An end-to-end workflow for accurate methylation detection

Lydia Bonar, Kristin Butcher, Michael Bocek, Holly Corbitt, Bryan Hoglund, Cibelle Nassif, Patrick Cherry, Derek Murphy, Jean Challacombe, Esteban Toro

TWIST BIOSCIENCE

Abstract

Introduction DNA methylation at CpG nucleotide sites in eukaryotes is a key epigenetic mark that helps regulate gene expression. Specific changes in CpG methylation occur in many human cancers, making them a promising biomarker for early cancer detection. However, existing assays can be costly, lack specificity to regions of interest and often provide only semi-quantitative estimates of the methylation fraction. Here, we present new tools to address these challenges including the use of a targeted methylome panel to decrease costs associated with next generation sequencing (NGS), methylation controls to calibrate quantitative assays and UMIs for accurate deduplication in low-diversity samples.

Experimental Procedures Genomic DNA (gDNA) was prepared for sequencing using the Twist Methylation Detection System consisting of enzymatic methyl conversion library generation and hybrid capture using the Twist Human Methylome Panel. Twist CpG methylation specific controls were spiked into gDNA with different methylation rates prior to library preparation and taken through the Twist Methylation Detection System using homologous panels to demonstrate how these controls can be used to calibrate detection assays. Additionally, libraries were generated using cell free DNA (cfDNA) and either conventional or UMI-containing adapters during the library preparation ligation step to investigate the impact on quantitative detection and total unique coverage.

Data Sequencing the Twist Human Methylome Panel to 100-250X raw coverage achieves uniform coverage with low off-bait Picard metrics. 6.59 million CpG sites were detected at a minimum depth of 10X. Methylome target capture allows for informative CpG calling of up to 196 samples on a single Illumina Novaseq S4 flowcell. In contrast, performing traditional whole genome bisulfite sequencing (WGBS) without target enrichment, only 28 samples could be run on the same flowcell. The Twist CpG methylation specific controls are constructed of 48 unique contrived sequences that contain a total of 8 different levels of methylation, ranging from 100% to 0%. Including these controls allows for quantitation of methylation levels in the experimental samples and qualification of the enzymatic conversion process.

Conclusions Our study uses several new products found in the Twist methylation detection portfolio that interrogate genome-wide methylation patterns for various applications. Not only have we found ways to better control assays, but we have also determined methods that lower costs compared to WGBS.

Twist Methylation Detection System

The Twist Methylation Detection System consists of (1) library preparation with enzymatic conversion and (2) target enrichment with a methylation panel such as the Twist Methylome Panel (**Figure 1**). Conventional library preparation protocols convert unmethylated cytosines to uracils using bisulfite, which causes unwanted DNA breaks that complicate downstream sample preparation and, methylation detection. Twist Bioscience partnered with New England Biolabs to offer the NEBNext Enzymatic Methyl-seq (EM-seq) kit as part of the Twist Methylation Detection System. This innovative process accomplishes the same conversion results as bisulfite treatment without the harshness of chemical conversion, yielding a superior end result.

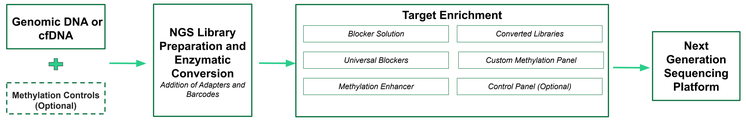

Figure 1. Twist Methylation Detection System Overview

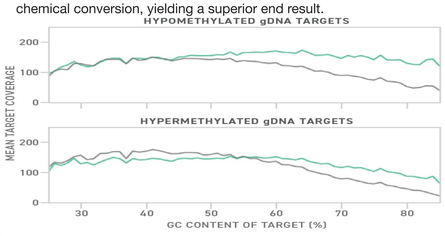

Figure 2. Coverage by Target GC Content for Hypo- & Hypermethylated DNA Libraries Prepared with Bisulfite and Enzymatic Conversion. Improved coverage uniformity is observed across all GC bins when using enzymatic conversion to prepare libraries, regardless of the target methylation state. (EpiScope Unmethylated HCT116 DKO gDNA and EpiScope Methylated HCT116 gDNA, Takara Bio USA).

Compared to chemical bisulfite conversion, enzymatic conversion with the NEBNext EM-seq kit results in high-quality DNA libraries with improved yields and longer insert sizes, each of which is crucial for maximizing sequencing and mapping efficiency. Further, bisulfite treatment is harsh to GC-rich DNA targets since conversion takes place at unmethylated cytosines. This ultimately results in reduced coverage at high-GC target regions that are of great interest in methylation sequencing applications. The approach taken by enzymatic conversion results in more uniform coverage across targets of varying GC content without sacrificing methylation detection sensitivity (**Figure 2**).

Target enrichment can occur before or after library conversion. While post-capture conversion simplifies probe design, this approach requires large amounts of DNA input since PCR amplification does not preserve DNA methylation and cannot take place before conversion. Therefore, pre-capture conversion is often the preferred approach, especially for low-input applications of methylation sequencing, such as cell-free DNA (cfDNA).

Twist Methylome Panel

The Twist Human Methylome Panel targets 3.98M CpG sites across 123Mb of genomic content to target biologically relevant methylation markers. The panel targets several types of differentially methylated regions (DMR) including CpG islands, shores, shelves and open seas (**Figure 3**). Throughout all DMR, the Twist Human Methylome Panel covers 84% of CpG island bases found within the human genome. 54Mb of target bases overlap five types of Ensembl regulatory features most critical for methylation, including CTCF binding sites, enhancers, open chromatin regions, promoters and transcription factor binding sites. In addition to Ensembl, DMR and regulatory features, the design takes into consideration data found in the Genetic Testing Registry (GTR) and genes found within PubMed. In the case of PubMed, 99.9% of regions that have been cited at least 150 times are included in the panel design.

The Twist Human Methylome Panel, in combination with the Twist Methylation Detection System, demonstrates best-in-class capture performance. Picard metrics show that the Twist end-to-end solution gives excellent coverage across the genome at between 100-250X downsampling, with Fold-80 Base Penalty, Percent Target Bases at 30X, Percent Off-Bait, and HS Library Size all reaching gold standard values (**Figure 4A).** Downsampling to 150X coverage, 6.59 million CpG sites were detected at a minimum depth of 10X. Comparing the sequencinging necessary for capture with traditional Whole Genome Bisulfite sequencing (WGBS) makes the methylome panel economic over other assays. (**Figure 4B).** Researchers can spend less resources in sequencing coverage, and still get optimal methylation detection calls.

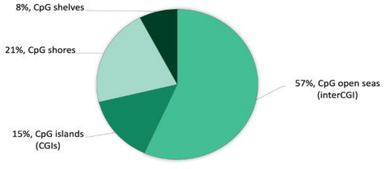

Figure 3. CpG Site Coverage across the Twist Human Methylome Panel. The panel is designed to contain highly curated content of DMRs in order to create an opportune choice for academic research and clinical investigation for discovery.

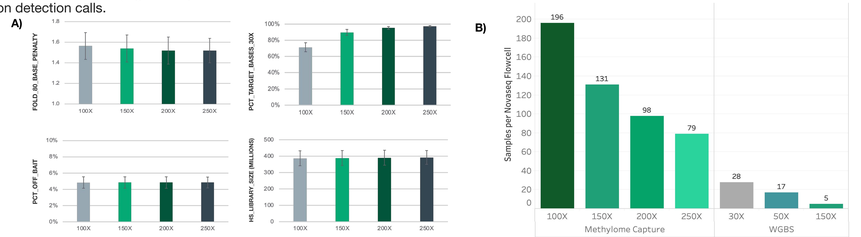

Figure 4. A-B Picard Metrics when using the Twist Human Methylome Panel Libraries were prepared from 200ng of gDNA input (NA12878, Coriell) with the NEBNext EM-seq kit. Target enrichment was performed using 200ng of gDNA library input, a 63°C Fast Wash 1 Buffer temperature, and 16-hour hybridization reactions. Sequencing was performed with the Illumina NovaSeq platform, using NovaSeq V1.5 chemistry kits and 150 bp paired-end reads. Data was downsampled to various aligned coverage (100X, 150X, 200X, 250X) relative to probe territory and analyzed with BWA-meth/MethylDackel and Picard HsMetrics. Figure show common picard metrics (A), and the number of samples that can be multiplexed on a Novaseq 6000 S4 flow cell based on sequencing coverage depth (B).

CpG Methylation Specific Controls

Methylation sequencing experiments are inherently challenging, as experimental variation in conversion can confound the results and produce false positives. By using an inline methylation control these confounding effects can be detected, establishing a best-in-class method for calibrating methylation levels and identifying patterns using methylation target enrichment. Twist's methylation controls (**Figure 5**) are composed of 48 different sequences with no homology in the Refseq database and are similar in length to cfDNA (~167 bp) to be compatible with liquid biopsy workflows. Each sequence is designed to have a GC content similar to the human genome and to avoid stretches of homopolymers. These controls can be spiked directly into gDNA or cfDNA and taken through library preparation and target enrichment with the addition of a panel complementary to the controls. Measured and expected percent methylation levels are compared to determine the success of the conversion process (**Figure 6**).

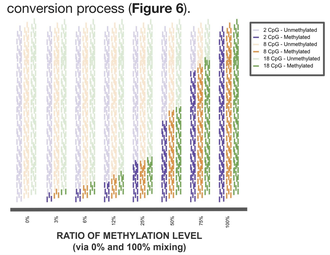

Figure 5. Eight Specific Levels of Methylation with 3 Distinct Numbers of CpG Sites per Sequence. Within each methylation level containing a unique sequence, the material is made up of either low (2, purple), medium (8, orange) or high (18, green) CpG sites. Each pool of controls contains different ratios of methylation levels in order to symbolize 0%, 3%, 6%, 12%, 25%, 50%, 75% or 100%. This is achieved by mixing unique sequences that are 0% methylated material (lighter color) and 100% methylated material (darker color).

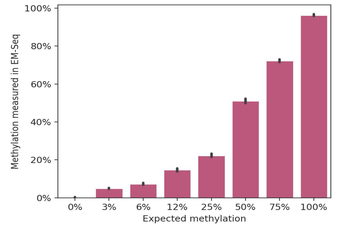

Figure 6. Measured vs. Expected % Methylation of Each Methylation Level Using The Twist Methylation Detection System. 1ul Methylation controls were spiked into 200 ng gDNA then taken through NEBNext EM-seq library preparation and Twist Targeted Methylation sequencing workflow. Sequencing was performed on an Illumina 550 instrument using 2x150 paired end reads. Quantitation of each methylation level was measured using a BWA-meth/methyldackel analysis workflow to determine the amount of methylation for each sequence.

Unique Molecular Identifiers in Methylation Capture

Liquid biopsy has become an increasingly important tool in the early detection of cancer and cancer-associated DNA methylation changes are promising for the development of novel biomarker tests. Using cfDNA as a template for methylation detection can pose additional challenges due to the low abundance of available material and the structure of cfDNA itself. The majority of cfDNA fragments are distributed around 167 bp in length corresponding to the length of DNA wrapped around a nucleosome. This fragmentation structure can lead to a decrease in stop-start diversity of molecules and increases the probability of calling unique molecules as duplicates. Using the Twist Methylation Detection system helps investigators overcome these challenges as using enzymatic conversion preserves library fragment length and performing capture post conversion allows for low mass input library preparation. Additionally, methylated unique molecular identifiers (UMIs) can be added to the system to improve deduplication in low diversity insert size samples such as cfDNA. The Twist Methylated UMI adapters are designed to be compatible with the EM-seq library preparation protocol, Twist target enrichment, and Illumina sequencing.

Here, cfDNA enzymatically converted libraries were captured using the Twist Alliance Pan-Cancer Methylation panel-1.5MB which covers 126K CpG sites and contains targets relevant to 31 different cancers. This workflow produces optimal Picard metrics when tested with standard methylated adapters and the Twist Methylated UMI adapters (**Figure 7**). When using an analysis pipeline that integrates UMI information into alignments to deduplicate data, UMI informed deduplication provides higher mean target coverage over standard dupsmarking alone (**Figure 8**).

Figure 7. Picard Metrics from cfDNA Captured with a Cancer Specific Methylation Panel Libraries were prepared from 10 ng or 30 ng cfDNA input (BioIVT) with the NEBNext EM-seq kit using standard adapters or Twist methylated UMI adapters. Target enrichment was performed using the Twist Alliance Pan Cancer Methylation Panel. Sequencing was performed on a Illumina Nextseq 2000 platform, using 2x100 paired-end reads. Data was downsampled to 1000X aligned coverage relative to probe territory and analyzed with BWA-meth/MethylDackel and Picard HsMetrics. Consistent performance between UMI and non-UMI containing libraries can be seen across coverage uniformity, percent on-target rate, HS library size, and percent targets at 30X coverage.

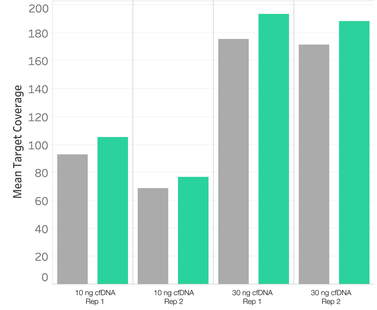

Figure 8. Mean Target Coverage using UMI for Deduplication Comparison of mean depth over panel targets using either position-based deduplication (as offered in GATK MarkDuplicates) or UMI-informed dupsmarking, which splits reads into families based on both duplex UMIs and mapping positions and then marks duplicates on the basis of this information. After dupsmarking with both strategies we measured the increase in mean-target coverage that could be obtained by avoiding collapse of coincidental library fragments.

Analysis workflow

The analysis workflow comprised a series of steps to align reads to a converted genome while preserving UMI information. First, FASTQ files were pre-processed to extract UMI information, followed by alignment to a converted version of hg38 using bwa-meth.

After alignment, we marked duplicates either based on position alone, or based on both position and UMI sequence using custom Python scripts and the pysam library. We also collect sequencing and enrichment metrics using Picard to evaluate the success of the target enrichment portion. Plotting is done with custom Python scripts.

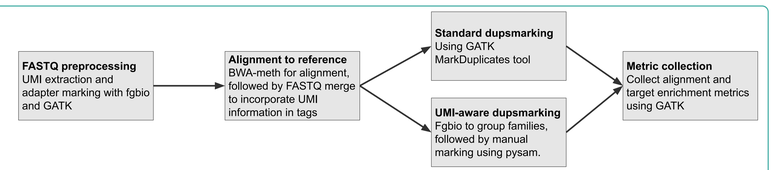

Figure 9. Summary of The Bioinformatic Processing Steps to Mark UMI duplicates
Raw reads were processed using fgbio and GATK to mark adapter sequences and to extract UMI information. Reads were then aligned to the human reference genome (hg38) using BWA-meth, after which we dupsmarked either using positions (GATK MarkDuplicates) or using UMI information (using fgbio and custom Python scripts). After dupsmarking, we compared the results using standard metrics calculated through GATK.

Conclusions

The Twist Methylation Detection System provides an end-to-end sequencing solution for a diverse range of applications. The expansive content of the Twist Human Methylome panel makes it an ideal choice for investigators to explore the methylation fraction across the full genome. The Twist Methylation Controls can be used to aid the accurate quantification of methylation states. Targeted methylation panels like the Twist Alliance Pan-Cancer Methylation panel in combination with UMI adapters provide additional benefits in the design of cfDNA methylation detection assays.

Conflict of interest statement

All authors are employees and shareholders of Twist Bioscience

Poster#: 6009