Genes & Diseases (2023) 10, 698–700

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases

RAPID COMMUNICATION

# A revision of the InfiniumPurify R package for genome-wide correction of tumor purity in Infinium DNA methylation array data

Understanding DNA methylation (DNAm) alterations in tumors is important to develop novel therapeutic targets and powerful biomarkers, and to gain insights into tumorigenesis processes. As many studies have highlighted the confounding effects of tumor purity in DNAm analyses,[1–3] various tools have also been developed to tackle the issue.[2–4] Among them, InfiniumPurify R package is a popular and widely cited tool,[2,3,5] consisting of a set of statistical methods to estimate and account for tumor purity in cancer DNAm analyses.

While many previous studies did not take into account tumor purity in genomic analyses, a common practice to alleviate biases introduced by tumor purity is including estimated purity as a continuous covariate in regression models in differential analyses. However, commonly used tools for differential methylation (e.g., Probe Lasso) or unsupervised clustering analysis (e.g., ConsensusClusterPlus) do not have the capacity to model covariate variables. In addition, tumor purity might have multiplicative rather than additive effects in differential expression/methylation analyses.[3] Therefore, a genome-wide tumor purity correction in DNAm could potentially overcome these shortcomings. To this end, InfiniumPurify introduces a function to correct for tumor purity for all CpG probes based on linear regression.[5]

Intuitively, for differentially methylated CpG probes, the uncorrected tumor beta values are expected to lie between the normal and the purity corrected beta values. This is because the contamination attenuates the difference between normal and "purified" tumor tissue samples. More specifically, if the CpG probes are hypermethylated in uncorrected tumors compared to normal, the difference between tumors and normal should be amplified after being "purified", i.e., the corrected beta values should be higher than in uncorrected tumor. The same pattern is expected for hypomethylated CpG probes in tumors, but in the opposite direction. For CpG probes whose tumor beta values do not differ to those in normal, tumor purity correction is not expected to have substantial effects on the beta values.

In Figure 1, Qin et al from InfiniumPurify,[5] CpG probes cg00340958 and cg02237119 displayed slightly higher beta values in uncorrected tumors compared to normal. However, after purity correction with InfiniumPurify, the "purified" tumors have much lower beta values than uncorrected tumors and normal. This observation is inconsistent with the aforementioned expected pattern of purity correction in DNAm. We conjecture that this error occurred due to the wrong purity correction model used in the InfiniumPurify package.

In the InfiniumPurify function (https://rdrr.io/cran/InfiniumPurify/src/R/InfiniumPurify.R), a linear regression model was used to correct for tumor purity for each CpG probe $i$ in samples $s$ of $t$ tumors and $n$ normal:

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ y'_{i1} \\ \vdots \\ y'_{in} \end{bmatrix} = m \begin{bmatrix} 1-\pi_1 \\ \vdots \\ 1-\pi_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} + c + \varepsilon_s(*),$$

where $(y_{i1}, \cdots, y_{it})$ are the arcsine transformed uncorrected beta values for tumors, and $(y'_{i1}, \cdots, y'_{it})$ are for normal samples. Here, $\pi_t$ represents the estimated sample purity for tumor $t$, $m$ is the regression coefficient, $c$ is the intercept, and $\epsilon$ is the residue of the regression model.

Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2022.08.003

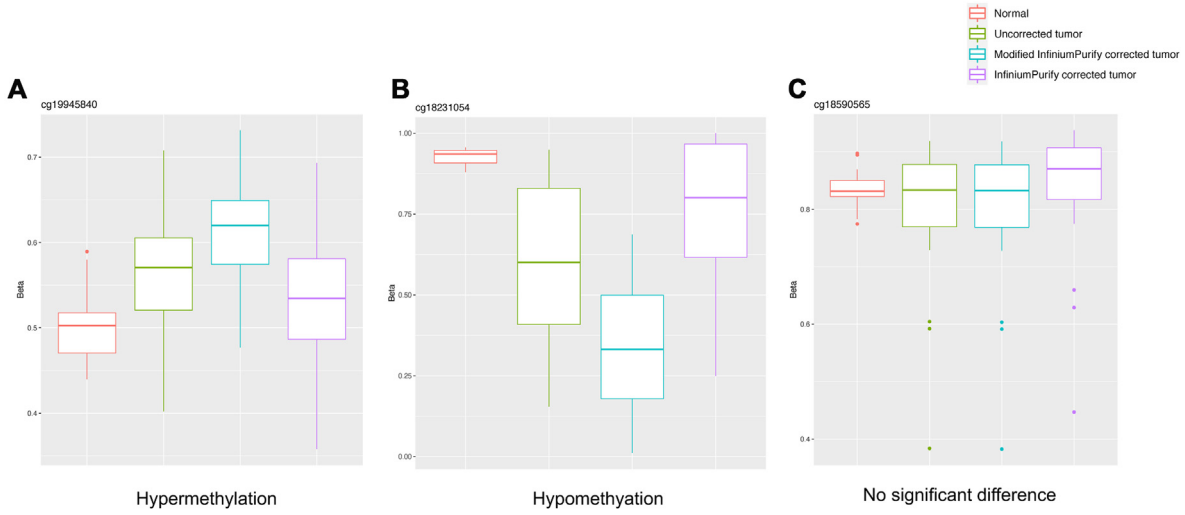


**Figure 1** Comparing the effects of genome-wide tumor purity correction by the original InfiniumPurify and modified InfiniumPurfy functions on beta values of selected probes. The raw data was provided as an example from InfiniumPurify R package. **(A)** Example of a probe that is hypermethylated in uncorrected tumors compared to normal. **(B)** Example of a probe that is hypomethylated in tumors compared to normal. **(C)** Example of a probe that is not differentially methylated between uncorrected tumors and normal. See Supplementary Code for more details.

The purity corrected arcsine transformed beta values $Z_{it}$ for tumor $t$ was then calculated as. $Z_{it} = \widehat{c} + \widehat{\epsilon_t}$.

However, if we assume that normal samples are not contaminated with tumor contents (i.e., $\pi_n = 0$), equation $(*)$ should correspond to:

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{it} \\ y'_{i1} \\ \vdots \\ y'_{in} \end{bmatrix} = m \begin{bmatrix} 1-\pi_1 \\ \vdots \\ 1-\pi_t \\ 1 \\ \vdots \\ 1 \end{bmatrix} + c + \epsilon_s$$

After modification, we applied this to an example subset of lung cancer adenocarcinoma data from The Cancer Genome Atlas (TCGA) provided within InfiniumPurify R package (Supplementary Code), we were able to obtain the expected patterns of corrected beta values (Fig. 1; blue box plots are modified version and purple box plots are computed by InfiniumPurify). For hypermethylated (Fig. 1A) and hypomethylated probes (Fig. 1B), InfiniumPurify correction could diminish the differential methylation signal originally observed between uncorrected tumors and normal. For CpG probes that do not show differential methylation between uncorrected tumors and normal (Fig. 1C), InfiniumPurify correction could result in these probes erroneously being identified as differentially methylated.

In summary, we modify the code of InfiniumPurify R package used to correct for purity in genome-wide DNAm data. The revised function can be downloaded from the Supplementary Code. Given that the package has been used in many publications, we warn researchers the potential consequences for downstream analyses. In addition, we clarify the expected patterns of tumor purity correction's effects on DNAm beta values. We take this opportunity to urge researchers in the field to take additional care when accounting for tumor purity in their future analyses.

## Conflict of interests

The author declares no conflict of interests.

## Funding

This work was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, USA.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2022.08.003.

## References

1. Sun W, Bunn P, Jin C, et al. The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic Acids Res*. 2018;46(6):3009—3018.
2. Zhang W, Feng H, Wu H, et al. Accounting for tumor purity improves cancer subtype classification from DNA methylation data. *Bioinformatics*. 2017;33(17):2651—2657.

3. Zheng X, Zhang N, Wu HJ, et al. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol*. 2017;18(1):17.
4. Zheng X, Zhao Q, Wu HJ, et al. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol*. 2014;15(8):419.
5. Qin Y, Feng H, Chen M, et al. InfiniumPurify: an R package for estimating and accounting for tumor purity in cancer methylation research. *Genes Dis*. 2018;5(1):43–45.

Phuc H. Hoang

*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA*

*E-mail address:* phuc.hoang@nih.gov

1 July 2022
Available online 22 August 2022