Kaushal *et al. BMC Bioinformatics* 2015, **16**(Suppl 15):P7
http://www.biomedcentral.com/1471-2105/16/S15/P7

BMC Bioinformatics

POSTER PRESENTATION Open Access

Which methods to choose to correct cell types in genome-scale blood-derived DNA methylation data?

Akhilesh Kaushal[1], Hongmei Zhang[1*], Wilfried JJ Karmaus[1], Julie SL Wang[2]

From 14th Annual UT-KBRIN Bioinformatics Summit 2015
Buchanan, TN, USA. 20-22 March 2015

Background

High throughput methods such as microarray and DNA-methylation are used to measure the transcriptional variation due to exposures, treatments, phenotypes or clinical outcomes in whole blood, which could be confounded by the cellular heterogeneity [1,2]. Several algorithms have been developed to measure this cellular heterogeneity. However, it is unknown whether these approaches are consistent, and if not, which method(s) perform better.

Materials and methods

The data implemented in this study were from a Taiwan Maternal and Infant Cohort Study [3,4]. We compared five cell-type correction methods, including four methods recently proposed: the method implemented in the minfi R package [5], the method by Houseman et al. [6], FaST-LMM-EWASher [7], RefFreeEWAS [8]) and one method using surrogate variables [9] (SVAs). The association of DNA methylation at each CpG site across the whole genome with maternal arsenic exposure levels was assessed adjusting for the estimated cell-types. To further demonstrate and evaluate the methods that do not require reference cell types, we first simulated DNA methylation data at 150 CpG sites across 600 samples based on an association of DNA methylation with a variable of interest (e.g., level of arsenic exposure) and a set of latent variables representing "cell types". We then simulated DNA methylation at additional CpG sites only showing association with the latent variables.

Results

Only 3 CpG sites showed significant associations with maternal arsenic exposure at a false discovery rate (FDR) level of 0.05, without adjusting for cell types. Adjustment by FaST-LMM-EWASher did not identify any CpG sites. For other methods, Figure 1 illustrates the overlap of identified CpG sites. Further simulation studies on methods free of reference data (i.e., FaST-LMM-EWASher, RefFreeEWAS, and SVA) revealed that RefFreeEWAS and SVA provided good and comparable sensitivities and specificities, and FaST-LMM-EWASher gave the lowest sensitivity but highest specificity (Table 1).

Conclusions

The results from real data indicated RefFreeEWAS and SVA were able to identify a large number of CpG sites, and results from SVA showed the highest agreement with all other approaches. Simulation studies further

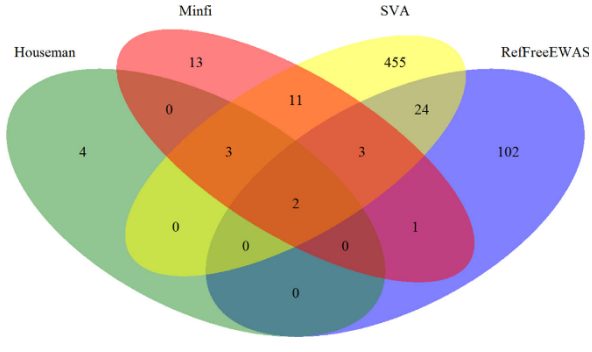

Figure 1 Venn diagram illustrating the overlap of significant CpG sites at FDR level of 0.05 after adjusting for cell types by different methods for the association study of maternal arsenic exposure with DNA-methylation.

* Correspondence: hzhang6@memphis.edu
[1]Division of Epidemiology, Biostatistics, and Environmental Health, University of Memphis, Memphis, TN 38152, USA.
Full list of author information is available at the end of the article

Table 1 Sensitivity and specificity with respect to truly identified variables using 100 simulated data; CI: confidence interval

	Sensitivity: Median (95% CI)	Specificity: Median (95% CI)
FaST-LMM-EWASher	0.00 (0.00, 0.52)	1.00 (0.99, 1.00)
RefFreeEWAS	0.98 (0.00, 1.00)	0.94 (0.93, 1.00)
SVA	1.00 (0.98, 1.00)	0.94 (0.93, 0.94)

confirmed that RefFreeEWAS and SVA are comparable and perform better than FaST-LMM-EWASher. Overall, the findings support a recommendation of using SVA to adjust for cell types due to its highest agreement with other methods and appealing findings from simulation studies.

Authors' details
[1]Division of Epidemiology, Biostatistics, and Environmental Health, University of Memphis, Memphis, TN 38152, USA.. [2]Division of Environmental Health & Occupational Medicine, National Health Research Institutes, Miaoli, 360, Taiwan.

Published: 23 October 2015

References

1. Adalsteinsson BT, Gudnason H, Aspelund T, Harris TB, Launer LJ, Eiriksdottir G, Smith AV, Gudnason V: **Heterogeneity in white blood cells has potential to confound DNA methylation measurements.** *PloS one* 2012, **7(10)**:e46705.
2. Talens RP, Boomsma DI, Tobi EW, Kremer D, Jukema JW, Willemsen G, Putter H, Slagboom PE, Heijmans BT: **Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology.** *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 2010, **24(9)**:3135-3144.
3. Lin L-C, Wang S-L, Chang Y-C, Huang P-C, Cheng J-T, Su P-H, Liao P-C: **Associations between maternal phthalate exposure and cord sex hormones in human infants.** *Chemosphere* 2011, **83(8)**:1192-1199.
4. Wang S-L, Su P-H, Jong S-B, Guo YL, Chou W-L, Päpke O: **In utero exposure to dioxins and polychlorinated biphenyls and its relations to thyroid function and growth hormone in newborns.** *Environmental health perspectives* 2005, 1645-1650.
5. Jaffe AE, Irizarry RA: **Accounting for cellular heterogeneity is critical in epigenome-wide association studies.** *Genome biology* 2014, **15(2)**:R31.
6. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT: **DNA methylation arrays as surrogate measures of cell mixture distribution.** *BMC bioinformatics* 2012, **13**:86.
7. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J: **Epigenome-wide association studies without the need for cell-type composition.** *Nature methods* 2014, **11(3)**:309-311.
8. Houseman EA, Molitor J, Marsit CJ: **Reference-free cell mixture adjustments in analysis of DNA methylation data.** *Bioinformatics* 2014, **30(10)**:1431-1439.
9. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** *PLoS genetics* 2007, **3(9)**:e161.

doi:10.1186/1471-2105-16-S15-P7
Cite this article as: Kaushal *et al.*: Which methods to choose to correct cell types in genome-scale blood-derived DNA methylation data? *BMC Bioinformatics* 2015 **16**(Suppl 15):P7.