachment chemistry results in complex libraries that faithfully represent the base composition of the starting material.

We examined the number of unique library molecules present in Accel-NGS 2S PCR-free libraries and libraries made with Illumina's kit at different input amounts and sequencing depths (**Figure 1**) using Estimate LibraryComplexity, a computational approximation of unique library molecules provided by the Picard MarkDuplicates tool (picard.sourceforge.net). Consistent with the expectation that higher input quantities provide greater starting complexity available for library preparation, we observed a slight increase in library complexity with the Accel-NGS 2S PCR-free kit from 100 ng to 500 ng of input. We also observed an increase in complexity when sequencing the libraries more deeply, indicating improved accuracy of the algorithm with more reads. When controlling for coverage depth, Accel-NGS 2S PCR-free libraries exhibited between 3-5 billion more unique library molecules than Illumina libraries, despite being made from 2-10X less starting material.

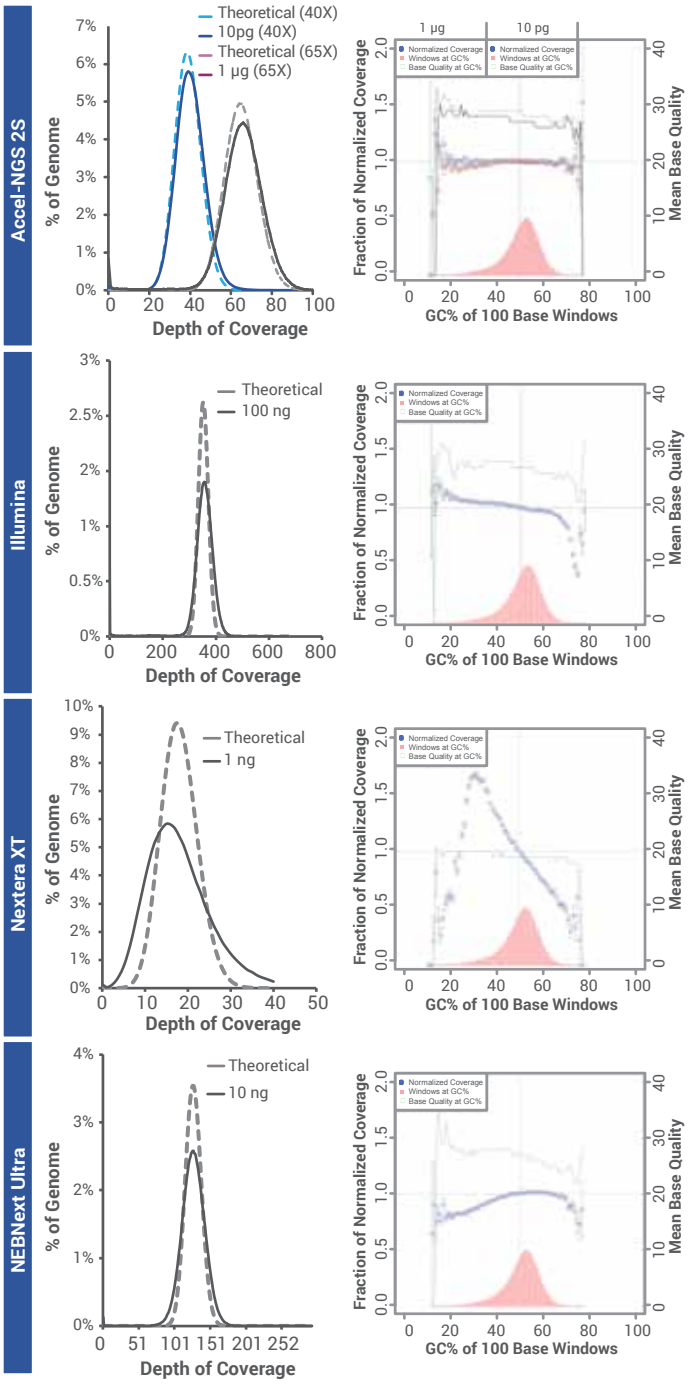
Figure 1: Impact of Library Preparation Method on Library Complexity



Library complexity was obtained at various sequencing depths for Accel-NGS 2S PCR-free libraries compared to libraries made with an Illumina kit. All libraries were made PCR-free from HapMap DNA NA12878 obtained from Coriell and sequenced on the HiSeq®. Estimated library size was calculated by Picard MarkDuplicates (picard.sourceforge.net).

Base composition can affect adapter ligation efficiency, a mechanism that provides insight into the impact of adapter ligation technology on library complexity and evenness of coverage[1]. Library preparation methods commonly struggle to convert fragments with extremely AT- or GC-rich sequences into sequenceable library molecules. To determine genome coverage with respect to base composition, we compared *E. coli* WGS libraries made with Accel-NGS 2S kits vs. Illumina, NEBNext® Ultra,™ and Nextera® XT (**Figure 2**). Despite the balanced composition of the *E. coli* genome, we observed variability in the coverage of GC extremes with the kits tested, in particular noting a significant underrepresentation of GC-rich sequences by the Nextera XT product.

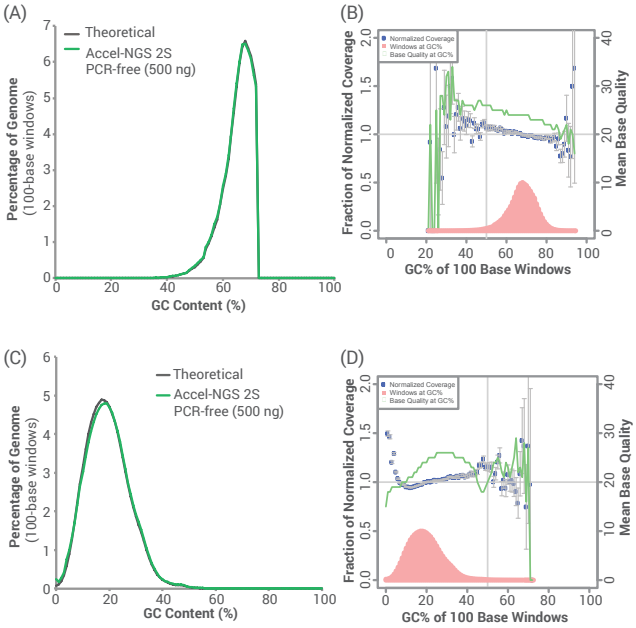
Figure 2: Coverage of a Balanced Microbial Genome with Accel-NGS 2S Kits and Three Other Leading Library Kits



E. coli *WGS libraries were constructed using Accel-NGS 2S Kits at 1 µg PCR-free or 10 pg with 15 cycles of PCR. Coverage was even across GC content and comparable to theoretical distribution, regardless of input. Similar results were observed for 100 pg with 12 cycles of PCR, 1 ng with 9 cycles of PCR, 10 ng with 6 cycles of PCR, and 100 ng PCR-free; data not shown. Libraries were also constructed using Illumina's product at its recommended 100 ng input with PCR, NEBNext Ultra at its recommended 10 ng input with PCR, and Nextera XT used at its recommended 1 ng input with PCR. Sequencing was performed with Illumina MiSeq® V2 reagents. Data was analyzed using BWA (Li and Durbin, 2010) and Picard (picard.sourceforge.net).*

In contrast, Accel-NGS 2S libraries demonstrated coverage across the genome close to the theoretical average coverage predicted by Poisson statistical distribution and relative coverage close to 1 at nearly all GC content. We also received data from an external evaluator who tested the Accel-NGS 2S PCR-free kit with the GC-rich *B. pertussis* genome (68% GC, **Figure 3A** and **3B**) and the extremely AT-rich *P. falciparum* genome (19% GC, **Figure 3C** and **3D**). These results revealed balanced, near-theoretical coverage of these extreme microbial genomes.

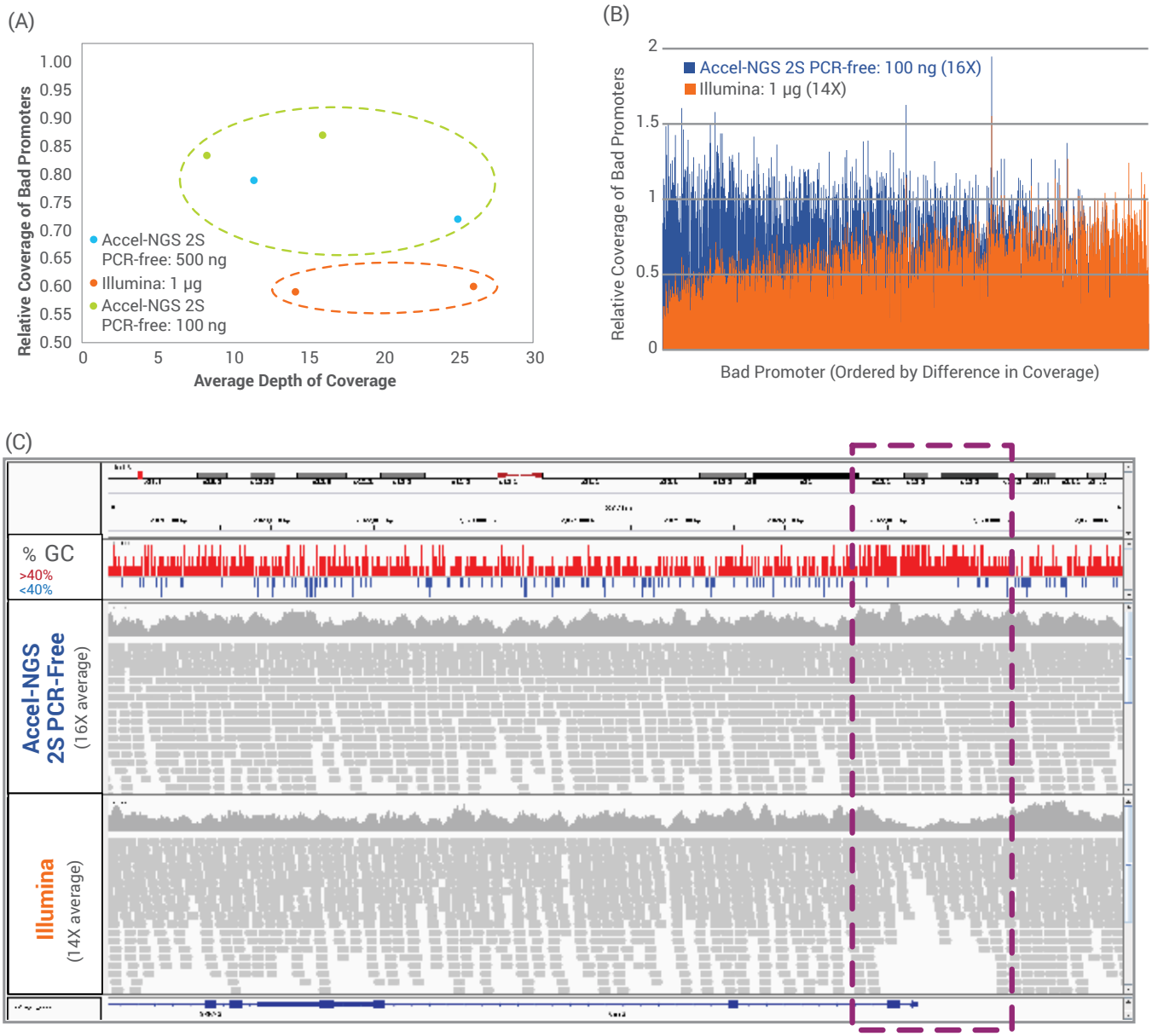
Figure 3: Coverage of the Extreme Microbial Genomes *B. pertussis* (68% GC) and *P. falciparum* 3D7 (19% GC) with the Accel-NGS 2S PCR-free Kit



PCR-free libraries were prepared from 500 ng in each case using the Accel-NGS 2S PCR-free DNA Library Kit. Shown are the GC content of each read vs. theoretical in silico distribution and coverage (a = B. pertussis*, c =* P. falciparum *3D7) and the normalized coverage vs. GC content (b =* B. pertussis*, d =* P. falciparum *3D7).*

For human WGS, transcription start sites or first exons often have poor sequence coverage. Due to high GC content, the Broad Institute defined those with the lowest relative sequence coverage and termed them the "1,000 bad promoters"[1]. These bad promoters are GC-rich, averaging 79% GC composition, relative to the overall 41% GC human genome composition. We observed 20-40% higher relative coverage of the bad promoters by Accel-NGS 2S PCR-free libraries compared to Illumina's libraries, despite being made from 2-10X less input DNA (**Figure 4A**). When comparing coverage of each bad promoter individually, we found that 88.1% of the bad promoters were covered better with the Accel-NGS 2S PCR-free Kit than with Illumina (**Figure 4B**), and zooming in on the TCEB2 locus demonstrated GC-rich areas with more uniform coverage by the Accel-NGS 2S PCR-free Kit (**Figure 4C**).

Figure 4: Relative Coverage of the GC-rich "1000 Bad Promoters" by Accel-NGS 2S PCR-free Libraries Compared to Libraries Made with the Leading Kit



*Relative coverage was plotted as the average relative coverage obtained at various inputs and sequencing depths (**A**) and as the individual relative coverage for each bad promoter when comparing a 100 ng Accel-NGS 2S PCR-free library to a 1 µg Illumina library (**B**). Overall, 88.1% of the bad promoters were covered better with the Accel-NGS 2S PCR-free Kit than with Illumina. Using the Broad Institute's Integrative Genomics Viewer (IGV) to examine the reads on the locus of one of the bad promoters, TCEB2, revealed more consistent coverage with the Accel-NGS 2S PCR-free Kit, particularly in GC-rich areas (**C**). All libraries were made PCR-free from HapMap DNA NA12878 obtained from Coriell and sequenced on the HiSeq.*

CONCLUSIONS

Human WGS libraries made with Accel-NGS 2S library kits use 2-10X less starting DNA and exhibit significantly higher library complexity and coverage of AT- and GC-rich regions than the leading competing kits. **Table 1** illustrates the order of magnitude difference in input level supported for meaningful 30X sequencing of Accel-NGS 2S libraries versus the leading commercial kit. Balanced and extreme AT/GC composition microbial genomes also show excellent evenness of coverage when prepared with Accel-NGS 2S kits relative to other methods.

REFERENCES

1. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013 May 29;14(5):R51.
2. Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D, Ragoussis J, Ponting CP, Fischer MD. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. Genome Res. 2012 Jan;22(1):125-33.
3. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. 2014 Jul 19;15:247.

Swift Biosciences, Inc.
58 Parkland Plaza, Suite 100 • Ann Arbor, MI 48103 • 734.330.2568 • www.swiftbiosci.com