Reviewer's report

Title: Batch Effect Correction for Genome-Wide Methylation Data with Illumina Infinium Platform

Version: 1 Date: 30 August 2011

Reviewer: Shuang Wang

Reviewer's report:

General comments:

The authors compared three normalization methods developed for gene expression data on normalization of DNA methylation data. The authors worked on three datasets which are considered to have small batch effect, median batch effect and large batch effect. The authors further tried empirical Bayes normalization method also developed for gene expression data and concluded that some further normalization may need on DNA methylation data in order to remove the batch effect. Although the topic is interesting and needed when more researches are being conducted to examine DNA methylation, there are several problems with the manuscript.

My specific comments are:

1. On page 4, the authors said “no such method is available for methylation data and the appropriateness of these methods for the methylation data has not been evaluated”. However, the authors did not develop a new method to correct batch effect just for DNA methylation data but compared existing methods developed for gene expression on DNA methylation. This statement is misleading. Also, unlike gene expression, DNA methylation can be completely different for normal individuals.

2. The authors made conclusion about the batch effects without looking into the data with details and more carefully. For example,

   a. For dataset 1, did the authors look at in detail what samples were on chip22 and chip26 to see if the samples differ in terms of some characteristic variables like age? DNA methylation is known to change with age and other environmental conditions. Observing chip22 and 26 to be separated with PCA does not justifying the reason is batch effect. The authors should explore further.

   b. The same comment for datasets 2 and 3.

3. If the authors can conduct some simulation studies where only batch effects are introduced, but also introducing effects of other characteristics like age, but no association with the outcome of interest, the results will be much more meaningful and interesting.

4. The authors should comment on the possibilities of over normalization.
Especially, for datasets 2 and 3, “For Dataset 2 and 3, the EB correction alone was very effective in removing batch effects as no principal components or CpGs were associated with batch effects anymore after the step (Table 1). However, the numbers of CpGs associated with the outcome of interest were all lower (1,079 and 2,047 respectively for Dataset 2 and 3) compared to normalization and EB corrected data”. Is it possible that normalization procedure has removed some signals? The authors should comment. In Illumina Beadstudio menu, it says some normalization methods developed for gene expression can not be used for DNA methylation because the underlying assumptions are different. With gene expression, the assumption is most of the genes are not differentially expressed, while we can not make this assumption for DNA methylation. The authors also noted this in the manuscript, but should discuss/explain/explore the three methods +EB examined regarding the assumptions.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.