Improving Methods for High Throughput Epigenetic Studies

B Marosy[1], H Ling[1], P Zhang[1], J Gearhart[1], J Paschall[1], C Ongaco[1], M Mawhinney[1], L Aker[1], A Robinson[1], B Craig[1], Z Xu[2], K Doheny[1]

[1]Center for Inherited Disease, Johns Hopkins Genomics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD

[2]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC

Introduction

The Center for Inherited Disease Research (CIDR) provides high-throughput genomic services and statistical genetics consultation to investigators working to discover genes that contribute to diseases. In response to the growing requests for epigenetic studies, we recently evaluated an alternative chemistry for methylation detection - an enzymatic library conversion of methylated sites (NEB EMseq) paired with a targeted methylome panel (Twist BioSciences). The performance of this method (Methyl-Seq) was compared with our existing methylation array (EPIC), which uses bisulfite conversion.

Our current EPIC array studies typically involve hundreds to a few thousand samples. However, the size of recently funded studies is expanding to the tens of thousands, necessitating the implementation of automated bisulfite conversion methods. We have successfully validated the Zymo EZ-96 DNA Methylation Lightning assay on our in-house automation, in preparation for population-scale studies.

Lastly, we have automated EPIC array sample level QC to date using *ewastools* and *minfi* packages. We are exploring the *ENmix* package to allow for a more in-depth QC on both sample and probe level, with the aim of providing our investigators with a dataset closer to "analysis ready."

Methods – Experimental Samples

Control samples were obtained from EpigenDx to provide methylated DNA at 0%, 10%, 25%, 50%, 75% & 100% and HapMap Ashkenazim trio from Coriell (NA24143/HG004, NA24149/HG003, NA24385/HG002) Table 1 – lists the samples used in the comparison and how many replicates per method.

List of Samples	MethySeq	EPIC_array.v2
METHYL000	2	5
METHYL010	2	3
METHYL025	2	3
METHYL050	2	3
METHYL075	1	3
METHYL100	1	6
NA24143 (HG004)	2	3
NA24149 (HG003)	2	3
NA24385 (HG002)	2	3
Totals	**16**	**32**

Methods – Methyl-Seq

200ng of DNA was sheared to 300bp using the Covaris LE220. The NEBNext Enzymatic Methyl-Seq kit was then used to end repair and ligate adapters for library preparation. Un-methylated CpG positions were enzymatically converted and the subsequent libraries were amplified with index specific primers. Converted, indexed libraries were then pooled (200ng/sample & 8 samples/pool) for targeted capture using the Twist Fast Hybridization Protocol and Methylome probe panel. All hybridization and post-hyb capture & washes were performed according to the manufacturer's protocol. Post Hyb PCR was performed using 6 cycles. Enriched libraries were sequenced at 2x100 on the Illumina NovaSeq6000 platform. Sequencing data was processed using bcl2fastq to generate fastq files from raw data. Trim galore was used to trim adapters from the reads followed by alignment using bwa-meth and methylation calls generated by MethyDackel.

Methods - EPIC v2

500ng of DNA was used as input into the EZ DNA Methylation-Lightning Automation kit (Zymo) for conversion of unmethylated sites using bisulfite conversion. All steps were processed according to the manufacturer's protocol and were adopted for automation using the Biomek i5 liquid handler (Beckman Coulter). All volume from the eluted converted samples was used as input into the Infinium Methylation EPIC v2 BeadChip array (Illumina). Array data was analyzed using *minfi* (v1.34.0), *ewastools* (v1.7.2).

Results – Data QC

The Methyl-Seq and Illumina EPIC assays are designed to capture 3.98M and 935K CpG sites respectively. On average the Methyl-Seq data detected 6.5M CpG sites per sample with a mean coverage of 25-55x and 85-90% mapping efficiency. Of the 3.98M CpG sites in Methyl-Seq 741K overlap with EPIC v2. When comparing the detected 6.5M CpG sites, 784K overlap with EPIC v2.

Conversion rates for the two 0% methylated controls were calculated for the methyl seq data at 99.88% and 99.87%. For the automated EPICv2 array, 99.6% of CpGs were detected at p=0.01.

Results – Correlation

Pearson correlation was used to determine R2 when comparing HapMap non-duplicate and duplicate pairs, both within and between platforms. When comparing Methylation controls root mean square deviation (RMSD) was used to assess 0% and 100% methylation controls as R2 is more sensitive to background noise in these controls.

Methyl-Seq – HapMap

- Between subject correlation/non-duplicate pairs: 89-91% (no filter); 92-94% (50x filter)
- Within subject correlation/duplicate pairs: >98% (no filter); >99% (50x filter)

EPIC v2 – HapMap

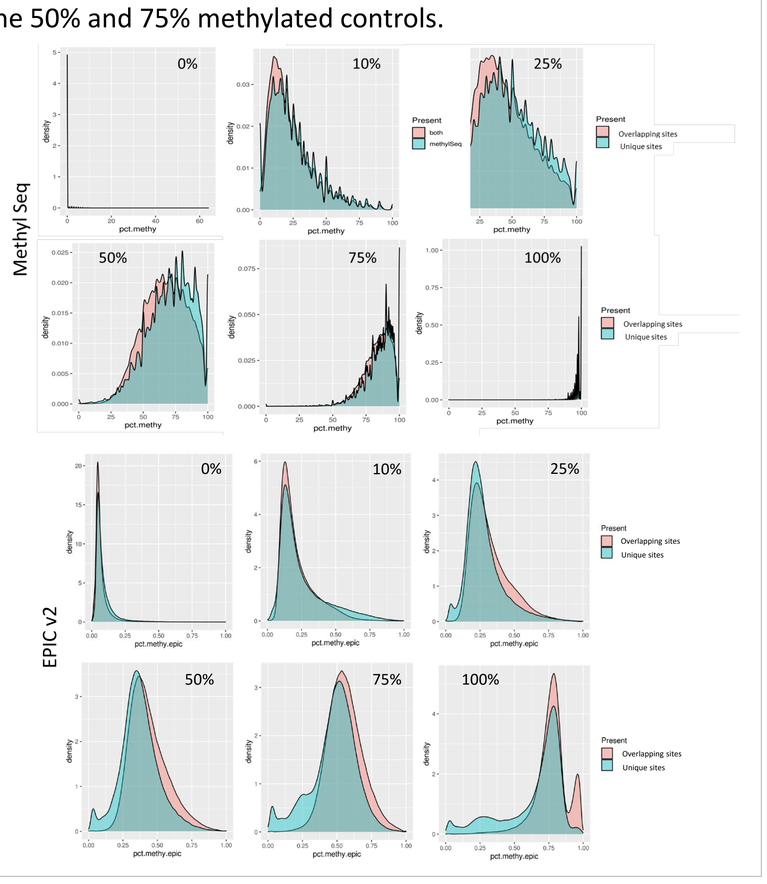
- Between subject correlation/non-duplicate pairs: 88-90%
- Within subject correlation/duplicate pairs: >99%

Methyl-Seq & EPIC v2

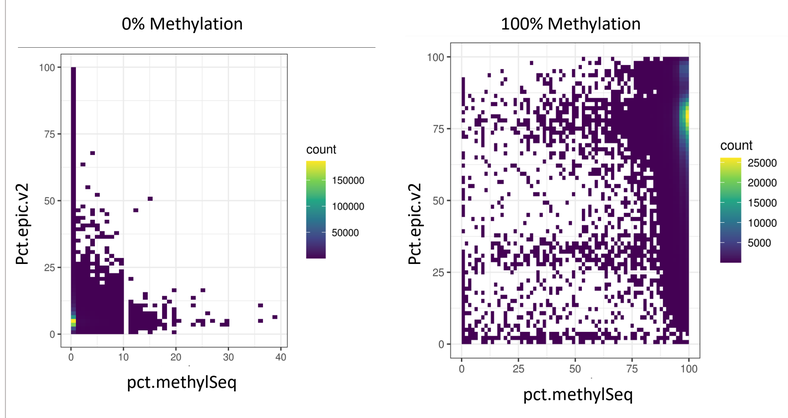
- HapMap Pairs: 96-97%
- Methylation control pairs (0 & 100%)
  - EPIC v2: 0.1318 and 0.2881, respectively
  - Methyl-Seq: 0.0099 and 0.0784, respectively

Results – Distribution of Methylation Level Estimates from Methyl-Seq and EPIC Array

The figure below shows the distribution of methylation level estimates for each methylated control processed with Methyl-Seq (top) vs EPICv2 array (bottom). Pink represents overlapping sites between Methyl-Seq or EPICv2 Array and blue represents sites unique to the respective platform. The difference in methylation level estimates are mostly driven by the technologies rather than by the sites. The distributions are quite different between the two platforms, but similar between the overlapping and unique sites within each technology. Compared to array, Methyl-Seq has lower background noise for 0% and 100% methylated controls, but tends to overestimate the 50% and 75% methylated controls.

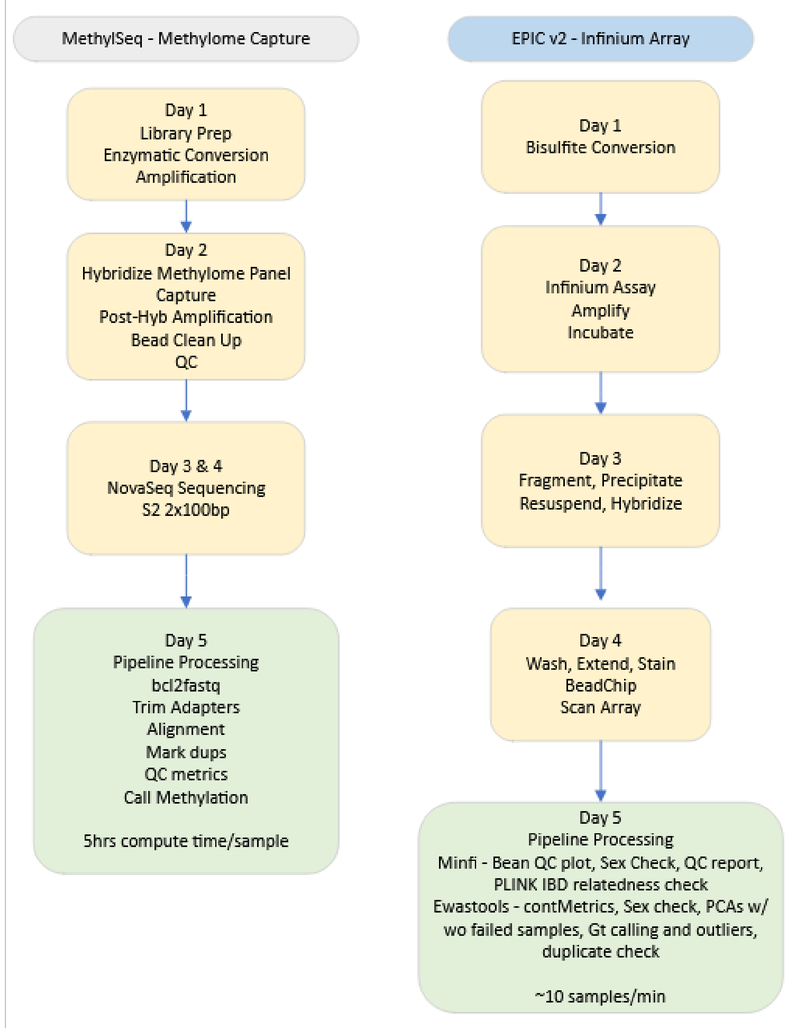


Results – Discordance



The heat map depicts discordance between the two methods, with the percent of methylation in Methyl-Seq data set represented on the x-axis, and the EPICv2 array data set is represented on the y-axis. In the left panel, 0% methylation is called similarly between the two methods ie low discordance. In the right panel, 100% methylation is being called at a higher percentage in the Methyl-Seq data compared to the EPICv2 array data. Overall, array has more noise compared to Methyl-Seq, and tends to over estimate 0% methylation and underestimate 100% methylation controls.

Workflow



Both workflows take a similar amount of time (4 days) on the wet bench (yellow squares). Post Data analysis is more time consuming when processing Methyl-Seq data compared to Infinium EPICv2 data (green squares). We've developed the post-array workflow for Infinium data using *minfi* to obtain the beta values, the number of detected CpG sites detected at p-values of 0.01 and 0.05. Additional QC includes sex check, Principle Components Analysis (PCA), duplicate and relatedness checks. We also use *ewastools* to obtain similar measures for comparison, all of which are combined in a QC report in csv format.

Discussion & Conclusions

- Methyl-Seq covers 7x more CpG sites than EPIC array, with reagent costs ~1.7x more for 50x coverage and ~2x more for 100x coverage compared to EPICv2 costs.
- EPIC array provides beta values for a pre-determined set of CpG sites; Methyl-Seq is unbiased, though CpG sites can vary by sample.
- Concordance is similar between and within Hapmap pairs for each method. RMSD is lower in Methyl-Seq than EPICv2 data when comparing 0% and 100% methylated controls.
- Data analysis is more complex and time-consuming for Methyl-Seq data.
- Array may be a better choice for larger studies needing to evaluate known CpG sites or to combine/compare results with previous studies. Methyl-Seq is more expensive, but offers exploration of novel CpG sites.

Next Steps:

- We plan to evaluate reproducibility of *ENmix* and other methods in a larger project with more duplicate pairs, as the current dataset may be of insufficient size to use methods like intra-class correlation coefficient (ICC).
- Comparison breakdown by annotation & site coverage
- Absolute beta difference for methylation controls