Scaling library preparation and maintaining quality

Summary

- When scaling up next-generation sequencing (NGS) library preparation methods, it is critical to maintain even coverage and minimize bias.
- Various factors affect library quality, including input DNA quantity and quality, fragmentation methods, enzymatic biases, and PCR cycles.
- GC bias is the biggest contributor to uneven coverage, but it can be minimized by using an optimal PCR master mix or PCR-free library preparation methods.

Introduction

The rapid adoption of next-generation sequencing (NGS) continues to open up new avenues of research, from single-cell genomics to metagenomic studies of complex environmental communities. The amount of input DNA available for library preparation varies widely, as does its quality. For example, single-cell whole-genome studies necessitate working with sub-nanogram amounts of DNA, typically requiring a preamplification step before library preparation. DNA from archived tumor samples, especially formalin-fixed, paraffin-embedded (FFPE) tissues, is highly degraded and poses unique challenges for library preparation. Therefore, it is unrealistic to expect that a single library preparation workflow will accommodate all types of samples. Current library preparation methods can be divided into two broad types: PCR-based and PCR-free. To generate high-quality data, both approaches must enable the construction of high-diversity libraries across a wide range of input levels.

An ideal library will have uniform coverage across the spectrum of GC content. Several factors affect library quality at low input amounts. The typical steps of DNA fragmentation, end repair, A-tailing, and adapter ligation all contribute to the quality of the sequencing library. The effect of these steps on library quality, independent of subsequent PCR amplification, can be assessed by the conversion rate: the percentage of input DNA that is converted to adapter-ligated fragments that can be sequenced successfully. As the conversion rate increases, the number of required PCR cycles in the subsequent amplification step is reduced, thereby reducing potential bias.

GC bias

GC bias is the most common problem introduced by PCR, and it can pose serious challenges for data analysis. In microbiome studies that rely on whole-genome sequencing of complex communities, a comparison of PCR-based and PCR-free library preparation methods showed significant variation in taxonomic representation [1]. Further, GC bias hinders *de novo* genome assembly and single-nucleotide polymorphism (SNP) discovery [2]. In addition to GC bias, amplification-related artifacts can contribute to reduced library quality. These artifacts arise from PCR stochasticity, template switching, and polymerase errors [3].

It should be noted that, even with PCR-free library preparation methods, bias can be introduced during cluster amplification and even from the sequencing chemistry itself [1]. Continual refinement of these steps by sequencing instrument manufacturers will help to minimize these sources of bias.

PCR-free workflow

Ideally, a sequencing library should reflect the diversity of the original DNA sample by representing all regions of the genome equally. Amplification bias results in over- or underrepresentation of GC-rich regions, thereby reducing library diversity; in extreme cases, some regions of the genome may be completely absent in the final sequencing data. Due to the problems associated with PCR amplification outlined above, considerable attention has been devoted to developing efficient and scalable PCR-free library preparation methods.

The advantages of a PCR-free workflow are significant. Since PCR amplification is the main contributor to GC bias, eliminating PCR during library preparation results in vastly improved library diversity. This improvement in library quality has made PCR-free library preparation methods popular in studies of population-scale genomics and human disease. However, the main disadvantage of PCR-free library preparation methods is their requirement for higher input amounts compared to PCR-based methods. They can also pose challenges for quality control and quantitation [4]. Optimization of PCR-free methods has lowered the input requirements. For example, Invitrogen™ Collibri™ PCR-free DNA library preparation kits require 500 ng input for physically sheared DNA and 100 ng input for enzymatically sheared DNA.

PCR workflow

To address the bias concerns with typical PCR-based library preparation, Thermo Fisher Scientific has developed a master mix optimized for NGS library amplification. The Invitrogen™ Collibri™ Library Amplification Master Mix (2X) is a ready-to-use solution designed for the amplification of NGS libraries compatible with Illumina™ sequencing platforms. The master mix includes Invitrogen™ Platinum™ SuperFi™ DNA Polymerase in combination with an optimized reaction buffer that contains all the necessary components for efficient and uniform library amplification, regardless of GC content, helping to improve coverage across GC-rich sequences and other complex regions. The master mix helps preserve data quality across a range of input amounts and number of PCR cycles. Thus the Invitrogen™ Collibri™ DNA library preparation kits require only 1 ng of sample input for physically or enzymatically sheared DNA.

Methods

Libraries were prepared using the following kits for physically sheared DNA: the Collibri Physically Sheared (PS) DNA Library Prep Kit, KAPA™ HyperPrep Kit, and NEBNext™ Ultra II DNA Library Prep Kit. All steps were performed according to the manufacturers' protocols. The Collibri DNA libraries were prepared using the Collibri Library Amplification Master Mix. The NGS libraries were prepared using the reference genomic DNA from the Coriell Institute for Medical Research (accession number: NA12878). Corriell NA12878 samples were sequenced on an Illumina™ NovaSeq™ 6000 system using a NovaSeq™ S4 flow cell. Reads were normalized to 186M for physically sheared DNA and sequenced at 2 x 150 bp. GC bias was calculated using Picard tools v2.7.1.

Results

The combination of the optimized PCR master mix and Collibri library prep kit produced libraries with highly uniform coverage across the GC spectrum, compared to other methods (Figure 1). The Collibri DNA library prep kit, containing Collibri Library Amplification Master Mix, provided comparatively even coverage across a broad range of GC content, while the other library preparation kits showed significant variation in coverage between the two workflows.

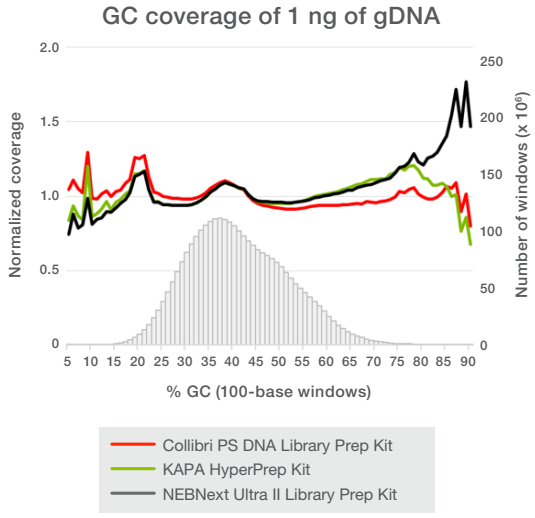

Figure 1. Consistent coverage across GC content with the Collibri PS DNA Library Prep Kit. The Collibri PS DNA library preparation kit, containing Collibri Library Amplification Master Mix, provided consistent, uniform coverage across GC content. Libraries were prepared following the manufacturers' instructions with the recommended PCR enzymes. Only Collibri DNA library preparation kits use Collibri Library Amplification Master Mix containing Platinum SuperFi DNA Polymerase. Libraries were sequenced on a NovaSeq 6000 system using a NovaSeq S4 flow cell.

invitrogen

Conclusions

- GC bias is a significant concern when using standard PCR-based library preparation methods.
- Collibri DNA library preparation kits, in combination with Collibri Library Amplification Master Mix, enable uniform coverage across a wide range of GC content.

References

1. Jones MB, Highlander SK, Anderson EL et al. (2015) Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci USA* 112:14024-14029.
2. Aird D, Ross MG, Chen W-S et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18.
3. Kebschull JM, Zador AM (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* 43:e143.
4. UC Davis Genome Center. NGS Library Construction. **genomecenter.ucdavis.edu**

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
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

CD = combinatorial dual, UD = unique dual, H/M/R = human/mouse/rat

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

ThermoFisher SCIENTIFIC