Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 9 **Hypothesis**

Integration of pre-normalized microarray data using quantile correction

Takashi Yoneya[1]*, Tatsuya Miyazawa[2]

[1]Drug Discovery Research Laboratories, Kyowa Hakko Kirin Co Ltd, 1188, Shimotogari, Nagaizumi-cho, Sunto-gun, Shizuoka, 411-8731, Japan; [2]Innovative Drug Research Laboratories, Kyowa Hakko Kirin Co Ltd, 3-6-6 Asahi-machi, Machida-shi, Tokyo 194-8533, Japan; Takashi Yoneya- Email: takashi.yoneya@kyowa-kirin.co.jp; *Corresponding author

Received October 30, 2010; Accepted January 19, 2011; Published February 07, 2011

Abstract:
An enormous amount of microarray data has been collected and accumulated in public repositories. Although some of the depositions include raw and processed data, significant parts of them include processed data only. If we need to combine multiple datasets for specific purposes, the data should be adjusted prior to use to remove bias between the datasets. We focused on a GeneChip platform and a pre-processing method, RMA, and examined simple quantile correction as the post-processing method for integration. Integration of the data pre-processed by RMA was evaluated using artificial spike-in datasets and real microarray datasets of atopic dermatitis and lung cancer. Studies using the spike-in datasets show that the quantile correction for data integration reduces the data quality at some extent but it should be acceptable level. Studies using the real datasets show that the quantile correction significantly reduces the bias. These results show that the quantile correction is useful for integration of multiple datasets processed by RMA, and encourage effective use of public microarray data.

Keywords: data integration, quantile correction, microarray, RMA, GeneChip

Background:
Microarray is one of the most conventional methods of comprehensive expression profiling and various platforms are currently available. GeneChip is one of the most popular platforms and various pre-processing methods are currently proposed **[1]**. A huge amount of microarray data has been collected by various groups and stored in public repositories, such as ArrayExpress **[2]** and Gene Expression Omnibus (GEO) **[3]**. Along with the accumulation of the data, various challenges to integrate the different datasets have been achieved so far **[4-9]**. One of the major approaches is a so-called meta-analysis **[4, 5]**. Features in a specific sample are extracted from a dataset and such features extracted from different experiments are merged and used for further analyses. These methods mainly integrate results of a similar purpose, e.g. cancer classification, to improve the statistical power of discrimination. Another approach is a direct connection of the probe intensities. First, probe in different platforms are linked using cross-references, e.g. EntrezGene, and multiple datasets are integrated. Next, the integrated data are transformed by an appropriate processing method, e.g. gene shaving, and then the specific features are extracted. Currently, both strategies mainly focus on the differentially expressed genes in the groups of interest. Although these approaches consider the integration of different platforms, there are also difficulties for the same platform. One of the problems is the bias of signal distribution. The bias is not only due to the experimental procedures but also due to the intrinsic differences of samples. In many cases, but not all cases, the same platform data could be handled like a single dataset if the data are processed all together from raw data and the bias is properly removed. We wondered if we should handle the data sharing the same platform like the meta-analysis or not when the raw data are unavailable. Then, we focused on and studied the integration of pre-processed data in this paper. Although various pre-processing methods have been proposed so far, MAS5.0 **[10]**, RMA **[11]** are two of the most conventional methods for GeneChip and a lot of data processed by these methods have been deposited. Because each sample is processed by MAS5.0 in a one-by-one manner, independent datasets processed by MAS5.0 separately can be compared with each other. On the other hand, the independent datasets processed by RMA separately cannot be compared because RMA calculates the corrected values using other samples' values. In this report, we focused on the data processed by RMA and evaluated quantile correction as a post-processing method for data integration.

Methodology:
An Affymetrix Latin Square (LS) dataset **[12]** and a Drosophila cDNA spike-in experiment (DR) dataset **[13]** were used for the first evaluation. These are triplicate artificial spike-in datasets. Two subsets generated from each dataset were processed all together or independently by RMA. The independently processed data were integrated without or with quantile correction (QC) as a post-processing method. Hereinafter, the datasets processed all together by RMA, the datasets processed independently and integrated without QC, and the datasets processed independently and integrated with QC are represented as S1, S2 and S3, respectively. The signals of triplicate samples in each subset were averaged and evaluated with Receiver Operating Characteristic (ROC) curves and areas under ROC curve (AUC). We adopted two- and four-fold as the thresholds although we had several options of the criteria of the LS dataset. We also used two thresholds, 1.2 and 2.0 fold, for the true positives although several fold-change spike-in transcripts are included in the DR dataset. The DR dataset consists of two conditions of triplicates and each sample contains same 3,860 RNA species. Concentrations of 1,309 transcripts vary from 1.2 to 4 folds between the two conditions and the concentrations of the remaining ones are constant. We originally assigned probesets to the transcripts using available information because the assignment is not shown in the original article. Therefore, the number of assigned probesets is slightly different from the original article **[13]**.

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 9 **Hypothesis**

An atopic dermatitis (AD) dataset and a lung cancer (LC) dataset are used for the second evaluation. The AD dataset is composed of two independent datasets, i.e. GSE5667 and GSE6710. The LC dataset is composed of three independent datasets, i.e. GSE3268, GSE6253 and GSE7670. These GSE datasets were processed all together (S1) or independently by RMA. The separately processed data were integrated without (S2) or with (S3) additional QC. At first, correlation analyses between the S1 and S2 or S3 were carried out. If the bias is properly removed, the slopes, intercepts and Pearson's correlation coefficients (PCCs) are close to 1, 0 and 1, respectively. For the LC dataset, GSE6253 and GSE7670 are used for the evaluation because the raw data of GSE3268 are not available and the S1 for GSE3268 cannot be calculated. Next, expression patterns of house keeping and disease specific genes were characterized. If the bias is properly removed, the house keeping genes would show similar distributions between the subsets which are represented as G1 to G5 in **Figure 1**, and the disease specific genes shows similar distributions between the analogous subsets of the independent experimental datasets. Only S2 and S3 are calculated for the LC dataset because the raw data of GSE3268 are not available.

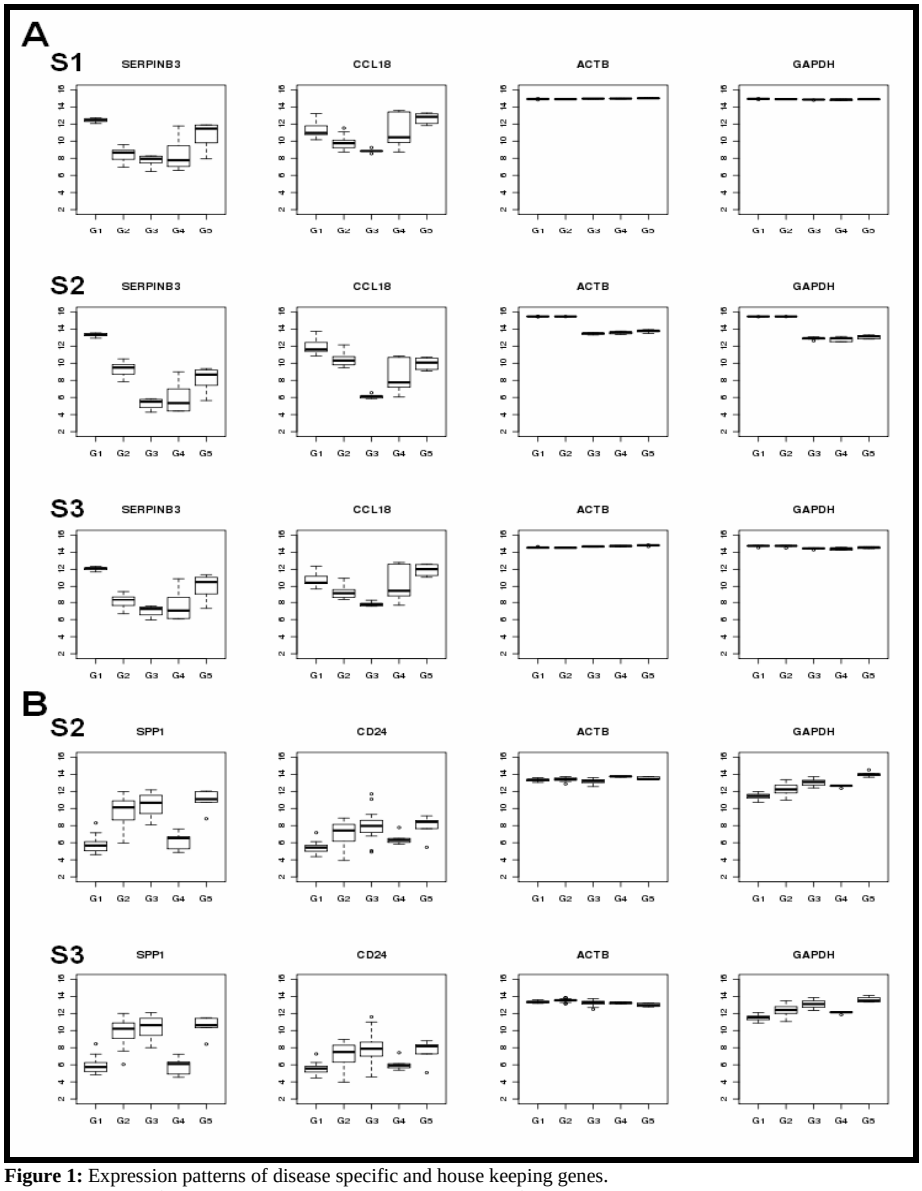


Figure 1: Expression patterns of disease specific and house keeping genes.
The distributions of CCL18, SERPINB3, ACTB and GAPDH of the AD dataset are shown in A, and the distributions of SPP1, CD24, ACTB and GAPDH of the LC dataset are shown in B. The S1 of the LC dataset is not shown because the raw data of GSE3268 are not available. (A) G1: lesional skin samples of patients (GSE6710), G2: non-lesional skin samples of patients (GSE6710), G3: normal skin samples of healthy donors (GSE5667), G4: non-lesional skin samples of patients (GSE5667), G5: lesional skin samples of patients (GSE5667). (B) G1: normal samples of patients (GSE7670), G2: tumor samples of patients (GSE7670), G3: tumor samples of patients (GSE6253), G4: normal samples of patients (GSE3268), G5: tumor samples of patients (GSE3268).

Discussion:

Evaluation using artificial spike-in datasets:

The LS dataset consists of fourteen conditions of triplicates with 42 spike-in transcripts **[12]**. The 42 transcripts are also divided into fourteen groups, i.e. three transcripts in each group, and the concentrations vary from 0.125 pM to 512 pM. The detail is described in Ref.12. We used three triplicate data, i.e. Exp 9, 10 and 11, from the dataset, and concentrations of the spike-in transcripts of Exp. 9 are two- and four-fold in combinations with Exp. 10 and Exp.11, respectively. The signals of group 6, and group 5 and 6 were removed in the first and second combinations, respectively. The ROC curves and AUCs are shown in **Supplementary Figure 1** and **Table 2 (see Supplementary material)**, respectively. The AUCs of the S2 and S3 are lower than the S1. It means that the integration processes without or with QC reduce the quality of data. The ROC curves and the AUCs are also shown in **Supplementary Figure 1** and **Table 2**, respectively. It also indicates that the additional integration processes without or with QC reduce the quality of data.

Evaluation using atopic dermatitis and lung cancer datasets:

At first, we calculated the correlation and regression parameters between the S1 and S2 (P1-2) or the S1 and S3 (P1-3) **(Table 1 see Supplementary material)** of the AD and LC datasets. Although the PCCs and slopes of P1-2 and P1-3 are close to 1, the intercepts of P1-3 are closer to zero than P1-2. These results show the QC removes or reduces the bias between the original datasets. Next, expression patterns of several genes were characterized (Figure 1). Subset 3 (G3) is the skin of normal donors (NS), G1 and G5, and G2 and G4 are the lesion (LS) and non-lesion skin (NLS) of AD patients, respectively. The expression levels of CCL18 and SERPINB3 are LS > NLS > NS and the results are consistent with previous reports **[14, 15]**. Although the expression levels of house keeping genes, beta action (ACTB) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH), are uniform in the S1 and S3, the levels of G1 and G2 are higher than G3, G4 and G5 in the S2. These differences are correlated with the discrepancies of the global distributions (data not shown) and the QC properly removes or reduces the bias in the S2. The expression levels of SPP1 and CD24 in cancer tissues (G2, G3 and G4) are higher than normal tissues (G1 and G5) and the results are also consistent with previous reports **[16, 17]**. The expression revels of GAPDH in cancer tissues are slightly higher than normal tissues although the expression revels of ACTB are uniform between the subgroups.

Although an efficient use of public microarray data is crucial, a significant part of the datasets is difficult to compare with each other. One of the reasons is due to no provision of raw data, and additional correction of the pre-processed data is important to integrate independent datasets. Although keeping data quality is an essential point, the results using the spike-in data show that the integration reduces the data quality. Various pre-processing methods are extensively evaluated using ROC curves and AUCs so far, and such results show that the performances of the methods are various. In consideration of the differences between the methods, the degradations caused by the integration would be an acceptable level. Indeed, data should be processed all together if the raw data are available. The AUCs of S3 are better than S2 in the DS dataset but there are no differences in the LS dataset. The DS dataset contains large bias because 1,309 in 3,860 transcripts are used as spike-in samples. Therefore, the QC reduces the bias of the DS dataset and the AUCs are improved. On the other hand, the LS dataset contains only 42 spike-in probesets and the effect of the spike-in samples to the distribution should be small. Therefore, no significant difference between S2 and S3 should be observed in the LS dataset. The correlations between the S1 and S3 of the AD and LC datasets are high. These results are also obtained with GCRMA **[18]**, but the correlations of data processed with MBEI **[19]** or PLIER **[20]** are significantly lower than RMA or GCRMA (data not shown). These results indicate that the integration using the QC is not applicable to all kinds of pre-processed data and user should confirm whether a set of pre-processed data is suitable for the integration or not using some raw data.

Conclusion:

We examined simple quantile correction as the post-processing method for integration of the data processed by RMA. Our results indicate that the integration using the QC is not applicable to all kinds of pre-processed data. It might be considered that the described results are not informative because our method is applicable to restricted methods. The GeneChip data processed by RMA shares significant part of the public data and the total amount is huge. Therefore, we believe our findings are quite informative for scientists who want to use such pre-processed data, and encourage effective use of public microarray data.

Acknowledgements:

The authors thank Toshio Ota, Masaya Obayashi, Tetsuo Yoshida, Kensuke Kojima and Makiko Shimizu of Kyowa Hakko Kirin for fruitful discussions and valuable comments.

References:

[1] RA Irizarry *et al. Bioinformatics* **22**:789 (2006) [PMID: 16410320]
[2] H Parkinson *et al. Nucleic Acids Res.* **37**: D868 (2009) [PMID: 19015125]
[3] T Barrett *et al. Nucleic Acids Res.* **37**: D885 (2009) [PMID: 18940857]
[4] R Shen *et al. BMC Genomics* **5**: 94 (2004) [PMID: 15598354]
[5] A Teschendorff *et al. Genome Biology* **7**: R101 (2006) [PMID: 17076897]
[6] S Ramaswamy *et al. Nat Genet* **33**: 49 (2003) [PMID: 12469122]
[7] P Warnat *et al. BMC Bioinformatics* **6**: 265 (2005) [PMID: 16271137]
[8] M Benito *et al. Bioinformatics* **20**: 105 (2004) [PMID: 14693816]
[9] H Jiang *et al. BMC Bioinformatics* **5**: 81 (2004) [PMID: 15217521]
[10] http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf
[11] RA Irizarry *et al. Biostatistics* **4**: 249 (2003) [PMID: 12925520]
[12] http://www.affymetrix.com/support/technical/sample_data/datasets.affx
[13] SE Choe *et al. Genome Biol.* **6**: R16 (2005) [PMID: 15693945]
[14] A Pivarcsi *et al. J. Immunol.* **173**: 5810 (2004) [PMID: 15494534]
[15] K Mitsuishi *et al. Clin. Exp. Allergy.* **35**: 1327 (2005) [PMID: 16238792].
[16] G Kristiansen *et al. Br. J. Cancer.* **88**: 231 (2003) [PMID: 12610508].
[17] S Schneider *et al. Clin. Cancer. Res.* **10**: 1588 (2004) [PMID: 15014008]
[18] Z Wu *et al. Journal of the American Statistical Association.* **99**: 909 (2004)
[19] C Li *et al. Proc Natl Acad Sci USA.* **98**: 31 (2001) [PMID: 11134512]
[20] http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf

Edited by P Kangueane

Citation: Yoneya & Miyazawa. Bioinformation 5(9): 382-385 (2011)

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 9 **Hypothesis**

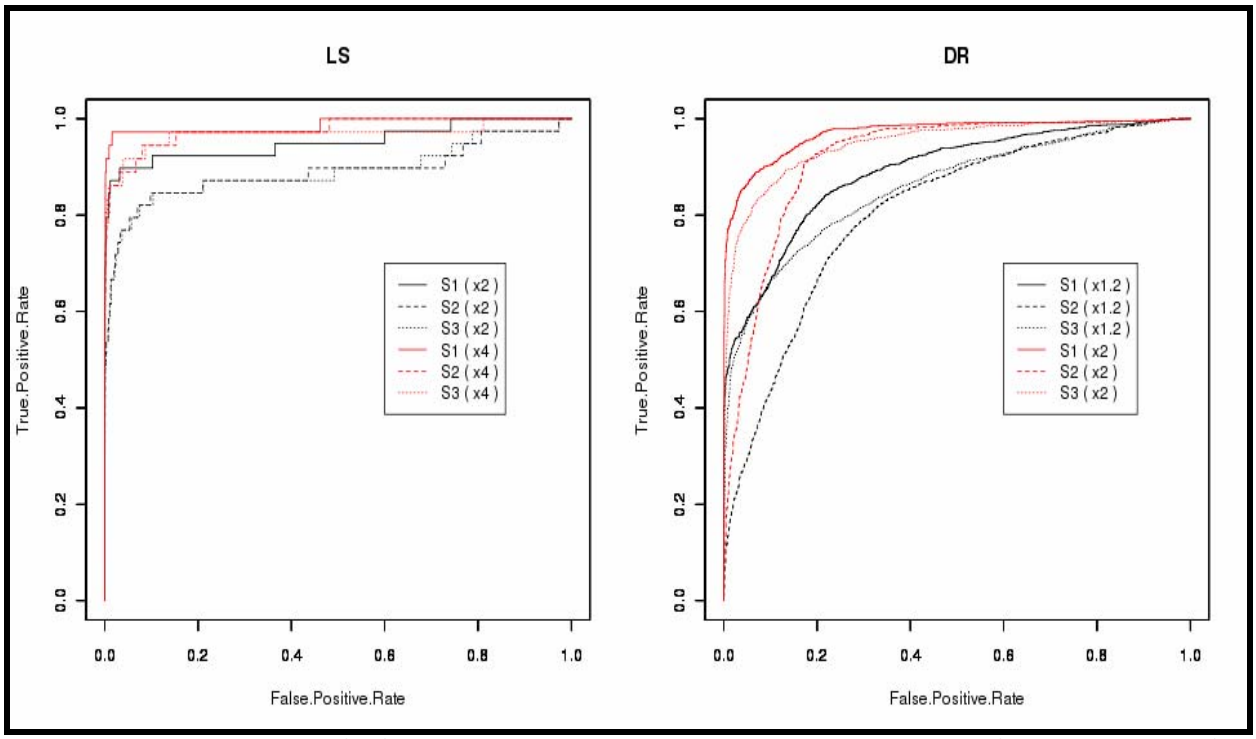
Supplementary material:

Table 1: Regression and correlation parameters of the AD and LC datasets; Slopes, intercepts and Pearson's correlation coefficients (PCCs) between S1 and S2 (P1-2) or S1 and S3 (P1-3) are shown. Intercepts, slopes and PCCs are calculated for each sample and the values belonging to the same GSE set are averaged.

Dataset	Combination	GSE No.	Intercept	Slope	PCC
AD	P1-2	GSE5667	-1.792	0.951	0.986
		GSE6710	0.628	0.998	0.996
	P1-3	GSE5667	-0.328	0.981	0.987
		GSE6710	-0.328	0.979	0.995
LC	P1-2	GSE6253	0.580	0.950	0.986
		GSE7670	-0.161	1.012	0.998
	P1-3	GSE6253	-0.049	1.008	0.986
		GSE7670	0.000	0.998	0.997

Table 2: AUCs calculated from the supplementary Figure 1

Dataset	LS		DR	
Threshold	x2	x4	x1.2	x2
S1	0.95	0.99	0.89	0.97
S2	0.89	0.98	0.81	0.92
S3	0.89	0.97	0.86	0.95



Supplementary Figure 1: ROC curves of artificial spike-in datasets; The ROC curves are calculated with two thresholds for each dataset, two- and four-fold for the Latin Square dataset (LS) and 1.2- and two-fold for the Drosophila cDNA spike-in experiment (DR) dataset. The S1, S2 and S3 are described elsewhere.