Improved Method for Reduced Representation Bisulfite Sequencing (RRBS)

Ben Schroeder, Mike Phelan, Lin Pham, Steve Kain, and Doug Amorese, NuGEN Technologies, Inc., San Carlos, CA, USA

INTRODUCTION

DNA methylation is a central component of epigenetic regulation, and has been found to play a role in development, environmental exposure, and diseases including cancer. In humans, the predominant form of DNA methylation is 5-methyl cytosine. When present, 5-mC occurs almost exclusively in the context of the CpG sequence motif. Reduced Representation Bisulfite Sequencing (RRBS) utilizes the restriction enzyme MspI, which recognizes and cleaves the sequence CCGG regardless of the methylation state of the central CpG.

Fig. 1

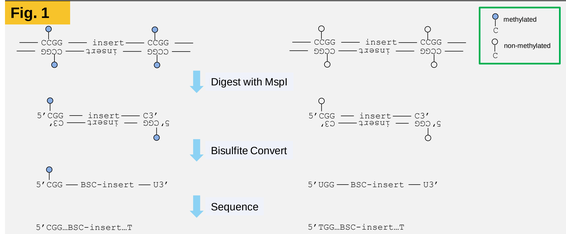


Figure 1: The basic principle of RRBS

Shown are two genomic regions, each flanked by MspI sites. The region on the left is methylated (5-mC shown with blue circle), while the region on the right is non-methylated. Genomic DNA is digested with MspI, bisulfite converted (a chemical reaction that converts non-methylated C to U, but leaves 5-mC unreacted), and sequenced. C in the sequence indicates the presence of 5-mC at that position. Note how the first base of every sequence contains a methylation measurement.

Reducing Representation to ~2%

After digestion with MspI, only ~2% of the genome is represented in fragments under 500 bp. However, many of these fragments are from CpG-rich regions. By focusing RRBS sequencing on small fragments, the methylation status of these important regions can be determined with much less sequencing than would be required using a whole genome bisulfite sequencing approach.

Fig. 2

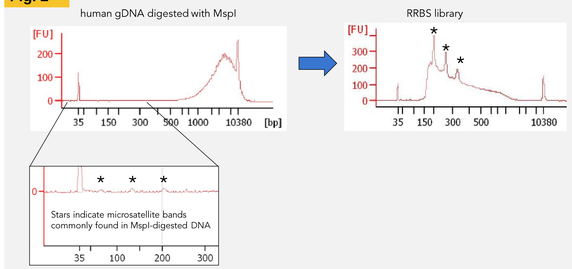


Figure 2: Bioanalyzer traces of MspI-digested human genomic DNA (left), and the resulting RRBS sequencing library (right).

Fig. 3

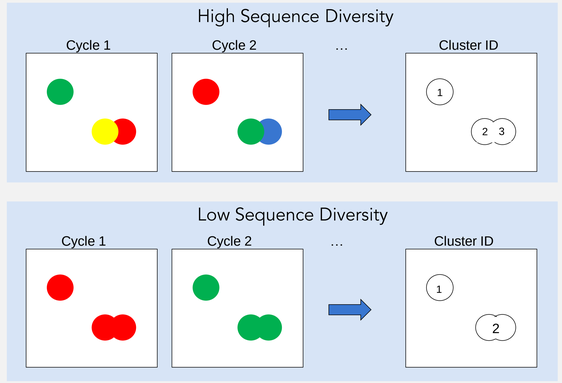


Figure 3: Cluster Identification Requires Sequence Diversity

The Illumina software identifies clusters over the first several cycles of sequencing. During sequencing of normal, high diversity clusters (top), overlapping clusters can be distinguished because they are different colors. If overlapping clusters contain the same sequence (bottom), they may be mistaken as a single cluster.

Challenges with sequencing RRBS libraries

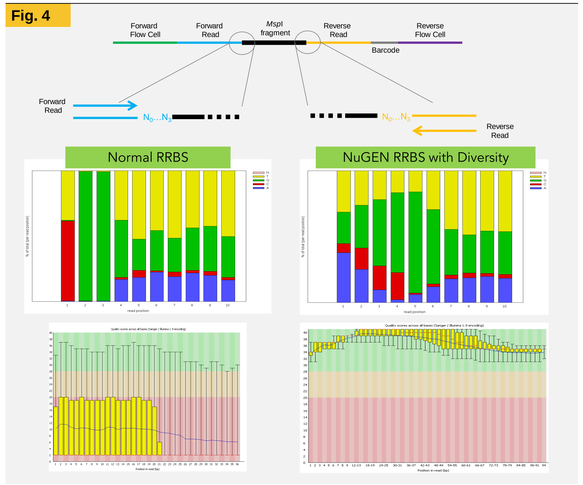
On Illumina sequencers, cluster identification as well as phasing and color matrix calculations can be negatively impacted by a lack of sequence complexity in the sample. RRBS libraries all begin with YGG (Y = C or T), making these libraries especially challenging to sequence. These issues can be mitigated by reducing cluster density and mixing in a high diversity library, such as a PhiX control library. Unfortunately, this approach will reduce the number of useful reads as PhiX levels of 40% or higher may be needed to ensure a successful run.

Table 1

	CpG Island	CpG Shore	Promoter	Genome Wide
Total CpG's in class	4,179,076	4,575,434	8,296,598	56,434,686
50nt SE RRBS	1,452,825	606,959	1,564,822	3,235,820
100nt SE RRBS	2,187,671	1,003,815	2,431,738	5,715,490

Table 1: CpG loci reachable by single end 50 or 100nt RRBS reads. The all CpG's in the hg19 human reference genome were assigned to the following classes: CpG Island (CpG-rich regions greater than 200bp, as defined in USCS Genome Browser), CpG Shore (2000bp on either side of a CpG Island), Promoter (1kb on either side of an Ensembl Transcription Start Site). Shown are the number of CpG's reachable by performing a 50nt or 100nt single end RRBS sequencing experiment. RRBS analysis considers uniquely mapping reads (Bismark) from 40-300bp MspI fragments and assumes 0% methyl-C and 100% bisulfite conversion.

Fig. 4



Adding Diversity to RRBS Reads

The challenges associated with sequencing RRBS libraries can be overcome by adding diversity in the form of 0 to 3 bases of sequence between the sequencing primer and the insert. This ensures all 4 bases are present during the critical first few cycles. In addition, the YGG signature is de-phased so that no cycle contains only G across all clusters.

Fig.5

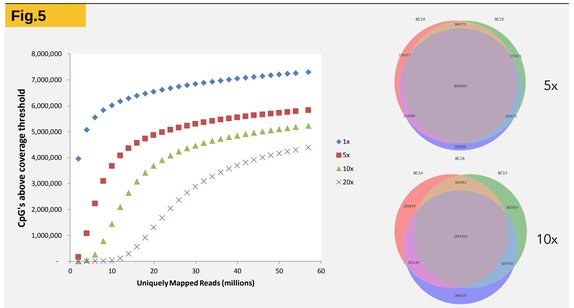


Figure 5: Saturation plots (left) and Venn diagrams (right) showing the number of CpG loci covered by the NuGEN RRBS system at the indicated depth. Venn diagrams calculated using 15 million reads from each barcoded replicate.

Table 2

	Replicate 1	Replicate 2	Replicate 3
Uniquely Mapping	67.60%	67.47%	66.98%
Non-Uniquely Mapping	23.12%	22.70%	23.93%
Total Mapping	90.72%	90.18%	90.91%
Methyl CG	44.05%	44.14%	43.17%
Methyl CHH	0.31%	0.31%	0.30%

Table 2: Bismark statistics for replicate NuGEN RRBS libraries from 25 ng IMR90 gDNA.

Fig.6

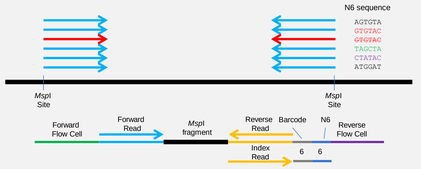


Figure 5: Random N6 sequence adjacent to the barcode can be used to mark and remove PCR duplicates.

Fig. 7

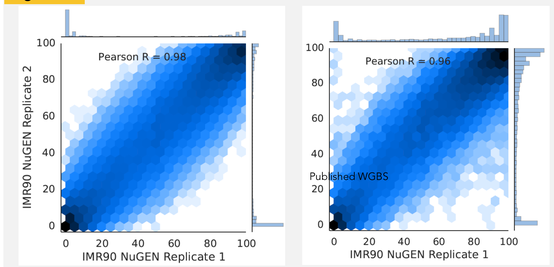


Figure 7: Concordance in methylation levels for CpG's covered at 20X or greater depth between NuGEN RRBS technical replicates (left) or between NuGEN RRBS (30 M reads) and published whole genome bisulfite sequencing (WGBS) from Lister et al., *Nature*, 2009, 462:315-322 (1180 M reads, right).

CONCLUSIONS

NuGEN has developed an enhanced method of Reduced Representation Bisulfite Sequencing that overcomes the challenges of sequencing on Illumina platforms. The method can be multiplexed up to 16 samples per lane, requires no gel purification steps, and has many advantages, including:

- Simplified workflow that can be completed in a single day
- Novel adaptor features which provide enhanced sequence quality without PhiX spike in, and the ability to identify and remove PCR duplicates
- Highly reproducible data that displays excellent concordance with published methylation profiles

NuGEN
imagine more from less®

www.nugen.com