Bioinformatics, 31(9), 2015, 1519–1520
doi: 10.1093/bioinformatics/btv090
Advance Access Publication Date: 13 February 2015
Letter to the Editor

OXFORD

Sequence analysis

A cautionary note on using binary calls for analysis of DNA methylation

Agnieszka Prochenka[1,2,*], Piotr Pokarowski[3], Piotr Gasperowicz[2], Joanna Kosińska[2], Piotr Stawiński[4], Renata Zbieć-Piekarska[5], Magdalena Spólnicka[5], Wojciech Branicki[6] and Rafał Płoski[2,*]

[1]Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, Warsaw, Poland, [2]Department of Medical Genetics, Medical University of Warsaw, Pawińskiego 3c, Warsaw, Poland, [3]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, Warsaw, Poland, [4]Department of Immunology, Center for Biostructure Research, Medical University of Warsaw, Banacha 2, Warsaw, Poland, [5]Central Forensic Laboratory of the Police, Aleje Ujazdowskie 7, Warsaw, Poland and [6]Institute of Forensic Research, Westerplatte 9, Krakow, Poland

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on November 5, 2014; revised on January 16, 2015; Accepted February 6, 2015

Contact: a.prochenka@phd.ipipan.waw.pl or rploski@wp.pl

In this article, 'A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data' Cheng and Zhu proposed a classification procedure based on a mixture of binomial model to make binary calls for methylation status. Whereas we find this approach interesting and competitive to method described in Lister *et al.* (2011) under the used assumptions, we advise to use the proposed methodology with caution because we disagree with some of the generalization made. Namely, the authors state that Cytosine (C) positions can be either methylated or not and that ideally after polymerase chain reaction amplification there are only C or T reads for each covered C position of interest, depending on the methylation status. In particular, it is stated that obtaining C reads at unmethylated sites can be observed only due to incomplete conversion, sequencing or other systematic errors.

Whereas true in many situations, these assumptions have important exceptions. Methylation can indeed be described by a binary variable when a single C in one DNA molecule is considered. However, even in a single diploid cell all the autosomal chromosomes exist in two copies and at some loci methylation may occur at one but not the other chromosome with imprinted loci being a well recognized example (Adalsteinsson and Ferguson-Smith, 2014). The heterogeneity of methylation at a given C can be even more prominent when mixture of DNA molecules extracted from a population of cells is studied (the most popular experimental design). In such setting when many reads are obtained at a site (sequencing depth higher than one), Cs in some DNA molecules can be methylated and in some not generating a proportion which reflects a true biological phenomenon rather than technical artifacts.

Using such proportions for prediction of phenotypic features, for example age, has been widely described in the literature. In Hannum *et al.* (2013) and Garagnani *et al.* (2012), a microarray-based methylation analysis shows that treating methylation as a continuous variable is effective for identifying the correlation between subject age and methylation. In Garagnani *et al.* (2012) proportions of methylated Cs were used to predict age of 64 subjects and the *ELOVL2* gene showed a progressive increase in methylation with age with the *Spearman's* correlation coefficient equal to 0.92.

The dependence between age and methylation proportions for the purpose of the age prediction is also the subject of our own studies. Using yet another technology (PyroMark platform) to analyse full blood, we recently confirmed usefulness of *ELOVL2* methylation as an age marker—the final linear regression model included two Cs in *ELOVL2* and enabled prediction with $R^2 = 0.859$ (Zbieć-Piekarska *et al.*, 2014). Finally, we observed similar dependences using reduced representation bisulfite sequencing (RRBS), a method directly discussed by Cheng and Zhu. A plot of methylation frequency against age for the *ELOVL2* gene as obtained from RRBS versus the methylation status calling (MSC) procedure described in Cheng and Zhu (2014) is presented in Figure 1. Observe that in our data, according to the MSC procedure, the position is 'methylated' if and only if the methylation frequency directly counted from RRBS is greater than 0.45.

Incomplete conversion (for our data estimated as ∼0.3%), sequencing and systematic errors cannot be responsible for these findings reported in several studies using different methods. Thus, we would like to argue that making binary calls as proposed in

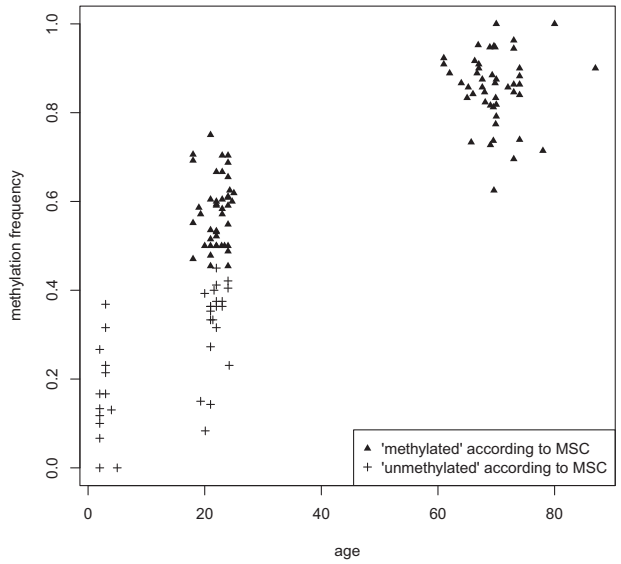

Fig. 1. A plot of methylation frequency/status against age for the *ELOVL2* C (position 11044867, chromosome 6) as obtained directly from RRBS versus the MSC procedure. Each data point illustrates methylation frequency calculated by read counts from the RRBS experiment with the *Spearman's* correlation coefficient equal to 0.86. Triangles and crosses denote samples whose methylation status was defined as 'methylated' or 'unmethylated' by the MSC procedure, respectively

Cheng and Zhu (2014) should, at least in some cases, be used cautiously remembering about the continuous alternative.

Funding

The study was supported by a grant from the National Centre for Research and Development, no. DOBR/0002/R/ID1/2012/03 and by research fellowship within 'Information technologies: research and their interdisciplinary applications' agreement number POKL.04.01.01-00-051/10-00.

Conflict of Interest: none declared.

References

Adalsteinsson,B.T. and Ferguson-Smith, A.C. (2014) Epigenetic control of the genome lessons from genomic imprinting. *Genes*, **5**, 635–655.

Cheng,L. and Zhu,Y. (2014) A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data. *Bioinformatics*, **30**, 172–179.

Garagnani,P. *et al.* (2012) Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*, **11**, 1132–1134.

Hannum,G. *et al.* (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell*, **49**, 359–367.

Lister,R. *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68–73.

Zbieć-Piekarska,R. *et al.* (2015) Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Sci. Int. Genet.*, **14**, 161–167.