Ulrich Mahlknecht, MD, PhD

Editor-in-Chief

Clinical Epigenetics

Nov 17, 2014

Dear Dr. Mahlknecht:

Thank you for your letter of Nov 13, 2013 regarding our manuscript entitled “Identification and validation of the methylation biomarkers of Non-small cell lung cancer (NSCLC)”. Our appreciation also goes to the reviewers for their helpful comments. We have revised the manuscript following the reviewer’s comments and your instructions.

Enclosed please find the revised version of the manuscript along with a point by point description of our responses to the reviewer’s comments. We hope that the manuscript is now acceptable for publication in Clinical Epigenetics. Thank you again for your letter and for your editorial assistance.

Sincerely yours,

Jiucun Wang, Ph.D.

Professor,

School of Life Sciences

Fudan University

220 Handan Road, Shanghai 200433

People’s Republic of China.

Tel: (+) 86-21-55665499

Fax: (+) 86-21-55664885

E-mail: [jcwang@fudan.edu.cn](mailto:jcwang@fudan.edu.cn)

**Responses to Reviewer 1’s comments**

We first thank the reviewer for the helpful comments. In the revised manuscript, we have incorporated the reviewer’s comments.

**Comments**: “1. In the discovery stage, could you please explain the definition of “batch” in your results session? “Batch effect was significantly existed among the datasets which was showed in the first and second principle components.” The “batch” here indicates “individual study” or the real “batch” within each study? Similarly, what kind of “batch” effect “Combat” removed?”

**Respon**se: Thank you so much. It is really a good question. Microarray measures some specific signals, such as gene expression, DNA methylation and so on, of thousands of genes in a single assay. It is a revolutionary tool for identifying genes whose methylation changes in response to specific situation, such as different development stages, physiological or pathological status. However, microarray results can be affected by minuscule differences in any number of non-biological variables, such as methods for RNA isolation, amplification and target labeling, and array processing and scanning, reagents from different companies, different technicians or even changing atmospheric ozone levels can impact. Here, the term “batch” refers to microarrays processed at one site over a short period of time using the same platform. The cumulative error introduced by these time and place-dependent experimental variations is referred to as batch effects. In terms of different study, the methylation array data was created in different time, place, technician and so on, therefore, the main variation of among data would be showed as batch effect.

In present study, batch is equivalent to “each individual study” and Combat algorithm could remove such noise (batch signal or each individual study) from the dataset, adjusted with additional and multiple effect of the batch information. We have made it clear in the background section.

2. Did the authors use pooled-data from the three public datasets, or meta-analyze the individual results from separate datasets? Or both? Please clarify.

**Response:** Thank you so much. It is really a good question. In the present study, pooled-data from the three public datasets followed by batch effect elimination with Combat algorithm was used to the following statistical analysis. We have made it more clear in the background section.

3. Multi-cellular issue is important in the methylation study. Please discuss how you handle this issue.

Response: Thank you so much. It is really a good question. Multi-cellular issue is a great challenge in epigenetic related study. Usually, cancer tissues from surgery operation would include some normal tissue to guarantee the complete remove of the tumor. However, in the present study, the null hypothesis is that the methylation level in the cancer tissue (mixed) is same with normal tissue (mixed), therefore, the alternative hypothesis is that the methylation level in the cancer tissue (mixed) is different with normal tissue (mixed), therefore, we can use paired *t*-test to test the mean of the methylation between cancer tissue and normal tissue. In the present study, we do not have the multi-cellular problem.

**Discretionary Revisions**

Though it is mentioned that “no significant association was observed between any of the 5 genes with age, smoking, TNM stage, lung cancer differentiation and lung cancer subtype” in the current study. I would further adjust for “TNM stage, lung cancer differentiation and lung cancer subtype” in the association between methylation and NSCLC status at least as a sensitivity analysis.

**Response:** Thank you so much. It is really a good question. We conducted the analysis again according your suggestion. The result is same with our previous result: No significant association was observed between any of the 5 genes with age, smoking, TNM stage, lung cancer differentiation and lung cancer subtype both in univariate and multivariate association models.

**Responses to Reviewer 2’s comments**

We first thank the reviewer for the helpful comments. In the revised manuscript, we have incorporated the reviewer’s comments.

**Comments:** The p-values in Table 2 do not agree with the log10(OR) and 95% CI. Take SLC5A8 as an example, the 95%CI is 2.51-5.09. The standard error (SE) calculated from the CI is 0.658. The z score can be calculated from the log10(OR) and SE, which is 5.775, and the p-value corresponding to the z score is 9E-7, roughly 3 orders of magnitudes less significant than the reported p-value (4.8E-12).

**Response:** Thank you so much for you careful checking. the beta and P-value were come from the logistic regression directly from R code.

glm<- glm(y~gender+age+x,data,family=binomial(logit))) (4)

beta and se were estimated with above R code/result. And then OR and the 95% CI of the OR were calculated with the following function.

OR <- exp(beta) (1)

Up <- exp(beta+1.96\*se) (2)

Low <- exp(beta-1.96\*se) (3)

Take SLC5A8 as an example, the 95%CI is 2.51-5.09. The beta is ln(OR). The standard error (SE) calculated from the 95% CI is 1.1515 rather than 0.658. The z score can be calculated from the log10(OR) and SE, which is 6.9099, rather than 5.775, and the p-value corresponding to the z score is 4.8E-12 (2\*(1-pnorm(6.9099))), rather than 9E-7(2\*(1-pnorm(5.775))). The main problem may be caused by the difference between ln and log(10).

**Comments: Minor Essential Revisions**

Several important details are missing in this manuscript. For example

a) how was the 'FDR adjusted p-value" calculated.

b) how does the 'combat' method work (a citation or some descriptions are necessary)

c) sensitivity, specificity and accuracy usually depend on a tunable threshold. The authors should discuss how the threshold was chosen.

d) it's unclear which dataset (discovery? validation) was used in the 'Sensitivity, specificity and accuracy of the diagnosis panel' section

**Response:**

Thank you so much, we have added these information into the manuscript as the following:

1. False discovery rate (FDR) correction was used for multiple test correction with the R function of p.adjust with fdr as parameter. In the article, it is showed as “False discovery rate (FDR) correction was used for multiple test correction with the R function of p.adjust with fdr as parameter”.
2. Microarray measures some specific signals, such as gene expression, DNA methylation and so on, of thousands of genes in a single assay. It is a revolutionary tool for identifying genes whose methylation changes in response to specific situation, such as different development stages, physiological or pathological status. However, microarray results can be affected by variety of non-biological variables, such as methods for RNA isolation, amplification and probe labeling, processing and scanning, reagents from different companies, different technicians or even changing atmospheric ozone levels. Usually, the term “batch” refers to microarrays processed at one site over a short period of time using the same platform. The cumulative error introduced by these time and place-dependent experimental variations is referred to as batch effects. In terms of different study, the methylation array data was created in different time, place, technician and so on, therefore, the main variation of among data would be showed as batch effect. In present study, batch is equivalent to “each individual study” and Combat algorithm could remove such noise (batch signal or each individual study) from the dataset, adjusted with additional and multiple effect of the batch information. We have provide the reference for Combat algorithm in the 14th and 15th reference: Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C: Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. Plos One 2011, 6:e17238 and Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007, 8:118-127.
3. It is really a good suggestion. You are quite right. The sensitivity, specificity and accuracy depend on a tunable threshold at one stage model design. However, in the two stage design which includes training and test stage, a cut-off is choose in the training stage to obtain the best prediction model, and then the cut-off was applied in the test stage. In the present study, the optimized prediction models were built with the best prediction accuracy in training dataset and then the sensitivity, specificity, accuracy were obtained from logistic regression, SVM, random forest and Bayes tree model in the test dataset. In the manuscript, it is showed as “The optimized prediction model were built with the best prediction accuracy in training dataset and then the sensitivity, specificity, accuracy were obtained from logistic regression, SVM, random forest and Bayes tree model in the test dataset with previous parameters applied in the training stage.”
4. Thank you so much. It is really a good suggestion. We have made it clear with the following change in the manuscript, “Several classification methods including logistic regression model, random forest, support vector machine (SVM), and Bayes tree were used to construct effective diagnosis models for cancer prediction based on MSD-SNuPET result.”