Review of: "BSImp: Imputing Partially Observed Methylation Patterns for Evaluating Methylation Heterogeneity"

Hongmei Zhang[1]

1 University of Memphis

Potential competing interests: The author(s) declared that no potential competing interests exist.

The authors developed an imputation approach based on estimated probabilities for a methylation site being fully methylated. The probability calculation is based on an assumption that the pattern of methylation is similar for cells within a population and an assumption that the behavior of cells or the methylation statuses of a cell at a given position can be predicted by the statuses nearby and cells nearby. By applying the methods to whole genome bisulfite sequencing (WGBS) data and reduced representation bisulfite sequencing (RRBS) data, it is demonstrated that the proposed method can increase coverage by up to 15% with higher accuracy compared to the existing approaches for methylation pattern recovery (PReLIM and column mean) and methylation level imputation (lower depth and METHimpute). Methylation imputation is an important research field and has a strong potential to benefit epigenetic studies.

Multiple aspects may need to be considered should the authors decide to further improve the method or their developed computing programs, which are outlined below:
1) How sensitive are the results to the selection of window size? What is the optimized window size? A thorough assessment will be greatly appreciated and beneficial to users.
2) Equation (1) (the manuscript noted equations (1), (2), and (3), but it should be just one equation) is for the probability of methylation status at site j. In the Results section of data analyses, the authors also presented results on methylation level accuracy (bias). It is unclear how methylation level is imputed. If it is from equation (1), that is, using the probability of methylation status as a methylation level, then the bias is expected to be larger. I wish the authors had provided more details.
3) Many DNA methylation association studies rely on array-based data, e.g., Illumina EPIC arrays or 450K arrays. Often, the data are generated from multiple batches with each batch having some CG sites with missing values due to various reasons. Will this approach be applicable to that type of DNA methylation data?
4) I had wished a discussion on computing expenses and a comparison with existing methods on this regard.
5) It is great that the authors posted Python programs with instructions on GitHub. It may help more users if an R package can be developed.

I think this article has a typo, which may mislead or confuse some readers. The right side of Equation (1) (again, the manuscript noted equations (1), (2), and (3), but it should be just one equation) should be m_{i,j} rather than m_{-i,j}, i.e., p(m_{i,j}=1|p_{i,-j}=s_1)p(p_{i,-j}=s_1)+ …+ p(m_{i,j}=1|p_{i,-j}=s_n)p(p_{i,-j}=s_n). Currently, it is written as p(m_{-

i,j}=1|p_{i,-j}=s_1)p(p_{i,-j}=s_1)+ …+ p(m_{-i,j}=1|p_{i,-j}=s_n)p(p_{i,-j}=s_n), which is not equal to the left side, p(m_{i,j}=1).

Overall, I enjoyed reading the article and believe the proposed method and published GitHub program will benefit researchers in this field.