Correcting for cell-type heterogeneity in Epigenome-Wide Association Studies: revisiting previous analyses.

Shijie C Zheng [1,2], Stephan Beck [3], Andrew E. Jaffe [4,5], Devin C. Koestler [6], Kasper D. Hansen [7,8,9], Andres E. Houseman [10], Rafael A. Irizarry [11,12] and Andrew E. Teschendorff [1,13,14,*]

1. CAS Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai, China.
2. University of Chinese Academy of Sciences, 19A Yuquan Road, Beijing, China.
3. Medical Genomics, Paul O'Gorman Building, UCL Cancer Institute, University College London, 72 Huntley Street, London, United Kingdom.
4. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America.
5. Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland, United States of America.
6. Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS, USA.
7. Center for Epigenetics, Johns Hopkins University School of Medicine, USA.
8. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, USA.
9. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA.
10. School of Biological and Population Health Sciences, College of Public Health and Human Sciences, Oregon State University, Corvallis, OR, USA.
11. Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, USA.
12. Department of Biostatistics, Harvard T.H. Chan School of Public Health, USA.
13. Statistical Cancer Genomics, Paul O'Gorman Building, UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, United Kingdom.
14. Department of Women's Cancer, University College London, 74 Huntley Street, London, United Kingdom.

Corresponding author: Andrew E. Teschendorff: a.teschendorff@ucl.ac.uk

To the Editor:

Recently, there has been growing interest in statistical algorithms designed for tackling intra-sample cellular heterogeneity (ISCH) in Epigenome-Wide Association Studies (EWAS) [1]. Such algorithms can be broadly classified as either reference-based (if they use reference DNA methylation (DNAm) profiles of representative cell types) [2], or reference-free (if they don't require such reference profiles) [3-6]. Reference-free methods can be further subdivided into those that use the phenotype of interest in the inference process (this includes algorithms such as Surrogate Variable Analysis (SVA) [4, 7] and RefFreeEWAS [3]), and those that do not (e.g. EWASher [5] and RUV [6]). Comparisons between these different inference-paradigms is of paramount interest in order to inform the EWAS community on how best to approach the ISCH problem.

A recent study by Rahmani et al [8] presented a reference-free algorithm called ReFACTor, and suggested that it leads to improved estimates of cell type composition and power when compared to other competing algorithms. However, the approach on which ReFACTor is based could incorrectly remove the biological signal of interest if the latter is stronger than the variation associated with cell-type composition. We confirmed this by applying ReFACTor to additional datasets. Below we discuss key issues which any future methodological comparative study should pay particular attention to, to ensure robust and meaningful conclusions, which can then be used to guide the EWAS community.

In principle, an advantage of a reference-free method like ReFACTor is that it is applicable to any tissue type. It is important therefore to assess performance in tissue types other than blood, because assumptions valid in one tissue type may not be valid in others. For instance, ReFACTor relies on the assumption that the top components of variation are associated with changes in cell-type composition, effectively using these components to construct variables that account for variations in cell-type. While this assumption may be valid for EWAS conducted in whole blood [9], the generality of it to other tissue types remains to be shown. In essence, ReFACTor is similar in concept to Remove Unwanted Variation (RUV) [6] in that both select control genes that capture confounding variation. However, blind application of ReFACTor could lead to a substantial loss of power if control genes are misidentified as those carrying biological signal. Although these problems represent an intrinsic limitation of any reference-free method, it will be particularly acute for methods like ReFACTor or EWASher [5], which do not use phenotype information from the outset. We used normal mammary epithelial and breast cancer cell-line data to define a gold-standard set of true positive features and a breast cancer tissue EWAS for the evaluation of several methods. SVA [4] had a much better control of power,outperforming ReFACTor by as much as 70% (**Table-1, Supplementary Data 1-2, Supplementary Software 1-2**). While specificity is harder to estimate, the improved power of SVA over ReFACTor was at the expense of only a 10-20% lower specificity (**Table 1**). ReFACTor's loss of power in our cancer-tissue EWAS was due to the top components of variation correlating more strongly with disease status than with cell-type composition (**Supplementary figure 1**). Only lower-ranked components correlated with adipose cell content, which is the major source of cell-type variation in breast tissue (**Supplementary figure 1**). This problem could in principle be circumvented by applying ReFACTor to the normal samples only, as suggested by Rahmani et al., but it remains to be tested on more datasets. Hence, application of a method like ReFACTor demands that one must carefully consider the tissue and biological context.

A second key issue concerns the evaluation of a reference-free method in terms of modelling cell-type composition. In the case of ReFACTor, estimated components were added successively to a linear model, leading to an improvement in the fraction of variance explained (summarized with R^2 values). To avoid the problem of overfitting we used a nested models likelihood ratio test (LRT) (or adjusted R^2 values). We found little justification for the successive addition of components (**Supplementary Methods, Supplementary Software 1-4, Supplementary Data 3, Supplementary figure 2**). Alternatively, one could attempt to estimate the number of significant components of variation. In our hands entering such estimates into ReFACTor leads to a drop of as much as 20% in R^2 values, resulting in reduced modeling performance, when compared to reference-based methods (**Supplementary figures 3-4**). This indicates that application of ReFACTor with all estimated components could lead to overfitting. We confirmed this further using training/test set partitions (**Supplementary figure 5**).

Another issue is the use of a single or limited number of datasets with matched FACS data to benchmark a novel method against existing algorithms. In our experience, the complexity and unknown nature of the

sources of variation in EWAS data requires many datasets to reach unbiased conclusions. To demonstrate this, we performed cell composition analysis for an independent whole blood dataset, as well as an extensive analysis encompassing five different in-silico mixture experiments, drawing on 1573 purified blood cell-types from over 6 different studies (**Supplementary table 1**). These analyses demonstrate the strength of Houseman's reference-based method compared to ReFACTor (**Supplementary figure 2, Supplementary figures 6-9**). Further issues, including inappropriate choice of gold-standards in real data are discussed in Supplementary Methods.

In summary, we suggest that future studies proposing novel methods ought to (i) provide comprehensive comparisons to existing algorithms, (ii) use biological scenarios and datasets that allow objective comparisons, and (iii) when applicable, include tissues other than blood. We provide some recommendations in the accompanying **Supplementary Information and Supplementary table 2.** Briefly, we recommend reference-based methods for scenarios where the composition of tissues is relatively well known, and reference-free methods like SVA or RefFreeEWAS when reference DNAm profiles are not available. We point out that our recommendations are based on currently available data sets and approaches, which may change as the field continues to evolve.

Data Availability: All data analyzed is publicly available. See Supplementary Information for detailed accession numbers for all datasets analysed.

Acknowledgements: AET was supported by a Royal Society Newton Advanced Fellowship (award number 164914) and an NSFC grant (31571359) and by the Chinese Academy of Sciences. SB was supported by the EU-FP7 BLUEPRINT Project (282510). DCK was supported by the National Institute of Health (NIH) grants: (1KL2TR000119) and Kansas IDeA Network of Biomedical Research Excellence (K-INBRE) Bioinformatics Core, supported in part by the National Institute of General Medical Science award P20GM103418.

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

References:

1. Jaffe, A.E. & Irizarry, R.A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology* **15**, R31 (2014).
2. Houseman, E.A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**, 86 (2012).
3. Houseman, E.A., Molitor, J. & Marsit, C.J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431-1439 (2014).
4. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**, 1724-1735 (2007).
5. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nature methods* **11**, 309-311 (2014).
6. Gagnon-Bartsch, J.A. & Speed, T.P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539-552 (2012).

7. Teschendorff, A.E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**, 1496-1505 (2011).
8. Rahmani, E. et al. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature methods* **13**, 443-445 (2016).
9. Liu, Y. et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* **31**, 142-147 (2013).

Tables

	Unadj.	**SVA**	**ReFACTor (k=6,ncp=6)**	**ReFACTor (k=6,ncp=15)**	**ReFACTor (k=10,ncp=10)**	**ReFACTor (k=10,ncp=15)**
SE (P < 0.05)	0.90 (n=20876)	0.83 (n=19356)	0.09 (n=2066)	0.02 (n=412)	0.02 (n=410)	0.02 (n=410)
SE (FDR < 0.05)	0.89 (n=20667)	0.81 (n=18743)	0.04 (n=835)	≈0 (n=23)	≈0 (n=13)	≈0 (n=13)
SP (P < 0.05)	0.53 (n=16057)	0.70 (n=10274)	0.62 (n=12793)	0.92 (n=2603)	0.95 (n=1582)	0.95 (n=1582)
SP (FDR < 0.05)	0.58 (n=14146)	0.75 (n=8436)	0.84 (n=5571)	0.99 (n=115)	≈1 (n=11)	≈1 (n=11)

Table-1: *Table comparing the relative sensitivity (SE) and specificity (SP) of ReFACTor (for 4 different choices of k and ncp parameters: ncp=15, estimated using RMT [7] as described in* ***Supplemental Methods****), to SVA and to an unadjusted analysis. Sensitivities and Specificites were estimated using a set of n=23258 true positives and 34078 true negatives, respectively, and are shown at an unadjusted P < 0.05 and FDR corrected < 0.05.*