Impact of GC bias on library preparation

Summary

- The proper representation of GC-rich and GC-poor regions is important to understanding the scientific validity of next-generation sequencing (NGS) results.
- PCR amplification in library preparation is often the cause of over- or underrepresenting regions with extreme GC content [1].
- The Invitrogen™ Collibri™ PS DNA Library Prep Kit for Illumina™ sequencers shows balanced GC amplification with minimal loss of GC-rich regions.

Introduction

NGS is becoming a key approach for investigating the molecular basis of diseases because of its sensitivity and specificity. One of the challenges of using NGS is the need for equal coverage of all of the diverse regions in the human genome, especially in regions of extreme GC or AT content. These regions are particularly important because they contain many regulatory elements. To overcome the challenge of appropriate representation, the selection of NGS library preparation materials that accurately cover these difficult regions is crucial.

Preparing nucleic acids for NGS instruments involves a multistep library construction process. In the general workflow, the nucleic acid of interest is harvested, purified, fragmented, end-repaired, and A-tailed; adapters are then ligated, and the libraries are cleaned up, quantitated, normalized, and loaded onto a sequencer (Figure 1).

PCR-free library preparation protocols are usually the preferred method to create libraries that cover the extremes in GC and AT content [1-3]. The main shortcoming of PCR-free libraries is that they require a large amount of starting material, which, in many cases, is not available with precious or highly degraded samples. In these cases, the use of PCR amplification is often required because of the limited amount of starting material. Numerous factors need to be considered to achieve balanced library coverage; these factors can include the PCR enzyme and master mix used, the number of PCR cycles and conditions, and any PCR additives that may be used. Selecting materials that suit the specific needs of the researcher is important.

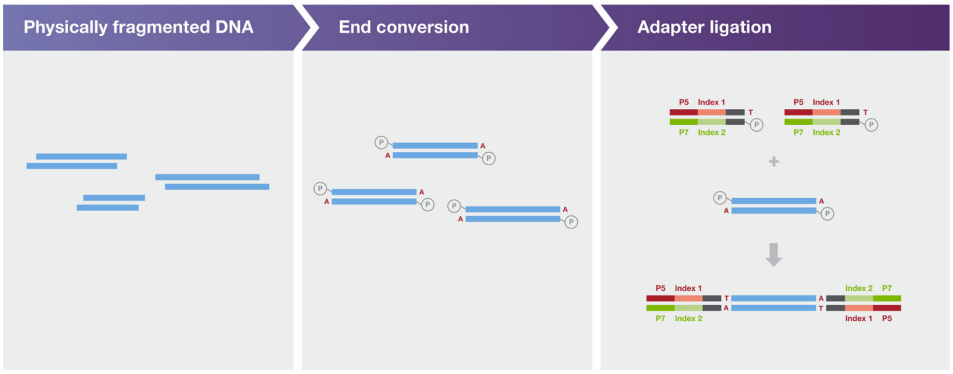

Figure 1. Overview of the Collibri PS DNA library preparation process.

In general, if samples are identified as being challenging, the easiest factors to modulate are the number of PCR cycles, the cycling conditions, and the addition of PCR enhancers. An increase in the number of PCR cycles in a protocol usually increases any bias caused by the PCR enzyme and master mix used; the recommended approach is therefore to minimize the number of cycles while ensuring that sufficient product is loaded onto the sequencer. Four to eight PCR cycles is typical with this approach.

A greater number of cycles and a higher annealing temperature tend to increase the specificity of the amplification protocol; however, the increase in specificity is achieved at the expense of losing regions with extreme AT content. Thus, the use of mid-range annealing and extension temperatures, such as 60°C and 72°C, respectively, is important. The PCR enzyme and master mix used may exert the greatest influence on the GC coverage of the library preparation [1-3].

Methods

Libraries were prepared using four different library preparation kits—the Collibri PS DNA Library Prep Kit for Illumina Systems, and older library prep kits including the KAPA™ HyperPrep Kit, NEBNext™ Ultra II DNA Library Prep Kit, and TruSeq™ DNA Nano Library Preparation Kit. All steps were performed according to the manufacturers' protocols. The NGS libraries were prepared using genomic DNA from the Coriell Institute for Medical Research (accession number NA12878), and with the Horizon™ Quantitative Multiplex Formalin Compromised (Moderate) Reference Standard (Cat. No. HD799). For comparison, 100 ng DNA samples of both types were used. Samples were sequenced on the NovaSeq™ 6000 Sequencing System with an S4 flow cell. GC bias was calculated using Picard tools v2.7.1. Formalin-compromised DNA was converted into sequencing libraries using the manufacturers' recommended protocols and sequenced at 2 x 150 bp using unique dual 8-base indexes for sample identification.

Results

The sequencing run resulted in >85% Q30 bases for both read 1 and read 2. The samples were normalized to 25x coverage and analyzed. The resulting libraries were analyzed for specific GC content as described [2]. In general, the Collibri and KAPA kits gave the most consistent coverage of extreme GC regions (Figure 2). The Collibri kit also showed the highest mean coverage for “bad promoters” and areas of the genome with greater than 75% GC content. Overall, the Collibri library preparation kit resulted in consistent, precise genomic coverage.

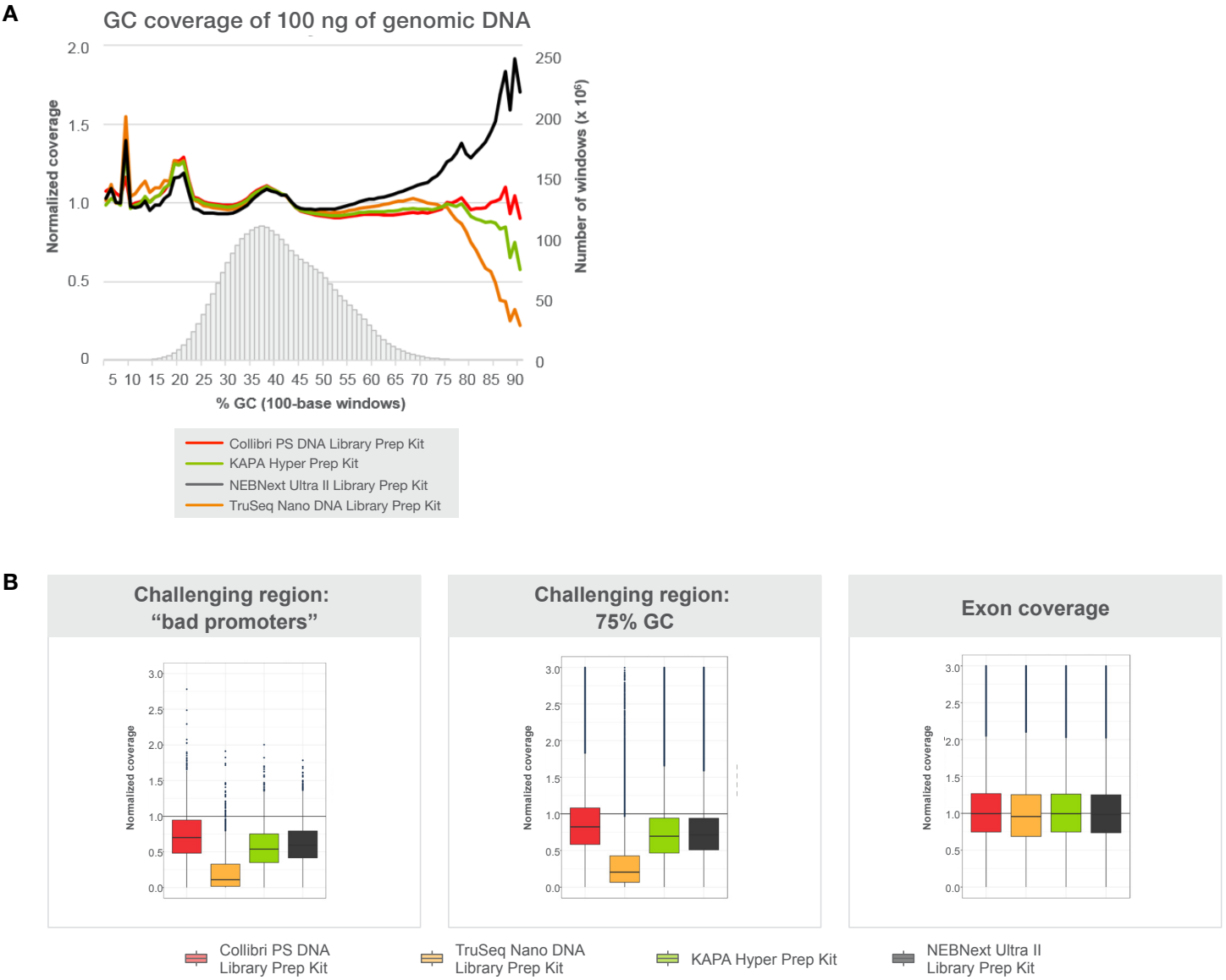

Figure 2. Graphs showing normalized coverage of challenging regions in the human genome. (A) Normalized coverage of the percent of mapped GC content of 100 ng of Coriell NA12878 DNA. **(B)** Coverage of promoters with challenging GC content and 75% GC content in Horizon FFPE DNA is higher and more even with the Collibri PS DNA Library Prep Kit. “Bad promoters” include 1,000 GC-rich human promoters that are exceptionally resistant to sequencing [2].

Conclusions

- Balanced coverage of the diverse regions in the human genome is important for the investigation of the molecular basis of diseases.
- PCR amplification of libraries prepared with older library prep technology can lead to coverage bias in regions with extreme GC and AT content.
- The Collibri PS DNA Library Prep Kit enables consistent genomic coverage.

Ordering information

Product		Quantity	Cat. No.
DNA-Seq kits for Illumina systems			
Collibri ES DNA Library Prep Kits	with CD Indexes	24 preps	A38605024
	with CD Indexes	96 preps	A38607096
	with UD Indexes, Set A (1-24)	24 preps	A38606024
	with UD Indexes, Set B (25-48)	24 preps	A43605024
	with UD Indexes, Set C (49-72)	24 preps	A43606024
	with UD Indexes, Set D (73-96)	24 preps	A43607024
Collibri PCR-Free ES DNA Library Prep Kits	with CD Indexes	24 preps	A38545024
	with CD Indexes	96 preps	A38603096
	with UD Indexes, Set A (1-24)	24 preps	A38602024
	with UD Indexes, Set B (25-48)	24 preps	A43602024
	with UD Indexes, Set C (49-72)	24 preps	A43603024
	with UD Indexes, Set D (73-96)	24 preps	A43604024
Collibri PS DNA Library Prep Kits	with CD Indexes	24 preps	A38612024
	with CD Indexes	96 preps	A38614096
	with UD Indexes, Set A (1-24)	24 preps	A38613024
	with UD Indexes, Set B (25-48)	24 preps	A43611024
	with UD Indexes, Set C (49-72)	24 preps	A43612024
	with UD Indexes, Set D (73-96)	24 preps	A43613024
	with UD Indexes, Set A-D (1-96)	96 preps	A38614196
Collibri PCR-Free PS DNA Library Prep Kits	with CD Indexes	24 preps	A38608024
	with CD Indexes	96 preps	A38610096
	with UD Indexes, Set A (1-24)	24 preps	A38609024
	with UD Indexes, Set B (25-48)	24 preps	A43608024
	with UD Indexes, Set C (49-72)	24 preps	A43609024
	with UD Indexes, Set D (73-96)	24 preps	A43610024
	with UD Indexes, Set A-D (1-96)	96 preps	A38615196

CD = combinatorial dual, UD = unique dual

invitrogen

Ordering information (continued)

Product	Quantity	Cat. No.
RNA-Seq kits for Illumina systems		
Collibri Stranded RNA Library Prep Kit for Illumina Systems	24 preps	A38994024
	96 preps	A38994096
Collibri Stranded RNA Library Prep Kit for Illumina Systems with H/M/R rRNA Depletion Kit	24 preps	A39003024
	96 preps	A39003096
ERCC RNA Spike-In Mix	1 kit	4456740
ERCC ExFold RNA Spike-In Mixes	1 kit	4456739
Library quantification		
Collibri Library Quantification Kit	100 rxns	A38524100
	500 rxns	A38524500
Qubit 4 Fluorometer, with WiFi	1 fluorometer	Q33238
Qubit 4 NGS Starter Kit, with WiFi	1 kit	Q33240
Library amplification		
Collibri Library Amplification Master Mix	50 rxns	A38539050
	250 rxns	A38539250
Collibri Library Amplification Master Mix with Primer Mix	50 rxns	A38540050
	250 rxns	A38540250

H/M/R = human/mouse/rat

References

1. Aird D, Ross MG, Chen WS et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18.
2. Ross MG, Russ C, Costello M et al. (2013) Characterizing and measuring bias in sequencing data. *Genome Biol* 14:R51.
3. Chen Y-C, Liu T, Yu C-H et al. (2013) Effects of GC bias in next-generation sequencing data on *de novo* genome assembly. *PLoS One* 8:e62856.

Find out more at **thermofisher.com/collibri**

ThermoFisher SCIENTIFIC