A Statistical Method to Estimate DNA Copy Number from Illumina High-density Methylation Arrays

Simon M Lin, Northwestern University
Gang Feng, Northwestern University
Xin Lu, Abbott Laboratories
Yue Yu, Northwestern University
Andrea Baccarelli, Harvard Unviersity
Hongmei Jiang, Northwestern University
Warren A. Kibbe, Northwestern University
Pan Du, Genentech Inc.
Lifang Hou, Northwestern University

Abstract

For the first time, we report that Illumina high-density methylation arrays can also be used to estimate DNA copy numbers. To demonstrate this statistical method, we used the Illumina HM450K methylation array to characterize the copy number aberrations of the HT-29 colon cancer cell line. The result was validated using the golden standard of Affymetrix SNP array. This statistical method will reduce the processing time and lower the cost of large-scale DNA copy number and methylation profiling studies, which is relevant to the CAMDA 2010 data set on profiling glioblastoma.

Introduction

High-density SNP genotyping arrays of both Affymetrix and Illumina have been used to characterize DNA copy numbers. We hypothesized that high-density Illumina methylation arrays, which are derived from the same genotyping principle as SNP arrays, can also be used to estimate DNA copy numbers.

Materials and Methods

The DNA of B-Lymphocyte from a healthy male donor, NA 10923, was obtained from Coriell, and the DNA of HT-29 colon cancer cell line (female donor) was obtained from ATCC.

The methylation of DNA samples were assessed using the Illumina HM450K array following the manufacturer's instructions. Raw hybridization images were analyzed using Illumina's BeadScan software with default parameters. The methylation status of each interrogated site was reported by signal A (intensity estimate of unmethylated DNA) and signal B (intensity estimated of methylated DNA). Following SNP arrays, we define the vector of total intensity R of all interrogated sites as

$$R = A + B,$$

where A and B are signal A and signal B, respectively. R is a proxy of DNA copy number. The variance of R was stabilized by $\log_2$ transformation. To make the $\log_2 R$ values across multiple arrays comparable, we used a constant-scaling normalization method to adjust the median of $\log_2 R$ on each array to a target value of 13 <ref>.

To derive a reference value of R for a normal diploid genome, we took the average of a cohort of 58 arrays that were normalized as described before. Each array measured the peripheral blood of a healthy male individual in a pesticide applicators study. We call this reference value as R_{REF}.

To adjust the probe effects, we define the relative copy number difference between the observation and the reference genome as the log2 R ratio (LRR), following the previous literature on SNP arrays <ref>,

$$LRR = \log_2(R_{OBS}/ R_{REF}) = \log_2 R_{OBS} - \log_2 R_{REF}$$

LRR was then subjected to the circular binary segmentation algorithm as implemented in the DNAcopy package in Bioconductor <ref>.

Results

The Illumina HM450K methylation array interrogates 485,577 sites in the human genome that include CpG sites (both in and outside of CpG islands) and some non-CpG sites. Thus, we collective call these sites are "interrogated sites" or simply "sites". We hypothesized that the total intensity R, which is the sum of signal A (intensity estimate of unmethylated DNA) and signal B (intensity estimated of methylated DNA) reflects the DNA copy number.

To characterize the reproducibility of R, we measured the same DNA sample of NA 10923 (B-Lymphocyte from a healthy male donor) on duplicated HM450K methylation arrays. Figure 1 suggests that log2 transformation is necessary to stabilize the variance across a range of R values. For all sites on the array, the median of the standard deviations is 0.20 (Figure 1B).

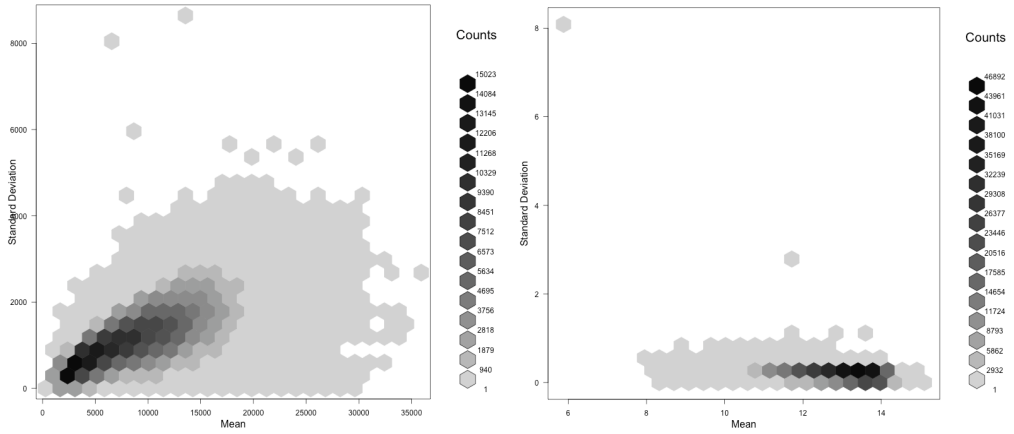

Figure 1. Mean and standard deviation of R on (A) original and (B) $\log_2$ scales. Each hexbin depicts the density of the points on a scatter plot where each point corresponds to a interrogated site on the HM450K array. There are a total of 485,577 points on each plot.

Previous studies of both Affymetrix and Illumina SNP arrays suggest that $\log_2$R can be a proxy of DNA copy numbers, but R can also be affected by the binding affinity of the probe designed for each interrogated site. As such, a direct plot of $\log_2$R against the genomic locations (Figure 2A) tends to be noisy. We utilized a reference $\log_2$R value derived from a cohort of 58 healthy individuals to derive the Log_2 R ratio (LRR), which is the Log_2 of the relative R ratio (R_{OBS}/ R_{REF}), where R_{OBS} and R_{REF} are the observation and the reference, respectively. Figure 2B suggests that LRR improves the signal to noise ratio when the spatial effect across multiple interrogated sites are assessed. We used the circular binary segmentation (CBS) algorithm implemented in the DNAcopy package in Bioconductor to segment the LRR.

To validate the DNA copy number estimates obtained from the HM450K methylation array, we compare the results with the golden-standard estimates using the Affymetrix mapping 100K array (115,389 SNP sites). Figure 2 suggests that the results are highly consistent. Due to the higher density of the HM450K array, we also observe the potentially true copy number aberrations of smaller fragments (Figure 2).

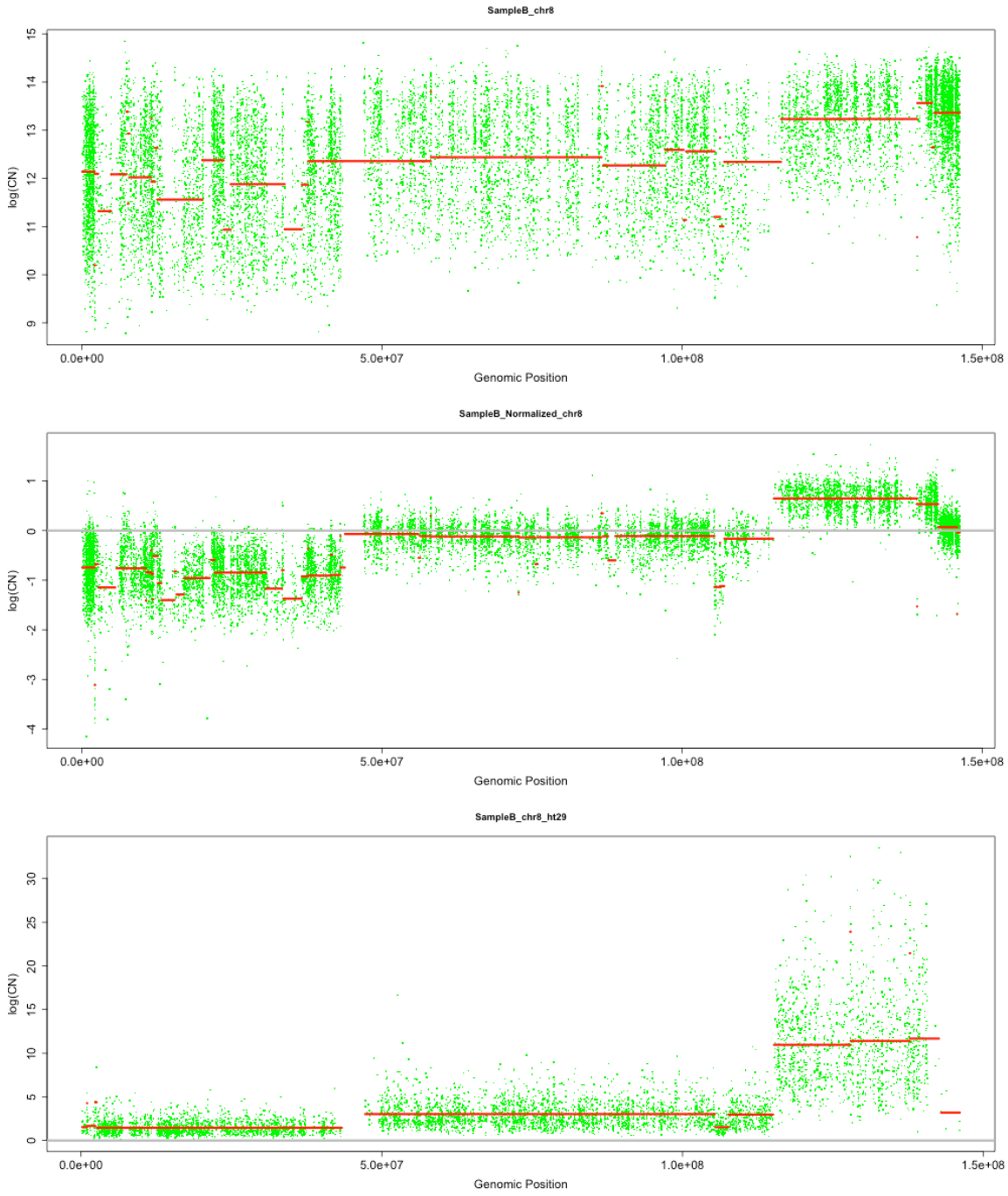

Figure 2. Copy number estimates of chromosome 8 of the HT-29 cell line using (A) $\log_2$R of HM450K, (B) log2 R ratio(LRR) of HM450K, and (C) Affymetrix SNP 100K array. Segmented results by the circular binary segmentation (CBS) algorithm are shown in redlines.

Discussion

The Illumina HM450K BeadChip uses both type I and type II chemistry on the same array (Table 1). We observed significant differences in their signals (Figure 3). As such, we only utilized type II sites in the current study. The calibration of type I and type II sites needs to be further investigated.

Table1 . Number of type I vs. type II sites on the HM450K BeadChip.

Chemistry	Number	Percent
Type I	13,5501	27.9%
Type II	35,0076	72.1%
Total	485,577	100%

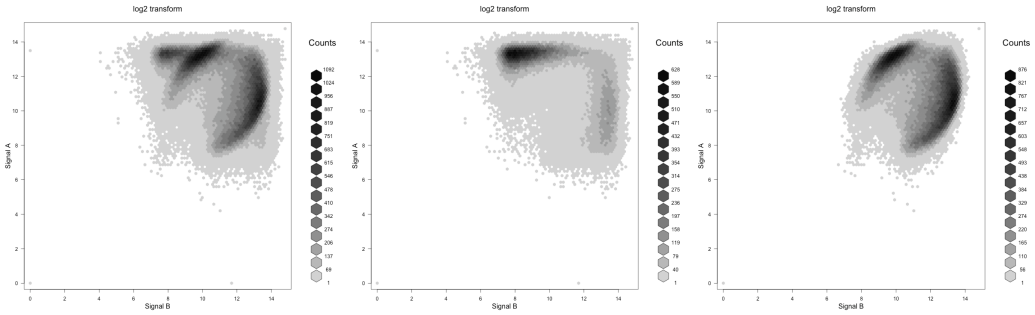

Figure 3. Signal A and signal B of sample NA 10923. (A) All sites on the BeadChip, (B) Sites measured by type I chemistry, (C) Sites measured by type II chemistry.

References

To be updated.