Genome-wide methylation data analysis on the SOLiD™ System

George Marnellos, Daniel Krissinger, Gavin D. Meredith, Miroslav Dudas, Kevin Clancy, Christopher Adams,
Life Technologies / R&D, 5791 Van Allen Way, Carlsbad, CA, USA, 92008.

ABSTRACT

DNA methylation is an epigenetic modification crucial for organism development and normal gene regulation; aberrations in methylation are, among others, characteristic of many cancers in mammals. Next-generation sequencing (NGS) technologies are enabling new methods for methylation profiling. Life Technologies has introduced a versatile methyl-CpG binding protein-based system (MethylMiner™) for the enrichment of methylated sequences from genomic DNA[1]. This enrichment step, along with the use of SOLiD™ System sequencing, allows for focused evaluation of genome-wide methylation patterns. This approach is an efficient and cost effective alternative to shotgun bisulfite sequencing of the entire genome to interrogate methylation marks, as only about 1% of the human genome is methylated and requires interrogation. Here we describe a comprehensive workflow for mapping and analyzing MethylMiner™-enriched fractions of genomic DNA as well as bisulfite converted reads sequenced on the SOLiD™ System, that employs freely-available public tools (e.g. Bowtie[2], SAMtools[3], MACS[4]) and our own scripts and programs. The workflow enables characterization of methylation patterns at different levels of resolution, from broad genome region comparisons and profile differences between samples to individual methyl C resolution. It provides the following functionality:

- Mapping of unconverted and bisulfite-converted reads
- Filtering of clonal reads
- Mapping statistics: distribution of reads on chromosomes, coverage and read depth statistics, C and CpG counts in mapped reads
- Methylation analysis: methylation status of C residues in various sequence contexts, and bisulfite conversion efficiency
- Peak-finding in MethylMiner™-enriched reads
- Level of enrichment in various genome regions (exons, introns, CpG islands, repeats, etc.)
- Visualization of mapped reads and MethylMiner™-enriched peaks with publicly available genome browsers (e.g. the UCSC or IGV browser).

We have implemented this analysis pipeline to analyze human data sets (IMR90 and MCF-7 cell lines), mainly MethylMiner™-enriched fractions, bisulfite-converted and unconverted reads. Results showed good agreement with publicly available methylation data: peaks in MethylMiner™ -selected reads have high coverage of genome regions with higher densities of published methyl-CpGs. This analysis workflow is a convenient and flexible solution for users which will allow the integration of methylation data with results from other SOLiD™ System applications (e.g. ChIP-seq and RNA-seq).

INTRODUCTION

DNA methylation is an epigenetic modification crucial for development and normal gene regulation. Aberrations in methylation are characteristic of many cancers in mammals.

Next-generation sequencing (NGS) technologies are enabling new methods for methylation profiling.

Life Technologies has introduced a methyl-CpG binding protein-based system, MethylMiner™, for the enrichment of methylated sequences from genomic DNA.

MethylMiner™ allows for focused evaluation of methylation patterns in genome-wide studies. This is more efficient than bisulfite conversion and sequencing of the entire genome, as only about 1% of the human genome is methylated.

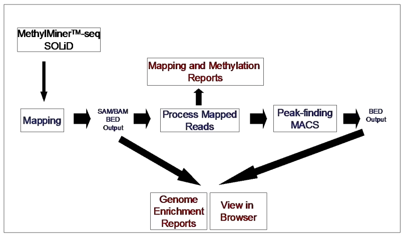
The purpose of the work described here has been to:

- Provide users with a workflow for mapping and analyzing MethylMiner™-enriched and unenriched genomic DNA sequenced on the SOLiD™.
- Enable characterization of methylation patterns at different levels of resolution, from broad genome region comparisons and profile differences between samples to individual methyl C resolution.
- Allow comparison of results with other SOLiD™ applications (e.g. ChIP-seq and RNA-seq).

DATA ANALYSIS PIPELINE

The analysis pipeline we have put together uses freely-available public tools, existing SOLiD™ System software and new auxiliary scripts and programs[5]. The pipeline functionality includes (see Figure 1):

Figure 1. Methylation data analysis pipeline



- **Mapping of unconverted and bisulfite-converted reads,** with existing color-space mapping programs
- **Filtering out of clonal reads** to eliminate amplified copies of a fragment/read mapping to the exact same genome position and strand as the original fragment.
- **Mapping statistics** including mapping rate, distribution of reads per chromosome and statistics on genome coverage and depth (see Tables 1 and 2, and Figs. 2 and 7).
- **Methylation analysis** at nucleotide resolution, reporting methylation status of C residues in CG and CH sequence contexts; analysis odf bisulfite conversion efficiency from mapping of control sequences.
- **Peak-finding** in MethylMiner™-enriched mapped reads, with unenriched control reas.
- **Enrichment of mapped reads** in various genome regions (exons, introns, CpG islands, repeats, etc.). See Fig. 6.
- **Viewing of mapped reads and derived peaks:** pipeline output files can be imported and viewed in genome browsers, e.g. the UCSC or IGV browser (see Fig. 3).

RESULTS

We used the analysis pipeline on human (shown below) and non-human data (mouse, Arabidopsis; not shown), bisulfite-converted and unconverted, MethylMiner™-enriched or unenriched. Results were in very good agreement with publicly available methylation data: among others, peaks in MethylMiner™-selected reads had high overlap with genome regions like CpG islands (see Table 1 and Fig. 5), and methyl-Cs extracted from unenriched bisulfite reads had excellent concordance with methyl-C's published by Lister *et al.* 2009[6] (see Table 2).

Figure 2. Methylation analysis pipeline output examples.
Data in this figure are from a MethylMiner™ 2000mM fraction of unconverted reads of the human MCF7 breast cancer cell line .
A. Distribution of uniquely mapped non-redundant reads on the chromosomes. **B**. Percent of genome covered at various read depths with non-redundant reads.

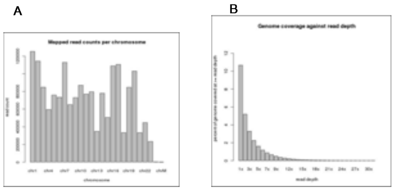


Figure 3. Examples of mapped reads and peaks of MethylMiner™-enriched unconverted human MCF7 reads viewed in the UCSC browser

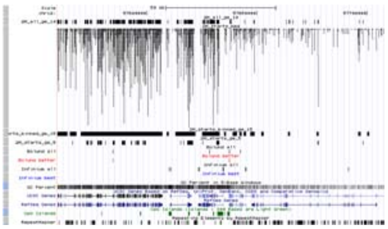


Figure 4. Higher MethylMiner™ salt fractions yield higher methyl-CpG regions.
Data shown are from human IMR90 fetal lung fibroblast cell line reads. Counts of the number of methyl CpGs reported by Lister *et al*. 2009[6] within 150 bp downstream of uniquely-mapped non-redundant reads show that different subsets of the genome are obtained with different MethylMiner™ fractions.

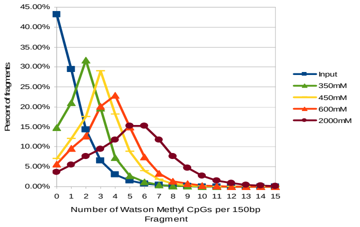


Table 1. Human IMR90 MethylMiner™ fraction mapping statistics and coverage of CpG islands

Fraction	Total reads	# uniquely mapped	% uniquely mapped	# unique starts	% CpG Islands covered (of 28,691)
input	86,033,985	52,804,281	61.40%	48,126,872	87.42%
350 mM	86,942,318	44,292,234	50.90%	37,632,630	48.19%
450 mM	83,790,547	40,991,116	48.90%	34,193,582	60.60%
600 mM	84,897,767	40,368,266	47.60%	33,372,789	55.34%
2 M	84,851,179	27,002,327	31.80%	21,223,792	51.16%

Figure 5. CpG island coverage by human IMR90 MethylMiner™ fractions (Table 1).

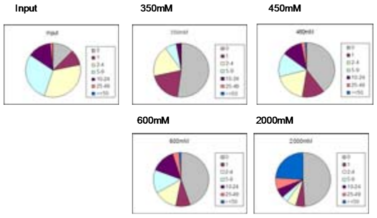


Figure 6. Human IMR90 MethylMiner fraction enrichment for various genome features (fold over whole-genome)
Differential enrichment for various, partially overlapping, annotated genomic features is obtained with MethylMiner™ fractions. CpG islands and shores, and exon sequences increase in relative representation in higher salt fractions.

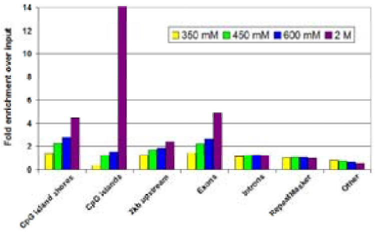


Figure 7. Human IMR90 genome coverage with whole-genome bisulfite reads.
Graph shows how coverage increases with increasing read input size.

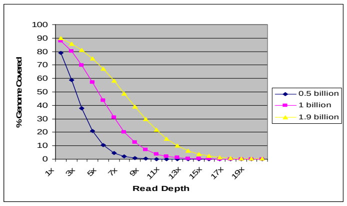


Table 2. Concordance of methyl-CpGs observed in human IMR90 whole-genome bisulfite reads with published data.
Bisulfite sequencing of IMR90 yields methyl-CpGs in high concordance with those reported by Lister *et al*. 2009[6] ("Salk mCG's" in table).

Total Reads	# Uniquely Mapped	Uniquely Mapped (%)	total mCG's	Salk mCG's	Concordance with Salk mCG's	% Salk mCG's
495,265,636	151,321,772	30.55%	5,632,473	5,339,933	94.81%	23.57%
515,699,253	160,975,121	31.21%	6,096,313	5,770,102	94.65%	25.47%
1,010,964,889	312,296,893	30.89%	12,359,341	11,692,335	94.60%	51.61%
1,911,550,949	524,998,672	27.46%	16,354,550	15,357,420	93.90%	67.79%

CONCLUSION

We have presented a simple and flexible pipeline for mapping and analyzing methylation data on the SOLiD™ system, focusing on MethylMiner™ enriched reads. The pipeline enables characterization of methylation patterns at different levels of resolution, from broad genome region comparisons and profile differences between samples to individual methyl-C resolution. We have successfully used the pipeline on human reads (and data from other organisms like mouse and Arabidopsis), with remarkably good concordance with published methylation data.

REFERENCES

1. http://www.invitrogen.com/methylation
2. B. Langmead *et al*., *Genome Biology* **10**, R25 (2009).
3. H. Li *et al*., *Bioinformatics* **25**, 2078 (2009).
4. Y. Zhang *et al*., *Genome Biology* **9**, R137 (2008).
5. http://solidsoftwaretools.com
6. R. Lister *et al*., *Nature* **462**, 315 (2009).

ACKNOWLEDGEMENTS

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.