Ulrich Mahlknecht, MD, PhD

Editor-in-Chief

Clinical Epigenetics

November 22, 2014

Dear Dr. Mahlknecht,

Thank you for your letter on Nov 13, 2014 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.

We have made changes on the Title page according to your instruction, including the title of the article, list the full names, institutional addresses and email addresses for all authors, indicating the corresponding author. In addition, we have used coloured text for all changes made in the revised manuscript.

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 information and editorial assistance.

Sincerely yours,

Jiucun Wang, Ph.D.

Professor

School of Life Sciences

Fudan University Jiangwan Campus

2005 Songhu Road, Shanghai 200438

People’s Republic of China.

Tel: (+) 86-21-51630606

Fax: (+) 86-21-51630607

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.

**Comment** 1. In the discovery stage, could you please explain the definition of “batch” in your results session? “Batch effect 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?”

**Response:** 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 a 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 the time and place-dependent experimental variations that are not related to biological signals is referred to as batch effects. In different studies, the methylation array data were created at different times, places and by different technicians. Therefore, the variation of methylation data may contain batch effects.

In the present study, batch is equivalent to “each individual study” and the Combat algorithm could remove such noises (batch signals arising from each individual study) from the dataset, adjusted with additional and multiple effects of the batch information. We have explained it in the section of Background.

**Comment** 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:** In the present study, pooled-data from the three public datasets followed by batch effect elimination with the Combat algorithm were used for statistical analysis. Detailed descriptions were included in the section of Background.

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

**Response:** Multi-cellular issue is a great challenge in epigenetic studies. On one side, cancer tissues include cancer cells (epithelial cells), mesenchymal cells and so. However, the proportion (at least 70% in general) of the tumor cells in cancer tissue is always much significantly higher than that of other cells. On the other side, normal tissues also include epithelial cells, mesenchymal cells and some others. In the present study, the null hypothesis is that the methylation level in the cancer tissue (mixed) is the same with normal tissue (mixed). The alternative hypothesis is that the methylation level in the cancer tissue (mixed) is different from normal tissue (mixed). Therefore, we can use a paired *t*-test to test the difference in the mean of the methylation between cancer tissue and normal tissue.

**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: We conducted the analysis again according to your suggestion. The results were similar to 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. We have incorporated the results in the Section of Methylation status validation with MSD-SNuPET

**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.

**Comment 1:** 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:** The beta and P-value were calculated by the logistic regression in the R code.

glm<- glm(y~gender+age+x,data,family=binomial(logit))) (1)

beta and the standard error (se) were estimated by the above R code. Be careful, the value in the column 4 is based on log10 rather than logE. We can calculate OR and the 95% CI of the OR by the following functions.

OR <- exp(beta) (2)

Up <- exp(beta+1.96\*se) (3)

Low <- exp(beta-1.96\*se) (4)

Taking SLC5A8 as an example, the 95%CI was 2.51-5.09. Then the standard error (SE) would be 1.1515 according to Functions (3) and (4). The z score could be calculated from the log10(OR) and SE, which was 6.9099, and the p-value corresponding to the z score was 4.8E-12 (i.e. 2\*(1-pnorm(6.9099))).

**Comment 2: 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:**

We have added the information into the revised manuscript as follows:

1. False discovery rate (FDR) correction was used for multiple test correction with the R function of p.adjust with fdr as a parameter. We have included the above description in the section of Statistical analysis and machine learning with coloured text.
2. ComBat algorithm estimated the additional and multiple effects of the batch information with empirical Bayes frameworks. The detail statistic description of Combat algorithm can be found in 14th and 15th references: 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. The sensitivity, specificity and accuracy depend on a tunable threshold at one stage in the model design. However, in the two-stage design which includes training and test stage, a cut-off was chosen 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 the training dataset and then the sensitivity, specificity, accuracy were obtained from logistic regression, SVM, random forest and Bayes tree model in the test dataset. We have incorporated the results in the Section of Statistical analysis and machine learning.
4. Validation data were used for the estimation of sensitivity, specificity and accuracy of the diagnosis panel. 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 the MSD-SNuPET result.”