3493

Read-level methylation pattern extraction for high-multiplex large-scale targeted NGS assay

Mingda Jin[1], Masatomo Kaneko[2], Steven Cen[3], Hongtao Li[2], Wei Guo[1], Xinyi Zhou[2], Atsuko Fujihara[2], Tsuyoshi Iwata[2], Lorenzo Storino Ramacciotti[2], Divyangi Paralkar[2], Giovanni E Cacciamani[2], Manju Aron[2], Osamu Ukimura[2], Inderbir S. Gill[2], Gangning Liang[2], Andre L. Abreu[2], Jeffrey Bhasin[1], Xiaojing Yang[1*], Xi-Yu Jia[1]

[1]Zymo Research Corp., Irvine, CA, [2]Department of Urology, University of Southern California, Los Angeles, CA, [3]Department of Radiology/Neurology, University of Southern California, Los Angeles, CA

Abstract

The abnormal changes in DNA methylation are linked to the early stages of carcinogenesis. Identifying these epigenetic changes in circulating tumor DNA (ctDNA) can reveal potential biomarkers for the early diagnosis of various cancers. However, analyzing such data poses bioinformatics challenges due to the lack of sensitivity in detecting the low abundance ctDNA signals in biopsy samples, which are often overwhelmed by the complexity of libraries containing hundreds of targeted regions. Read-level methylation analysis holds the promise of more in-depth DNA methylation detection due to the wide coverage and high sensitivity of rare signals. However, this approach is hindered by the absence of a standardized workflow capable of generating interpretable reports suitable for both bench scientists and professional bioinformaticians. Here, we present a bioinformatics workflow that examines next-generation sequencing (NGS) data and characterizes the read-level methylation patterns of amplicons. Compared to other currently available tools, our method is designed to work with high-multiplex, large-scale targeted assays. It effectively eliminates the undesired noise derived from sequencing byproducts such as false CpG calls, dimers, and off-target alignments. Additionally, to accommodate the substantial volume of data generated by state-of-the-art NGS platforms, the workflow enables parallel processing of samples compatible with both cloud-based and on-premises computing resources. This workflow provides a comprehensive per-sample visualization of DNA methylation patterns and reports read-level methylation results in a "pattern-as-a-feature" table. In this table, the occurrence of an amplicon epiallelic haplotype (pattern) for every sample is represented as a "feature column" and is aggregated with all patterns discovered in the experiment. These read-level patterns, along with other information, can be used to develop machine learning algorithms to reiteratively harvest true predictive features and penalize confounding signals in predicting cancer diagnosis.

Keywords: Read-level, Methylation pattern, Targeted-seq, Bioinformatics workflow

Background

Distinguish Predictive Signals

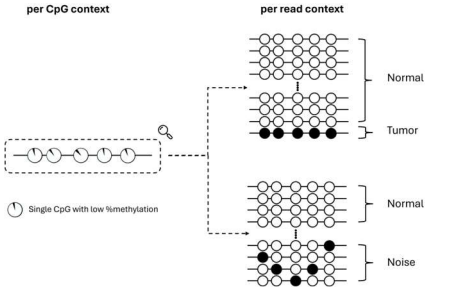

Figure 1. A zoom-in view of amplicon CpG methylation using read-level approach. Methylation patterns provide better distinction between noise and signals that carry true predictive power e.g., highly methylated reads, especially in low methylated regions.

Quantify Heterogeneity

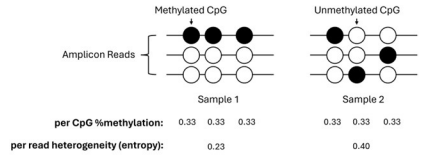

Figure 2. A new perspective to inspect amplicon methylation using read-level analysis compared to traditional CpG-level approach. Sample 1 and Sample 2 have the same CpG methylation ratios. But from a per-read view, they exhibit distinct epiallelic patterns indicating different levels of cell heterogeneity, which is measured here using entropy.

Methodology Workflow

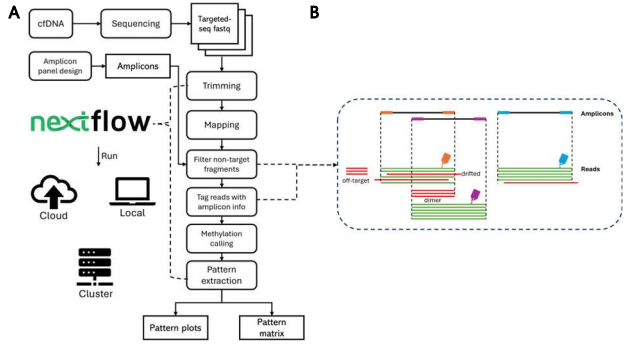

Figure 3. The bioinformatics workflow for read-level methylation pattern extraction from targeted sequencing assay. (A) An amplicon panel targeting various regions of interest is designed, and then used to amplify DNA from samples for targeted sequencing. The resulting FASTQ files undergo QC, trimming and mapping to generate raw BAM files. These mapped reads are further refined to reduce noise arising from library complexity. During methylation calling, a reference genome is supplied to eliminate false CpG calls. Finally, read-level CpG methylation is extracted for each amplicon, and generates outputs in both plots and data matrices. The bioinformatics workflow is orchestrated by Nextflow enabling the parallel processing of large-scale assays compatible across multiple computing platforms. **(B)** Reads derived from non-specific binding (red), such as primer dimers, off-target, and drifted reads are filtered out. Only the valid on-target reads (green) are retained and annotated with their corresponding amplicon information. This step not only greatly enhances the accuracy of subsequent pattern extraction but also improves the efficiency of workflow execution.

Results

Workflow Output

A

	Sample_1	Sample_2	Sample_3	...	Sample_391	Sample_392	Sample_393
Amp_1:CC	17	1397	1283	...	0	3	1
Amp_1:CT	2	46	8	...	0	1	0
Amp_1:TC	1	90	44	...	0	0	0
Amp_1:TT	0	5	0	...	0	0	0
Amp_2:CC	339	225	80	...	610	966	791
...	...	...	...	...	...	...	...
Amp_580:CTT	0	1	2	...	0	0	0
Amp_580:TCC	2	14	6	...	0	1	0
Amp_580:TCT	4	4	1	...	0	0	0
Amp_580:TTC	7	20	23	...	0	3	0
Amp_580:TTT	17	38	15	...	0	0	0

B

	Sample_1	Sample_2	Sample_3	...	Sample_391	Sample_392	Sample_393
Amp_1	0.265	0.170621	0.075263	...	NaN	0.375	0
Amp_2	0.612638	0.617131	0.65528	...	0.439652	0.438496	0.544735
Amp_3	0.338561	0.239873	0.448867	...	0.674747	0.726469	0.427197
Amp_4	0.681503	0.616542	0.705555	...	0.737095	0.866227	0.668113
Amp_5	0.726488	0.493894	0.369692	...	0.806122	0.613595	0.838648
...	...	...	...	...	...	...	...
Amp_576	0.24	0.111794	0.088982	...	0.5	0.5	0.277778
Amp_577	0.206598	0.256403	0.219187	...	0.493185	0.336424	0.411287
Amp_578	0.465374	0.418157	0.435956	...	NaN	0	0
Amp_579	0.355275	0.384761	0.35685	...	0.490978	0.492914	0.463864
Amp_580	0.713843	0.757929	0.740661	...	NaN	0.59375	0

Figure 4. The workflow outputs include both pattern visualizations and data tables for downstream analysis. (A) The table presents the read counts of epiallelic patterns for each amplicon across all samples. Each pattern is denoted by an amplicon name followed by a combination of the letters 'C' and 'T', where 'C' represents a methylated CpG and 'T' an unmethylated CpG. Patterns with unknown methylation state, such as 'N', are excluded from the table. The number of potential patterns for an amplicon expands exponentially with the number of CpG. Consequently, a large-scale, high-multiplex assay may produce a sparsely populated matrix with a massive number of rows. **(B)** The table delivers the epipolymorphism score of each amplicon for every sample. Pattern counts can be characterized using established metrics. The available options are epipolymorphism score and entropy. **(C)** Visualization of amplicon methylation patterns. **See Figure 5B.** The lollipop plot depicts methylation patterns of a single amplicon with circles along horizontal lines. The colors represent different methylation states as per the legend. Adjacent to it is a histogram showing the frequency of the corresponding patterns. A bar plot below summarizes methylation ratios at each CpG within the amplicon.

Results (cont'd)

Case Study

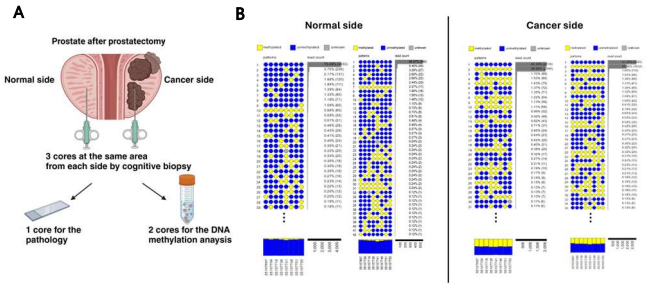

Figure 5. Read-level methylation to predict aggressive features in prostate cancer. (A) Prostate cancer patients were recruited to collect three-core needle biopsies at the same location from each lobe of the ex-vivo radical prostatectomy specimen. Two cores were analyzed for read-level DNA methylation while one was sent to pathologist for diagnosis. After screening more than 200 samples and over 580 regions, a specific gene as an example in this study was identified to have a strong positive correlation between the gene body methylation and aggressive features in PCa patients. **(B)** The amplicon methylation patterns of the informative region in four needle biopsies from the same patient. Some low-frequency patterns were truncated to fit the size of the poster. Cancer and non-cancer tissues are determined histopathologically. The samples from the cancer side exhibit a significantly higher content of fully and almost fully methylated reads with 8 and 7 methylated CpGs. Thus, the increased presence of such differentially methylated patterns may serve as an epigenetic marker for prostate cancer diagnosis.

Conclusions

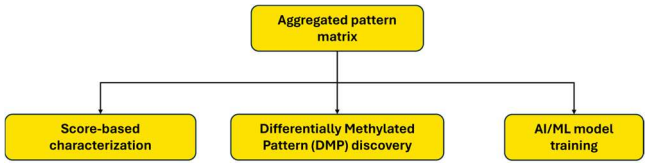

Figure 6. Potential analysis routes empowered by pattern matrix.

The study shows an effective and robust bioinformatics workflow to extract read-level methylation patterns for targeted sequencing assays. Here, we present three potential approaches researchers can utilize the result (As shown in Figure 6).

Compared with other similar bioinformatics workflows, our method has several advantages:

- ✓ It can work efficiently with high-multiplex amplicon panel and large cohort of samples to enable high-throughput screening.
- ✓ It removes the confounding data and possesses the sensitivity to accurately detect low-abundance signals in samples.
- ✓ It provides a quantitative result that can seamlessly integrate into further analysis to meet users' requirements.

All these features make it ideal for data-intensive applications such as AI-powered cancer screening.

References

- Wong, Nicholas C., et al. "MethPat: a tool for the analysis and visualisation of complex methylation patterns obtained by massively parallel sequencing." BMC bioinformatics 17 (2016): 1-14.