A Patient-Centric Methylation Pipeline

Methodology for detecting DNA methylation changes in microarray data.

Introduction

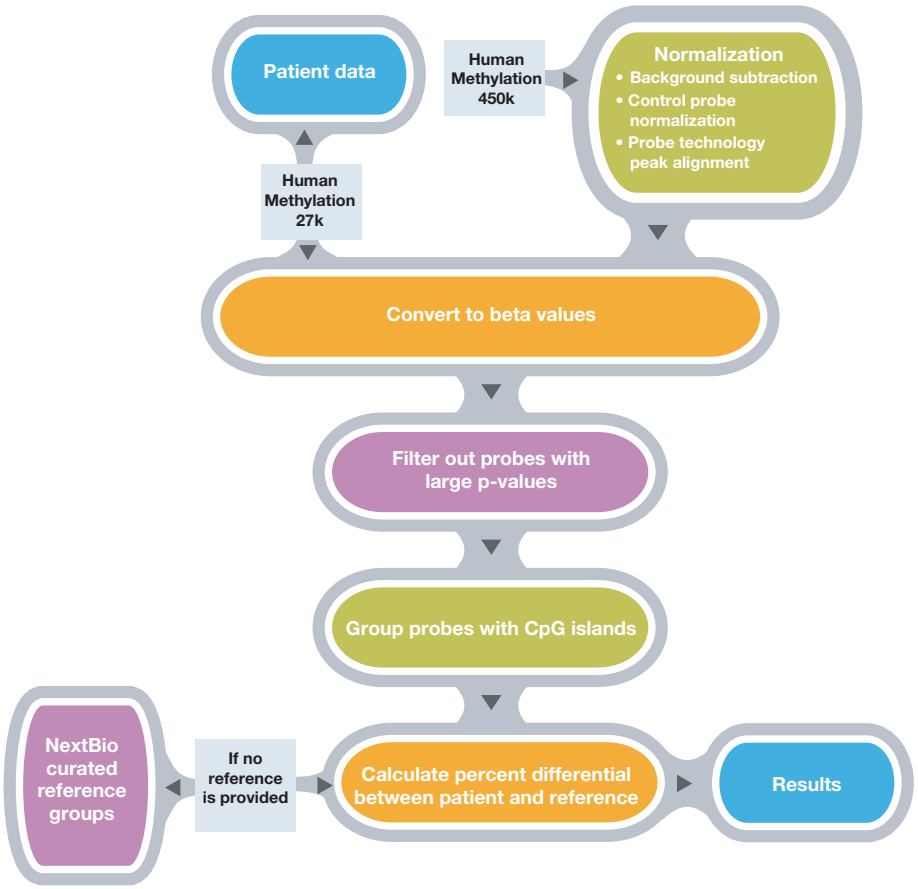
Methylation of DNA cytosine residues is a common epigenetic mark and is often found in the context of CpG dinucleotides. It is a crucial epigenetic modifier, implicated in regulating many cellular processes in normal and diseased cells[1,2]. As a result, there is a growing interest in profiling DNA methylation across the genome to improve understanding of this epigenetic factor's impact.

This application note outlines the methodology implemented in the NextBio platform (Figure 1) for detecting changes in DNA methylation for data produced on the Illumina Infinium® methylation assays. The method analyzes data in a patient-centric context, i.e., data are obtained for an individual patient, rather than as a batch. Therefore, the underlying analysis must be expanded to handle this complexity. The analysis pipeline is designed to compare a patient disease sample to a normal reference sample. The reference is derived from the average of a group of individual normal samples collected from patients (when available), or generated from highly correlated NextBio curated data sets from the same tissue as the disease sample.

Current Methylation Standards

Growing interest in DNA methylation has led to the rapid development of new technologies. This is visible in the cytosine coverage of the Infinium platforms, which has increased from 1,500 positions per platform to 27,000. With the introduction of the HumanMethylation450 BeadChip platform, this coverage now extends to over 480,000 positions.

Figure 1: Methylation Pipeline Workflow



On older arrays, each gene was typically tiled with 1–3 probes, with specific focus on the gene promoter regions. This design assumed that adjacent CpG sites had similar methylation status, thereby permitting a one-probe-to-gene strategy. The identification of differentially methylated genes was limited to a single-probe analysis which was, consequently, used to imply changes in gene transcription.

However, recent studies exploring CpG methylation patterns have shown associations between gene body methylation and transcriptional expression[3,4]. Further, there is evidence supporting the need to study relative change in methylation of a region from its own positional baseline signal. Thus, it is necessary to classify probe location and calculate differential methylation with respect to its own matched baseline. The Illumina HumanMethylation450 arrays have between 6 and 200 probes per gene and now investigate cytosine methylation across the entire gene region. This coverage allows researchers to interrogate the positional effects of methylation.

NextBio Methylation Analysis

The methylation analysis workflow processes data from the Illumina Infinium HumanMethylation27 and HumanMethylation450 platforms. After accounting for platform-specific normalization steps, the workflow pools neighboring CpG probes, allowing a more robust calculation of methylation signal and differential methylation. This measurement is especially important within a patient-centric context, as each probe signal is not an aggregate across replicates. Because the impact of CpG methylation is dependent on probe locations, probes are pooled within the same locally concentrated CpG island

Data and Ingestion

HumanMethylation27

Methylation data from the Infinium HumanMethylation27 arrays are reported as either beta (β) values or as individual methylated and unmethylated singals. The **limma** and **methylumi** packages in R are used to ingest data.

HumanMethylation450

Methylation data from the Infinium HumanMethylation450 array are reported as raw binary IDAT files generated by the iScan® system. These data are processed using the **methylumi** package in R. Individual methylated and unmethylated signals are extracted from this package, along with p-values calculated in comparison to negative control probes. Background correction is performed via a normal-exponential deconvolution approach using the **methylumi** package in R[5]. This method was also implemented by the The Cancer Genome Atlas (TCGA) consortium to generate their processed data.

Normalization and Correction

The HumanMethylation27 and HumanMethylation450 platforms differ in their detection of methylation signals from all CpG probes. Further, because of genomic variation in CpG methylation, commonly used normalization methods like quantile or Loess normalization can remove real biological signal. In order to account for these two factors, Illumina has implemented a two-pronged normalization pipeline for DNA methylation that is platform-dependent.

HumanMethylation27

The HumanMethylation27 platform uses fluorescent dyes to read the methylation signal; however, the dyes do not correspond to DNA methylation status. As a result, typical dye bias corrections are not relevant to this case. The signal is converted to a β value using Equation 1. This formula normalizes the methylation signal to the total signal.

$$\beta\ value = \frac{MethylatedSignal}{MethylatedSignal + UnmethylatedSignal} \quad (1)$$

HumanMethylation450

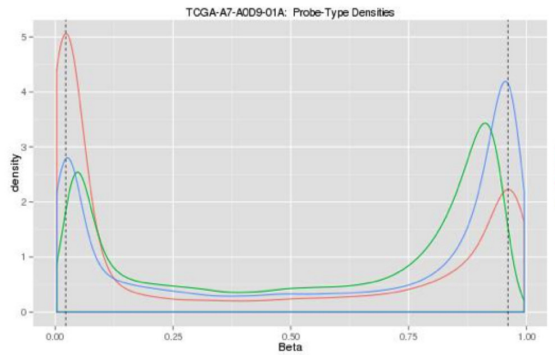
Probe signals from this platform are first background-adjusted and then normalized based on the platform's negative control probes. Infinium I technology for signal measurement is more sensitive to measurements near the tails of the distribution, while Infinium II technology tends to consolidate these extreme measurements to the center of the measurement distribution[6]. To account for these two signal measurement technologies, the methylation pipeline in the NextBio platform computes a β value as shown in Equation 2:

$$\beta\ value = \frac{MethylatedSignal}{MethylatedSignal + UnmethylatedSignal + 100} \quad (2)$$

A β value of 0 implies no methylation was observed while a value of 1 implies that the site is fully methylated[7–9].

To account for the differences in the β distribution of type I and type II probes, the type II probe's β distribution is adjusted so that the two peaks are closet to 0 and 1 (Figure 2). This peak correction of β value results in better signal capture[10].

Figure 2: Peak Correction

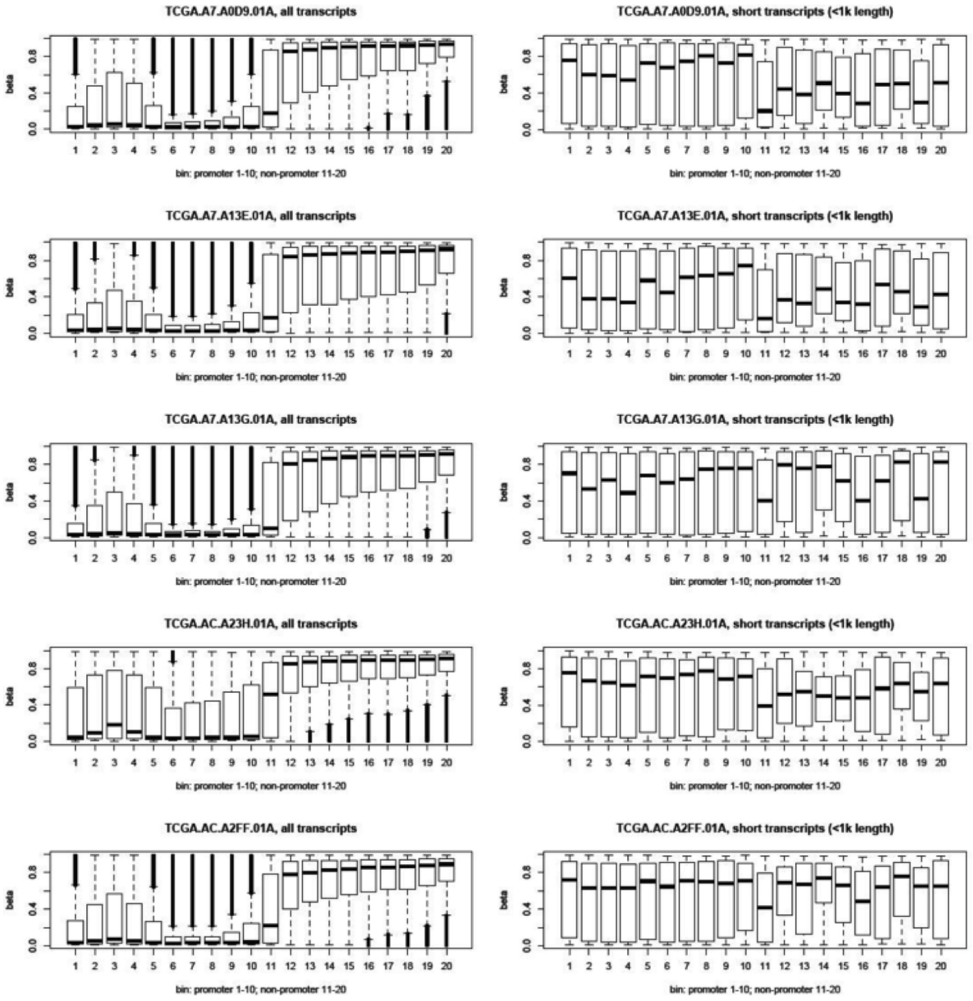


Peak correction for type II probes on the HumanMethylation450 array. Type I, type II, and adjusted type II probes are plotted in red, green, and blue, respectively.

Detection P-Value

Each probe is assigned a detection p-value by Illumina GenomeStudio® software. This value reflects the probability that the probe signal was produced from noise and is calculated using the negative control probes from the platform. Probes with a detection

Figure 3: Promoter Region Definition



Beta-value distributions for all transcripts at promoter and non-promoter regions of five TCGA samples. The β-value distributions for all transcripts at promoter and non-promoter regions of five TCGA samples were plotted. The left column shows analysis of all transcripts. The right column shows analysis of short transcripts. Bins 1-5 represents a total of 1,000 bp region upstream of the transcription start site (TSS), with each bin spanning 200 bp. Bins 6-10 are of the same size, spanning the TSS to the end of the promoter. Bins 11-20 represent the region between the end of the TSS and the beginning of the 3′ UTR. Each bin is of the same size.

p-value higher than a platform-specific threshold are removed from downstream analysis. On average, probes have extremely low p-values:

HumanMethylation27: p-value threshold > 0.01

HumanMethylation450: p-value threshold > 0.05

Pooling Probes

Because the probe placement strategy on the HumanMethylation27 and HumanMethylation450 platforms are different, the pooling strategy is platform-dependent. Usually, one or two probes on the HumanMethylation27 platform map back to the promoter region of the transcript. With the HumanMethylation450 platform, multiple probes map back to different regions spanning the length of the transcript. The pooling strategy accounts for these differences. For the HumanMethylation450 arrays, the analysis is restricted to the promoter region for ease of interpretation.

HumanMethylation27

The impact of CpG methylation is location-dependent. Therefore, an average of all methylation probes across the gene would over-smooth the data. Alternatively, reporting a change in methylation from data produced by a single probe would bias data towards outliers. Methylation signals for probes within most CpG Islands tend to be highly correlated. Moreover, most islands tend to cover analogous gene regions, i.e., one island covers the promoter and 5′ -untranslated region (UTR), while another is concentrated in the gene body. Thus, the methylation analysis pipeline calculates the average signal from probes within islands and uses this statistic for downstream analysis.

HumanMethylation450

In order to report a meaningful methylation signal for each transcript or gene represented on the HumanMethylation450 array, summary statistics are generated for methylation signals in the promoter region

of the transcript. This is due to a common observation that the methylation signal in the promoter region is negatively correlated to mRNA expression of transcripts[11]. In total, five definitions for **promoter regions** were tested for robustness in five TCGA samples:

1. The Bioconductor probe annotation.
2. Modified annotation based on Bioconductor probe-to-transcript mapping.
3. 1.5 kbp upstream and 0.5 kbp downstream of the transcription start site (TSS), based on Refseq transcripts annotation.
4. 1 kbp upstream and 1 kbp downstream of the TSS, based on Refseq transcripts annotation.
5. A modified annotation based on Refseq transcripts annotation (1 kbp upstream of the TSS and the minimum of 1 kbp, 5′ -UTR, or the first exon after the TSS.

The β-value distributions for all transcripts at promoter and non-promoter regions of five TCGA samples, based on the fifth definition for promoters, are plotted in Figure 3. The fifth definition yielded strong negative correlation with expression data and was adopted as the definition for promoter regions. Since methylation signals at nearby CpG sites are consistent[12,13], a clear methylation signal is expected to show a peak of β values across multiple probes. Smaller peaks of differentially methylated regions are filtered out if they have less than three probes that show a β value difference of least 0.2 between the sample and the reference. For promoter regions with no observable peaks, the methylation signal is calculated as the median β value across all probes in the region.

Selecting the Reference

Since differential methylation for a patient is calculated against a reference, it is essential to identify and use a reference that generates the most meaningful results. Ideally, the reference sample used belongs to the patient in question and is provided by the same laboratory submitting the patient sample. When a set of normal samples is provided, the average of probe signals for a region across all samples is used as a reference normal. In cases where a reference sample is not provided, the closest matching reference from curated NextBio studies is used. In such cases, the methylation pipeline requires at least 95% correlation between the percent methylation differential across all islands. Similarly, when no matched reference is present in the curated datasets, the methylation pipeline uses comparable tissue datasets to create a generic reference and requires a correlation of at least 87%.

CpG Island Differential

Once a reference dataset is identified, the percent differential for all CpG islands between the patient and reference group is calculated. The methylation signal between island groups is tested for significance using a two-sample t test. This test assumes that the patient data and reference sample have equal variance and the underlying β values follow a β distribution. This approach ensures that the differentials are bounded.

Conclusions

A number of human diseases are associated with aberrant DNA methylation. In particular, hypermethylation of CpG islands located within promoter regions of tumor-suppressor genes has been established as a common mechanism for gene regulation in cancer. Therefore, high-throughput profiling of the DNA methylation status of CpG islands is crucial for advancing the understanding of this epigenetic marker.

Illumina has developed a robust pipeline to process DNA methylation studies in a patient-centric context. The workflow leverages the unprecedented amount of curated NextBio data to create customizable reference groups that facilitate high-quality analysis.

References

1. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33(3s): 245–254.
2. Zhang FF, Cardarelli R, Carroll J, Zhang S, Fulda KG, et al. (2011) Physical activity and global genomic DNA methylation in a cancer-free population. Epigenetics 6(3): 293–299.
3. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK, et al. (2012) On the presence and role of human gene-body DNA methylation. Oncotarget 3(4): 462–474.
4. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13: 484–492.
5. Ritchie ME, Silver JD, Oshlack A, Holmes M, Diyagama D, et al. (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics 23: 2700–2707.
6. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, et al. (2011) Evaluation of the Infinium Methylation 450K technology. Epigenomics 3(6):771–784.
7. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 587.
8. Touleimat N, Tost J (2012) Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 4(3): 325–341.
9. Pidsley R, Wong CC, Volta M, Lunnon K, Mill J, et al. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 14: 293.
10. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, et al. (2011) Evaluation of the Infinium Methylation 450K technology. Epigenomics 3(6): 771–784.
11. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. (2011) Epigenetics 6(6): 692–702.
12. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, et al. (2006) DNA methylation profiling of human chromosomes 6, 20, and 22. Nat Genet 38(12): 1378–1385.
13. Li Y, Zhu J, Tian G, Li N, Li Q, et al. (2010) The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol 8(11): e1000533.
14. Ehrlich M (2002) DNA methylation in cancer: too much, but also too little. Oncogene 21(35): 5400–5413.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 970-2014-019 Current as of 05 December 2014

illumina®