HHS Public Access

Author manuscript

Epigenomics. Author manuscript; available in PMC 2017 July 10.

Published in final edited form as:

Epigenomics. 2015 ; 7(1): 13–15. doi:10.2217/epi.14.70.

Candidate gene methylation studies are at high risk of erroneous conclusions

Andrey A Shabalin and

Center for Biomarker Research & Personalized Medicine, Virginia Common-wealth University, PO Box 980533, Richmond, VA 23298, USA

Karolina A Aberg

Center for Biomarker Research & Personalized Medicine, Virginia Common-wealth University, PO Box 980533, Richmond, VA 23298, USA

Keywords

candidate gene study; DNA methylation; false discoveries; methylome-wide association study; principal component analysis

DNA methylation studies present a promising avenue for improving our understanding of common diseases and alleviating part of their public health burden. A commonly used approach involves testing sites in genes of interest for association with disease status. These genes are typically selected based on *a priori* ideas about their possible role in pathogenic processes. Compared with assaying many sites simultaneously, such candidate gene methylation studies are appealing because of their low costs. They also have the advantage of being relatively straightforward in terms of lab technical and statistical procedures. However, in this commentary we argue that specific properties of methylation studies present a serious challenge for the interpretation of findings originating from the candidate gene approach.

Common variation among large subsets of methylation sites

Recently a number of investigations assayed large sets of methylation sites simultaneously. A striking finding emerging from these studies is that the methylation statuses of large subsets of sites covary with each other [1,2]. This common variation is not restricted to specific chromosomal locations but involve methylation sites across the entire genome. Principal component analysis (PCA) provides a good approach to quantify this phenomenon.

Author for correspondence: Edwin JCG van den Oord, Center for Biomarker Research & Personalized Medicine, Virginia Common-wealth University, PO Box 980533, Richmond, VA 23298, USA, Tel.: +1 804 828 6098, Fax: +1 804 628 3991, ejvandenoord@vcu.edu.

Financial & competing interests disclosure

This work was supported by grants 1R03MH102723, 1R01MH099110 and 1R01MH104576. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript.

No writing assistance was utilized in the production of this manuscript.

PCA captures the common variation in methylation statuses among sites in a form of a set of uncorrelated components. The first principal component (PC) accounts for as much of the variation in the methylation data as possible, the second component captures as much of the remaining variance as possible in such a way that it is uncorrelated with the first component and so forth. Bell *et al.* performed PCA on methylation levels at 22,290 CpGs in lymphoblastoid cell lines from 77 individuals [1]. They found that the first PC explained 22% of the variation in methylation, and the first three PCs together explained 33%. This large amount of variation explained does not seem to be an artifact of the specific approach. Whereas Bell *et al.* [1] used an array, Aberg *et al.* [2] used a sequencing based approach to assay all 28 million common CpGs in the human genome in a sample of 1497 subjects. Their first PC explained 27% of the variation in methylation, and the first three PCs together explained 36%. These findings seem consistent with observations that global methylation levels may vary among subjects as a function of, for example, demographic variables, life style, nutrition or disease status [3,4]. Such global variations are only possible when individuals differ at many sites in a similar fashion.

Impact on association testing

Association testing typically starts with calculating a test statistic for each of the investigated methylation sites. If the test statistic is greater than a critical value, the null-hypothesis, assuming that the site has no effect, is rejected. The error of rejecting the null-hypothesis when it is true results in a false positive. Not rejecting the null-hypothesis when it is false results in a false negative.

To gain more insight into the impact of the common variation on association testing, we first performed a simulation study. For simplicity, we used a single PC. Mimicking the empirical data, this PC explained 20% of the variation. We simulated datasets of 500 samples each. To assess the impact on association testing, we calculated inflation factor λ; the ratio of the median test statistic value across all sites to the median test statistic value expected under the null hypothesis. Thus, if for all tests the null hypothesis is true, λ would need to be 1 for accurate statistical inferences.

In a first scenario, the PC was uncorrelated with the disease. We observed $\lambda = 0.873$. This λ indicates that test statistics were substantially lower (or p-values higher) than expected under the null hypothesis. This implies an increased risk of false negatives. In the second scenario we allowed a correlation of 0.1 between the PC and the disease. We now observed $\lambda = 1.712$. This λ suggests that test statistics were substantially higher (or p-values lower) than expected under the null hypothesis, implying an excess of false positives.

The simulation study clearly demonstrates the basic phenomena in a simplified scenario. In real life data there will likely be more than one strong PC, where some may be correlated with the disease and other not. The overall impact on the test statistics will be a sum of the effects of the individual PCs. In studies where PCs can be computed, an effective approach to correct for the test statistic inflation or deflation is to include the PCs as co-variates when performing the association tests. For example, inclusion of the PC as a co-variate reverted λ back to 1 in both scenarios from our simulation study. The only cautionary note about this

correction procedure involves the PCs that are correlated with the disease. Although these PCs may mainly capture confounding factors such as life style differences between cases and controls, this may not always be the case. For example, in addition to changing the methylation status of sites (e.g., side effects) that do not play any role in the disease, medications could also correct the aberrant methylation in sites that cause the disease. This approach of regressing out PCs is commonly used to avoid test statistic inflation caused by population stratification in genome-wide association studies with sequence variants. However, a major difference is that in methylation studies the PCs explain a much larger proportion of the variation. This has serious implications for candidate gene methylation studies.

Implications for candidate gene methylation studies

To give an impression of the implications for candidate gene methylation studies, we use findings from two real studies. Study 1 [Van den oord EJCG et al., Unpublished Data] involves a methylome-wide association study using samples from 75 patients and controls. In this study, five PCs explained most of the variation, where none of these PCs were associated with disease status. Consistent with the results from our simulation study, we observed $\lambda = 0.89$ when no PCs were included as co-variates in the association analyses. Study 2 involved the previously mentioned study by Aberg *at al.* conducting methylome-wide association study in 1497 cases and controls [2]. The first seven PCs explained most of the variation. Three of these PCs were associated with case–control status. Again consistent with our simulations, we observed $\lambda = 7.4$ when no PCs were included as co-variates in the association analyses.

In candidate gene studies, it is fairly typical to use a threshold p-value of 0.05 for declaring significance. Assuming accurate test statistics (i.e., $\lambda = 1$), this implies a probability of 5% that the study will produce a false positive. If multiple sites are tested, researchers will often use a Bonferroni correction that divides this threshold p-value by the number of tests performed. Regardless of the number of tests that are performed, such as correction would still ensure a probability of less than 5% that the study produces one or more false positives.

If a candidate study would have been performed in the same study sample as was used in Study 1, where none of the PCs were associated with the disease and test statistics were deflated, we would observe fewer false positives than expected based on the used p-value threshold. For example, allowing a probability of 5% that a test will produce a false positive (p-value threshold of 0.05), we would observe 3.4% of the sites being significant. Allowing this probability to be 1% (p-value threshold of 0.01, e.g., Bonferroni correction if five sites were measured), we would observe that 0.5% of all sites are significant. If we used the sample from Study 2 instead, where three PCs were associated with the disease and test statistics were inflated, we would observe many more significant tests than we would expect based on the p-value threshold that is used. Rather than 5% (p-value threshold of 0.05), we would observe that 46.7% of all sites are significant. Even at a more stringent 1% level (p-value threshold of 0.01), 31.3% of the sites would remain significant. Although a candidate gene study in this sample is very likely to produce a significant result, it would be

questionable whether that finding would provide any insight into the disease process or be caused by confounding effects.

These results suggest that depending on the correlations of PCs in the methylation data with diseases status, which would be unknown to the researcher in the case of a candidate gene study, we would see a higher than expect rate of false negatives or a flood of false positives.

Conclusion

In DNA methylation studies, large amounts of variation may be common among sites located across the entire genome. In candidate gene methylation studies where only few sites are assayed it is impossible to account for this variation through PCs or related statistical procedures. The implications depend on whether or not this common variation is associated with the disease. In the uncorrelated case, test statistics will be deflated resulting in too few sites with small p-values. In the correlated case, test statistics will be inflated resulting in too many sites with small p-values. Although candidate gene studies of sequence variants would in principle be affected by the same phenomena, a major difference is that in methylation studies the effects are much more severe because the correlations involve much larger subsets of sites. For example, we show on actual data that 30–50% of the tested methylation sites could be significant in candidate gene studies that use p-value thresholds of 0.01–0.05. Thus, results from candidate gene methylation studies are very difficult to interpret properly.

This lack of transparency is further enhanced by the fact that the choice of the p-value threshold in candidate gene studies often lacks proper statistical motivation. For example, because for the majority of tested sites the null hypothesis will be true, commonly used p-value threshold of 0.05 or 0.01 will often result in many more false positives than true positives [5,6]. When many tests are performed, methods can be used to empirically find the p-value threshold that provides a desired balance between false and true positives [7]. However, because these methods cannot be applied in candidate gene studies, it remains unclear what proportion of the significant findings produced by these studies are true or false.

Due to the unclear interpretation of results, we argue that candidate gene methylation studies are at high risk of erroneous conclusions. Rather than using candidate gene methylation studies as a discovery tool to detect initial associations, they are possibly better used to follow up significant findings from studies that can provide the insight into and properly handle the variance which is common among large subsets of methylation sites. For example, they may provide technical validation using a different technology or, provided that independent samples from the same study population are used, perform replication studies that can be informed by the previously generated knowledge of the common variance structure.

References

1. Bell JT, Pai AA, Pickrell JK, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011; 12(1):R10. [PubMed: 21251332]

2. Aberg KA, McClay JL, Nerella S, et al. Methylome-wide association study of schizophrenia: identifying blood biomarker signatures of environmental insults. JAMA Psychiatry. 2014; 71(3): 255–264. [PubMed: 24402055]
3. Hsiung DT, Marsit CJ, Houseman EA, et al. Global DNA methylation level in whole blood as a biomarker in head and neck squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev. 2007; 16(1):108–114. [PubMed: 17220338]
4. Lim U, Song MA. Dietary and lifestyle factors of DNA methylation. Methods Mol Biol. 2012; 863:359–376. [PubMed: 22359306]
5. van den Oord EJ, Sullivan PF. False discoveries and models for gene discovery. Trends Genet. 2003; 19(10):537–542. [PubMed: 14550627]
6. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst. 2004; 96(6):434–442. [PubMed: 15026468]
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B. 1995; 57:289–300.