Published in final edited form as:

Nat Methods. 2017 February 28; 14(3): 218–219. doi:10.1038/nmeth.4190.

Correcting for cell-type heterogeneity in DNA methylation: a comprehensive evaluation

Elior Rahmani¹, Noah Zaitlen², Yael Baran³, Celeste Eng², Donglei Hu², Joshua Galanter^{2,4,5}, Sam Oh², Esteban G Burchard^{2,4}, Eleazar Eskin^{6,7}, James Zou⁸, and Eran Halperin^{6,9}

¹Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv, Israel

²Department of Medicine, University of California, San Francisco, San Francisco, California, USA

³Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot, Israel

⁴Department of Bioengineering and Therapeutic Science, University of California, San Francisco, San Francisco, California, USA

⁶Department of Computer Science, University of California, Los Angeles, Los Angeles, California, USA

⁷Department of Human Genetics, University of California, Los Angeles, Los Angeles, California, USA

⁸Department of Biomedical Data Science, Stanford University, Palo Alto, California, USA

⁹Department of Anesthesiology and Perioperative Medicine, University of California, Los Angeles, Los Angeles, California, USA

To the Editor

Zheng *et al.*¹ discuss potential pitfalls in our evaluation of ReFACTor², a reference-free method to account for cell-type heterogeneity. Below, we reproduce their analysis and demonstrate that conclusions cannot be drawn on the basis of their results owing to conceptual and technical flaws in their analysis. We show with our reanalysis and further evidence from experiments on a total of 10 data sets that ReFACTor has improved performance over alternative methods, including the reference-based method of Houseman *et al.*³.

Zheng *et al.*¹ claim that more evidence needs to be provided to determine whether ReFACTor is applicable to tissues other than blood. They generated a “gold standard” set of

⁵Current address: Genentech, South San Francisco, CA, USA

Data availability statement: All data analyzed are publicly available. See Supplementary Methods for details.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Author Contributions: EH and ER designed the experiments. ER performed the experiments. ER, NZ, YB, CE, DH, JG, SO, EGB, EE, JZ, and EH wrote the manuscript.

Competing Financial Interests: The authors declare no competing financial interests.

“true positives” and “true negatives” for breast cancer differentially methylated CpGs (DMCs) and compared ReFACTor to Surrogate Variable Analysis (SVA)⁴ using EWAS data. There are multiple problems with this analysis. First, the list of “true positives” is unreliable owing to the fact that only two control individuals were used for its construction (Supplementary Note 1). We show through a simple permutation analysis that using only two controls is likely to result in tens of thousands of spurious “true positives” (Supplementary Fig. 1). Therefore, benchmarking on these “true positives” is an invalid approach. Second, Zheng *et al.*¹ report improved sensitivity for SVA; however, they do not report that a simple unadjusted analysis using a standard Bonferroni significance level achieves considerably better sensitivity and better specificity than SVA (Supplementary Table 1). Thus, the metric used to evaluate performance is also invalid, as the naive method that does not adjust for cell-type heterogeneity outperforms a method that does. A detailed description of this experiment as well as additional flaws in their analysis is given in Supplementary Note 1.

The focus of Zheng *et al.*¹ on the potential loss of power in the case of many true positives is of interest. Because a reliable gold standard is currently not available, we examined this scenario by splitting a large set of breast cancer samples ($n = 305$)⁵ into two groups on the basis of the reference-based cell-composition estimates provided by Zheng *et al.*¹. One group was labeled as controls, and differential methylation effects were added to all samples in the other group in more than 20,000 sites. The results (Supplementary Note 2 and Supplementary Tables 2–4) show that ReFACTor and SVA obtain similar sensitivity, but ReFACTor captured the cell composition substantially better than SVA and thus adjusted well for false positives, whereas SVA suffers from thousands of false positives. In contrast to the argument of Zheng *et al.*¹, when ReFACTor is correctly applied (Zheng *et al.*¹ did not follow our guidelines), the ReFACTor components are dominated by information about cell-type composition rather than disease status (Supplementary Note 1 and Supplementary Fig. 2).

Zheng *et al.*¹ next consider our original experiment in which FACS cell counts were available². They argue that successively adding components may cause overfitting. However, our point in that section of Rahmani *et al.*² was to evaluate the relative performance of different methods as a function of model dimension, and thus there is no issue of overfitting (Supplementary Note 3). They evaluated ReFACTor by measuring the correlation between each cell type and ReFACTor components selected via likelihood ratio test (LRT) and observed that ReFACTor only slightly improves upon the reference-based approach. However, LRT depends on sample size, hence we re-evaluated ReFACTor using LRT with all 560 samples in the data set (as opposed to a subset). Our analysis revealed more significant components, which leads to a substantial improvement, far outperforming the reference-based approach (Supplementary Note 3 and Supplementary Fig. 3).

Finally, Zheng *et al.*¹ try to demonstrate the advantage of the reference-based method³ using a very small data set with known cell composition ($n = 18$)⁶. However, in their analysis, Zheng *et al.*¹ did not correct for known batch effects, and we found that adjusting for batch information produces similar performance for ReFACTor and the reference-based method (Supplementary Fig. 4). Furthermore, such a small sample size cannot provide statistically

significant evidence for the improvement of any method. Specifically, using multiple subsampled FACS data sets of 18 samples, we observed that the performance of both methods was highly variable (Supplementary Fig. 5). Moreover, Zheng *et al.*¹ relied on a method for determining the dimension of the data (RMT)⁷. We found that the number of dimensions estimated by RMT is linearly determined by the sample size ($R^2 > 0.95$), making it inapplicable (Supplementary Fig. 6 and Supplementary Note 3).

Given that firm conclusions cannot be drawn based on small data sets, we further evaluated the performance of ReFACTor and the reference-based method using five large whole-blood data sets (minimum $n = 312$). We divided the samples in each data set into two groups on the basis of cell-composition distribution (Supplementary Note 3). Then, we conducted an EWAS on the assignment into groups as the phenotype. In this scenario, the assignment into groups is expected to be correlated with the true underlying cell composition, and an insufficient correction will lead to spurious associations. We found that ReFACTor consistently outperformed the reference-based method; particularly, the reference-based method resulted in more than 1,000 false positives in some cases, whereas ReFACTor resulted in a few dozen at most (Supplementary Note 3 and Table 1). This is not surprising given the inherent limitations of reference-based methods (Supplementary Note 4).

Comprehensive guidelines for using ReFACTor are provided in Supplementary Note 5, and a more comprehensive implementation of ReFACTor is available at <http://glint-epigenetics.readthedocs.io>. We also provide recommendations for the choice of algorithms for specific tasks in methylation analysis in Supplementary Table 5. We agree with Zheng *et al.*¹ that the community should constantly seek further evaluation of existing methods in more scenarios and using more data sets and tissues. Specifically, reasonable sample sizes should be used to draw significant conclusions. Our results provide multiple compelling evidence based on multiple large data sets that ReFACTor is currently the state-of-the-art approach for controlling for tissue heterogeneity in EWAS, even when compared to the reference-based method in whole blood.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University, the Israel Science Foundation (1425/13 to E.R. and E.H.) and US National Science Foundation grant 1331176 and United States Israel Binational Science Foundation grant 2012304 (to E.R., Y.B. and E.H.). E.R. was supported by Len Blavatnik and the Blavatnik Research Foundation. N.Z. was supported in part by a US National Institutes of Health (NIH) career development award from the NHLBI (K25HL121295). C.E., S.H., D.H., J.G., S.O. and E.B. were supported by the Sandler Family Foundation, the American Asthma Foundation, Hind Distinguished Professorships, and the following NIH grants: 1P60MD006902, 1R01HL117004, R21ES24844, R01HL128439, TRDRP 24RT-0025. E.E. was supported by NSF grants 1065276, 1302448, 1320589 and 1331176 and NIH grants R01-GM083198, R01-ES021801, R01-MH101782, R01-ES022282 and U54EB020403.

References

1. Zheng SC, et al. Nat Methods. 2017; 14:216–217. [PubMed: 28245219]
2. Rahmani E, et al. Nat Methods. 2016; 13:443–445. [PubMed: 27018579]

3. Houseman EA, et al. BMC Bioinformatics. 2012; 13:86. [PubMed: 22568884]
4. Leek JT, Storey JD. PLoS Genet. 2007; 3:e161.
5. Teschendorff AE, et al. Nat Commun. 2016; 7:10478. [PubMed: 26823093]
6. Koestler DC, et al. BMC Bioinformatics. 2016; 17:120. [PubMed: 26956433]
7. Teschendorff AE, Zhuang J, Widschwendter M. Bioinformatics. 2011; 27:1496–1505. [PubMed: 21471010]
8. Tan Q, et al. BMC Genomics. 2014; 15:1062. [PubMed: 25476734]
9. Liu Y, et al. Nat Biotechnol. 2013; 31:142–147. [PubMed: 23334450]
10. Hannon E, et al. Genome Biol. 2016; 17:176. [PubMed: 27572077]

Table 1
Numbers of false positives in whole-blood EWAS

	Unadjusted	Reference-based method ^a	ReFACTor ^b
GALA II FACS ($n = 84$) ²	27 (108)	5 (39)	0 (0)
GALA II ($n = 560$) ²	2,569 (6,845)	24 (1,341)	11 (158)
Tan <i>et al.</i> ($n = 312$) ⁸	1,592 (8,413)	231 (3,703)	8 (122)
Liu <i>et al.</i> ($n = 686$) ⁹	4,851 (14,541)	1,044 (5,191)	77 (2,266)
Hannon <i>et al.</i> A ($n = 675$) ¹⁰	1,711 (6,840)	24 (490)	33 (176)
Hannon <i>et al.</i> B ($n = 847$) ¹⁰	3,013 (9,672)	40 (749)	31 (142)

Numbers of false positives under Bonferroni correction and (in parentheses) FDR correction, averaged across multiple executions (Supplementary Note 3).

^aSeven cell types analyzed.

^bSeven components analyzed.