Improvements To Existing QC Tools For Methylation 450K Arrays

J. Romm, E. Pugh, K.F. Doheny

Center for Inherited Disease Research (CIDR), IGM, JHU-SOM, Baltimore, MD

Introduction

The Center for Inherited Disease Research (CIDR) provides high quality Next-Gen Sequencing (NGS), Genotyping and statistical genetics consultation to investigators working to discover genes that contribute to disease. At CIDR we use automated, real-time QC reports for genotyping and sequencing projects which allows the lab group to examine and compare dozens of metrics to detect failed or problematic samples and determine what, if anything, they share in common (e.g. failed array, position on a plate, DNA source). We wanted to create a similar report for Illumina's HumanMetylation450 array to supplement the QC information available within Illumina's raw data files.

Methods

To accomplish this we used the R and BioConductor package Minfi[1] to create an R script that calculates metrics for each sample that are not easily available in GenomeStudio, such as mean and standard deviation of beta, intensity of the methylated and unmethylated probes, detection P value, MDS distances, the mean and standard deviation of the values for Illumina's control probes and the mean and standard deviation of a subset of 25 methylated and unmethylated probes from the Y chromosome. We order the Y probes by beta intensity in decreasing order and use probes 6-30. We then create an R Data.Frame with the metrics and write them out as a delimited file. While most of this information is available in pdf reports available in Minfi, we found those reports cumbersome for large (100's to 1000's of samples) projects. This format allows us to combine the metrics with additional information about the sample, such as location on a plate and lab variables such as instrument used for processing. The combined information is imported into Spotfire® and used by the lab to determine outlying samples. This information is also provided to the investigator. With the investigator, we jointly determine whether the outliers represent sample failures or a subset of samples within the study that are meant to be outliers from a larger group (e.g. different sample preparation). Sample failures are redone one time in the lab and the metrics are recalculated and released to the investigator as part of our data release. Examples of some metrics evaluated are shown here.

Project 1 had ~3000 samples, which included fully methylated and fully unmethylated controls for QC. We were able to identify 3 full array failures and one sample with low overall intensity.

Project 2 had ~1000 samples, which also included fully methylated and unmethylated controls for QC. 12 % of the samples had less than 99% of CpG islands detected.

Project 3 had ~96 samples. We were very easily able to determine 2 groups of samples, corresponding to gender of the samples assayed.

Results

Project 1 had ~3000 samples, which included fully methylated and fully unmethylated controls for QC.

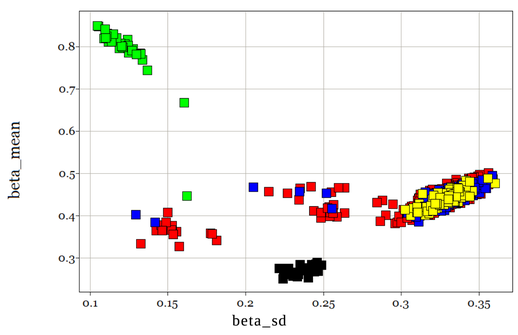

Figure 1:
Beta mean and beta standard deviation.
Green 100% methylated controls
Black 0% methylated controls
Red - coded as BWA by investigator
Blue - coded as CGF by investigator
Yellow - coded as DUP by investigator

Samples not in the main clusters were investigated as possible QC problems, and re-bisulfite converted and rerun.

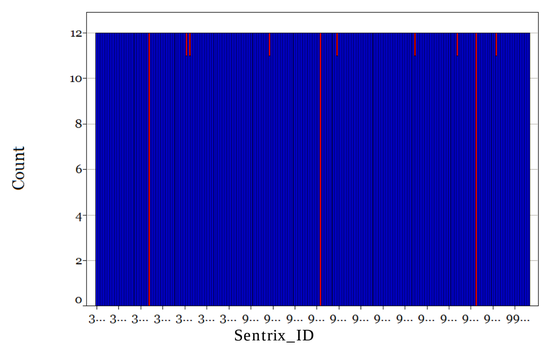

Figure 2:
Plot of the problematic samples by methylation array showing 3 failed arrays (entire bar is red) and random sample failures (red and blue in a bar).

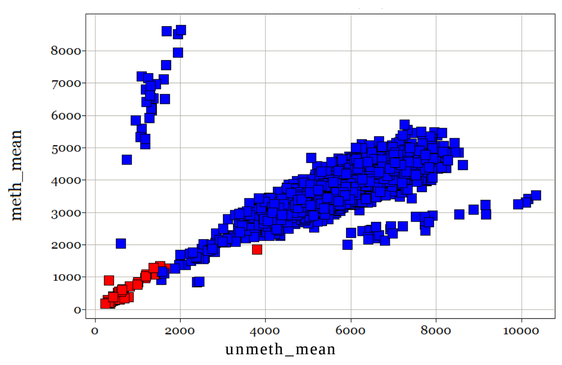

Figure 3:
Problematic (red) and other samples by the mean intensity of the methylated and unmethylated probes. Most of the problems are for samples where the both intensities were very low.

Project 2 had 1152 samples run, including 0%, 30%, 70% and 100% methylated controls

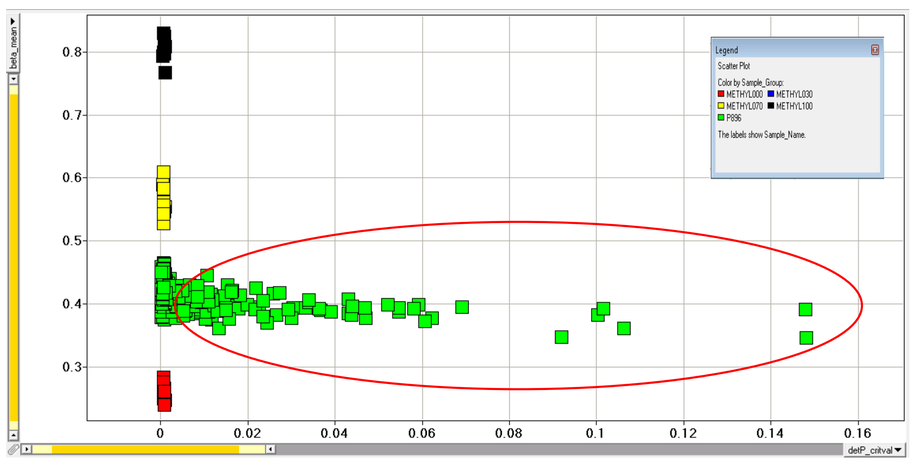

Figure 4:
12% of this project's samples had less than 99% of CpG islands detected. These were scattered across arrays and plates processed in the lab, indicating a sample difference. Plotted is detection P- value by beta mean. Also separated nicely are the project samples and the varying % methylated controls.

Project 3 was 96 samples, including 100% methylated and unmethylated controls.

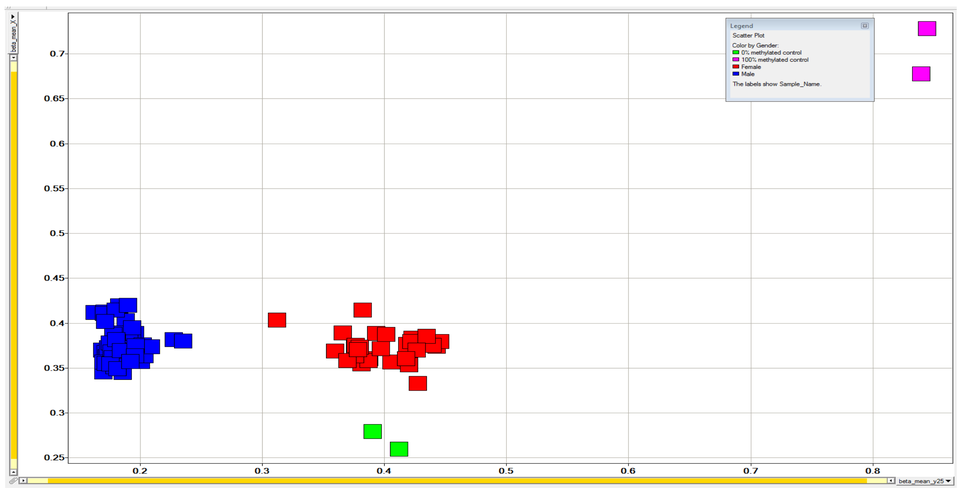

Figure 5:
Figure 5 is a plot of beta mean of the X chromosome by beta mean of a subset of y chromosome probes (y25). Samples nicely separate into 2 main clusters which correspond to gender. The fully methylated and fully unmethylated controls groups are also separated.

Conclusion

Using Minfi to gather quality control statistics for Illumina Methylation 450K array data is quite effective. Although some of the basic statistics regarding intensity and control codes are present in GenomeStudio, they are difficult to sift through for large projects. Many other important statistics and simply not available in GenomeStudio such as mean and standard deviation of beta, intensity of the methylated and unmethylated probes, detection P value, MDS distances, the mean and standard deviation of the values for Illumina's control probes and the mean and standard deviation of a subset of methylated and unmethylated probes from the Y chromosome. Using MinFi and Spotfire, we are able to verify the gender of the samples, control performance and relative % methylated related to the project data and easily detect outliers for the metrics described above. Samples can then be identified for redos and plate and array issues can be addressed.

References:
Martin J Aryee, Andrew E Jaffe, Hector Corrada Bravo,et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics, 30(10):1363–1369, 2014. doi:10.1093/bioinformatics/btu049, PMID:24478339.
Spotfire: http://spotfire.tibco.com/