Methylation Controls to detect for methylation level quantification in the Twist Targeted Methylation Sequencing workflow

Kristin Butcher, Michael Bocek, Patrick Cherry, Jean Challacombe, Esteban Toro*

Abstract

DNA methylation at CpG nucleotide sites in eukaryotes is a key epigenetic mark that helps control gene expression. Specific changes in CpG methylation occur in many human cancers, making them a promising biomarker for early cancer detection, especially in the context of liquid biopsy testing. However, assays for methylation detection are often only semi-quantitative for the methylated fraction, making it difficult to firmly establish the limit of detection for these assays and to define methylation levels in a given sequence.

To address these challenges, Twist Bioscience has designed and developed a CpG methylation specific control in order to help calibrate assays that determine site-specific methylation rates. By using our core DNA synthesis technology, we have constructed a pool of 48 unique contrived sequences that contain a total of 8 different levels of methylation at a CpG site, ranging from 100% to 0%. Each unique sequence also contains either a low, medium, or high amount of total CpG sites for direct comparisons of sequences with various CpG sites. These controls also mimic the DNA fragment lengths commonly found in cell-free DNA (cfDNA) and are therefore compatible with experimental workflows that target liquid biopsy applications. These pools can be taken through the Twist Targeted Methylation Sequencing workflow, using panels designed to target these controls as spike-in panels during hybrid capture.

This study demonstrates an improved method for calibrating detection assays across a range of methylation levels containing variable amounts of CpG sites. This aims to make a direct impact in the future of cancer care and analyzing patterns in data related to methylation levels.

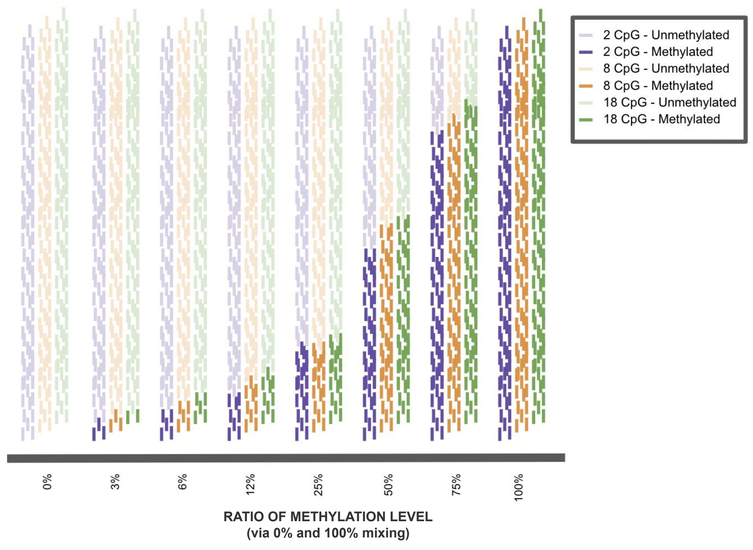
Methylation control design

Twist's methylation controls were carefully designed to enable accurate methylation calling in cfDNA samples. Sequences were chosen to be cfDNA-like in length (~170bp) and distinct from each other and any sequence in the non-redundant (nr) database. A total of 48 different sequences are present, representing:

- 8 methylation levels (0%, 3%, 6%, 12%, 25%, 50%, 75%, and 100%)
- 3 levels of CpG sites per molecule (2, 8, or 18 CpG sites per molecule)
- 2 replicate sequences for each combination of methylation level and CpG site number

All sequences are also designed to have a GC content similar to the human genome and to avoid stretches of homopolymers. As shown in **Figure 1**, specific methylation levels are achieved by mixing fully methylated and unmethylated populations of the same sequences.

Figure 1. 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).



Pool calibration and quality control

Methylation control preparation requires several distinct processes (**Figure 2**). Pools of controls are first made based on methylation level, mixing the low/medium/high CpG site sequences that correspond to that methylation level. This ensures that ratios of the different CpG site numbers are consistent in the final pool. After this first pooling, sequences are methylated and mixed with untreated material to get the different methylation levels shown in **Figure 1**.

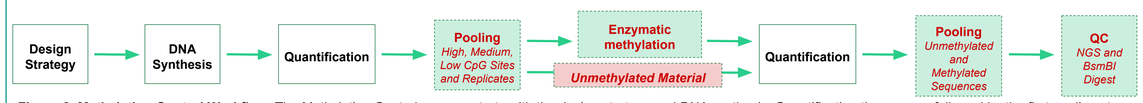


Figure 2. Methylation Control Workflow. The Methylation Control process starts with the design strategy and DNA synthesis. Quantification then occurs followed by the first pooling step. The pools are then enzymatically methylated. Finally, mixtures of different methylation levels are made using the methylated and unmethylated pools.

Two quality control (QC) measures are performed to guarantee performance of the methylation pools. During the design process, a BsmBI digestion site is added to each unique sequence. A small volume of the sequences designated as 0% methylated and 100% methylated are taken prior to the second pooling and a BsmBI digestion is done. BsmBI does not cut when the CpG in its recognition site is methylated (**Figure 3**). Therefore, this digestion is done to confirm that the methylation step generated fully methylated material prior to pooling the different methylation levels.

Once the pools are finished, next-generation sequencing (enzymatic methyl--seq) of the methylation controls is completed to ensure that pooling and methylation levels are correct.

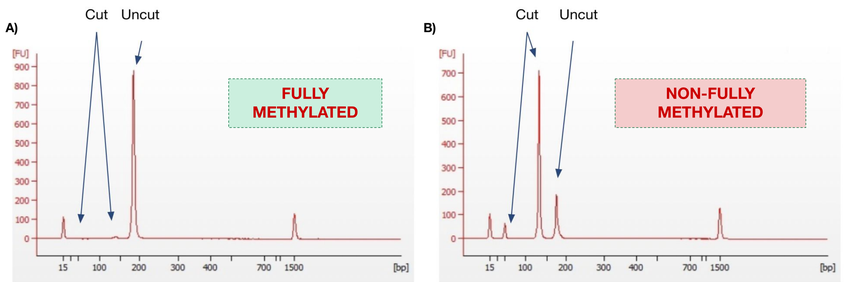


Figure 3. BsmBI Digestion on Fully Methylated Material. BsmBI is used to confirm full methylation post-enzymatic treatment. When CpG sites are methylated in the BsmBI recognition site, the enzyme doesn't cut. BioAnalyzer data post-BsmBI digestion shows the final QC trends when material is fully methylated (A) and non-fully methylated (B). Fully methylated material shows a single peak at approximately 170 base pairs. Non-fully methylated material shows several peaks, one peak demonstrating uncut material and two peaks demonstrating digested material.

Twist Methylation Detection System

Twist methylation controls are designed to be used in conjunction with the Twist Methylation Detection System, which includes target enrichment. The methylation controls are added to sample gDNA prior to library preparation. Target enrichment is then performed using a panel that combines probes designed against customer-defined targets and methylation control targets (**Figure 4**).

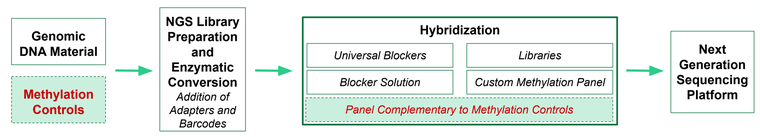


Figure 4. Twist Methylation Detection System Overview while Using the Methylation Controls. Methylation controls are spiked-in to sample gDNA and allowed to go through the library prep conversion process. Libraries then go through hybrid capture using customer-defined panels and a specific panel complementary to the methylation controls. Methylation controls are pulled down and sequenced alongside gDNA molecules and quantified using standard methyl-seq methods.

Conventional library preparation protocols convert unmethylated cytosines to uracils using bisulfite, which causes unwanted DNA breaks that complicate downstream sample preparation and, ultimately, 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.

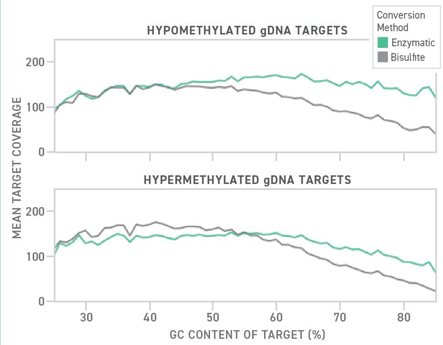


Figure 5. Coverage by Target GC Content for Hypo- & Hypermethylated DNA Libraries Prepared with Bisulfite and Enzymatic Conversion. Improved coverage uniformity is observed across all GC bins when using enzymatic conversion to prepare libraries, regardless of the target methylation state. (EpiScope Unmethylated HCT116 DKO gDNA Cat. # 3521 and EpiScope Methylated HCT116 gDNA Cat. # 3522, 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 gentle approach taken by enzymatic conversion results in more uniform coverage across targets of varying GC content without sacrificing methylation detection sensitivity (**Figure 5**).

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.

Methylation calibration and level quantification

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 and even potentially corrected. The methylation controls, in combination with the Twist Methylation Detection System, establish a best-in-class method for calibrating methylation levels and identifying patterns using methylation-based hybridization assay technology.

To test these controls, 1µL of control was added to 200ng of gDNA input (NA12878, Coriell) and the NEBNext EM-seq kit was used for library preparation. Target enrichment was performed using 200ng of library, a 65°C Fast Wash 1 Buffer temperature, and 2-hour hybridization reactions. A methylation control complementary panel was used as a spike-in during the hybridization process. Sequencing was performed with the Illumina NextSeq platform and 151 bp paired-end reads. Data were downsampled to 150X aligned coverage relative to probe territory and analyzed with BWA-meth/MethylDackel and Picard HsMetrics. Measured and expected percent methylation levels were compared (**Figure 6**) to determine if the expected and observed methylation levels match.

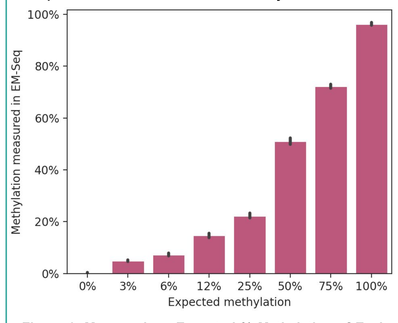


Figure 6. Measured vs. Expected % Methylation of Each Methylation Level Using The Twist Methylation Detection System. Quantitation of each methylation level was measured by taking the controls through library preparation and target enrichment. The controls were then analyzed using an analysis workflow to determine the amount of methylation for each sequence.

Twist standards are produced as either fully methylated or fully unmethylated molecules and then pooled to the desired fractions. Therefore, individual sequences should show a bimodal distribution of sites by methylation state. This is what we observe (**Figure 7**), with the vast majority of sequences being either fully methylated or fully unmethylated.

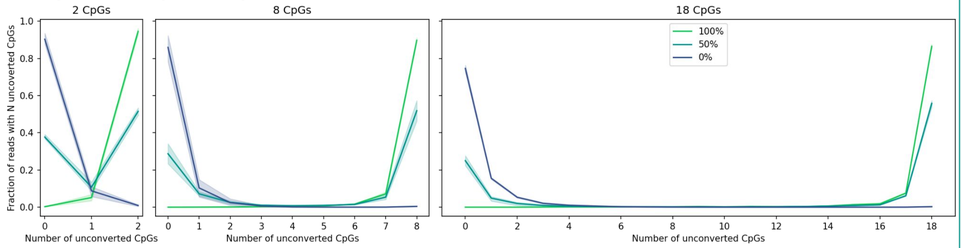


Figure 7. Quantification of Individual Methylated CpG Sites. All three amounts of CpG sites per methylation level were quantified to determine how many CpG sites are found in each sequence. Each figure reflects sequences that contain 2, 8 or 18 CpG sites. The data shows that the majority of reads are either fully methylated (100%) or completely unmethylated (0%), meaning that a 50% methylated control reflects half of the sites being fully methylated in half of the fragments, and half of sites being completely unmethylated. There are a small number of fragments either missing one site or with one additional site, which likely reflects mild inefficiencies in the process and/or sequencing or conversion errors.

Analysis workflow

We tested 3 methylation aligners (Bismark, Bwa-meth, and BsMapz) and 2 methylation callers (BsMapz and MethylDackel - Figure 8) Components were evaluated for performance and efficiency on a contrived dataset with known methylation states. On the basis of these metrics, we selected bwa-meth and MethylDackel for alignment and calling.

The pipeline used for analysis begins with read processing to downsample and remove adapters, followed by alignment (using bwa-meth) and methylation state calling (with MethylDackel - Figure 9). 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.

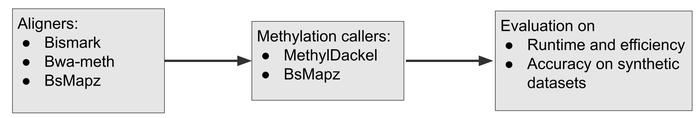


Figure 8.. Evaluation of Aligner and Methylation Caller Combinations.

Figure 9. Summary of The Bioinformatic Processing Steps Utilizing The BWA-meth/MethylDackel Analysis Workflow. Reads are downsampled, adapter trimmed, and aligned to the converted genome with BWA meth. After alignment, CpG methylation state is called with methyldackel, and other sequencing metrics (i.e., enrichment, GC bias) are collected with Picard.

Conclusions

Methylation patterns are an important component of the cancer detection toolbox. The Twist Methylation Spike-in Controls can be used with the Twist Methylation Detection System to aid the accurate quantification of methylation states, serving as useful process controls for methylation applications.

Conflict of interest statement

All authors are employees and shareholders of Twist Bioscience