Preprocessing Steps for Agilent MicroRNA Arrays: Does the Order Matter?

Li-Xuan Qin*, Huei-Chung Huang* and Qin Zhou
Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. *Equal contribution.

Supplementary Issue: Array Platform Modeling and Analysis (A)

ABSTRACT

MOTIVATION/BACKGROUND: Previous publications on microarray preprocessing mostly focused on method development or comparison for an individual preprocessing step. Very few, if any, focused on recommending an effective ordering of the preprocessing steps, in particular, normalization in relationship to log transformation and probe set summarization. In this study, we aim to study how the relative ordering of the preprocessing steps influences differential expression analysis for Agilent microRNA array data.

METHODS: A set of 192 untreated primary gynecologic tumor samples (96 endometrial tumors and 96 ovarian tumors) were collected at Memorial Sloan Kettering Cancer Center during the period of 2000–2012. From this same sample set, two datasets were generated: one dataset had no confounding array effects by experimental design and served as the benchmark, and another dataset exhibited array effects and served as the test data. We preprocessed our test dataset using different orderings between the following three steps: quantile normalization, log transformation, and median summarization. Differential expression analysis was performed on each preprocessed test dataset, and the results were compared against the results from the benchmark dataset. True positive rate, false positive rate, and false discovery rate were used to assess the effectiveness of the orderings.

RESULTS: The ordering of log transformation, quantile normalization (on probe-level data), and median summarization slightly outperforms the other orderings.

CONCLUSION: Our results ease the anxiety over the uncertain effect that the orderings could have on the analysis of Agilent microRNA array data.

KEYWORDS: microRNA, microarray, preprocessing, normalization, log transformation, probe set summarization

SUPPLEMENT: Array Platform Modeling and Analysis (A)

CITATION: Qin et al. Preprocessing Steps for Agilent MicroRNA Arrays: Does the Order Matter?. Cancer Informatics 2014:13(54) 105–109 doi: 10.4137/CIN.2014.521830.

RECEIVED: November 13, 2014. RESUBMITTED: December 21, 2014. ACCEPTED FOR PUBLICATION: December 25, 2014.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Original Research

FUNDING: This work was supported by NIH grants CA008748 and CA151947 (LXQ, HCH, and QZ). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: qinl@mskcc.org

COPYRIGHT: © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

PAPER SUBJECT TO INDEPENDENT EXPERT BLIND PEER REVIEW BY MINIMUM OF TWO REVIEWERS. ALL EDITORIAL DECISIONS MADE BY INDEPENDENT ACADEMIC EDITOR. UPON SUBMISSION MANUSCRIPT WAS SUBJECT TO ANTI-PLAGIARISM SCANNING. PRIOR TO PUBLICATION ALL AUTHORS HAVE GIVEN SIGNED CONFIRMATION OF AGREEMENT TO ARTICLE PUBLICATION AND COMPLIANCE WITH ALL APPLICABLE ETHICAL AND LEGAL REQUIREMENTS, INCLUDING THE ACCURACY OF AUTHOR AND CONTRIBUTOR INFORMATION, DISCLOSURE OF COMPETING INTERESTS AND FUNDING SOURCES, COMPLIANCE WITH ETHICAL REQUIREMENTS RELATING TO HUMAN AND ANIMAL STUDY PARTICIPANTS, AND COMPLIANCE WITH ANY COPYRIGHT REQUIREMENTS OF THIRD PARTIES. THIS JOURNAL IS A MEMBER OF THE COMMITTEE ON PUBLICATION ETHICS (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Array normalization, data transformation, and probe-set summarization are the three most common preprocessing steps for microarray data analysis.1,2 In practice, the three steps have been implemented in different orders depending on the specific data preprocessing pipeline. For example, the RMA pipeline follows the order of quantile normalization, log transformation, and probe-set summarization for the analysis of Affymetrix gene expression arrays,3 while AgiMicroRNA uses the order of log transformation, probe-set summarization, and quantile normalization for the analysis of Agilent microRNA (miRNA) expression arrays.4 It remains uncertain whether the ordering affects downstream data analysis results.

In our previous works, two datasets were generated on the same set of tumor samples, where one dataset had no confounding array effects by experimental design and served as the benchmark, and another dataset exhibited array effects and served as the test data for evaluating normalization methods.5,6 Using the two datasets, we compared the relative performance of different array normalization methods and showed that quantile normalization performs better relative to the other normalization methods we examined.6 During the analyses, we preprocessed the test dataset following the order of log transformation, array normalization, and median summarization. In this follow-up paper, we set out to further evaluate the performance of quantile normalization when combined with log transformation and probe-set summarization in different orderings.

Methods

Data collection. A set of 192 untreated primary gynecologic tumor samples (96 endometrial endometrial tumors and 96 serous ovarian tumors) were collected at Memorial Sloan Kettering Cancer Center during the period of 2000–2012. The samples were profiled using the Agilent Human miRNA Microarray (Release 16.0), following the manufacturer’s protocol. Two datasets were originated from
the same set of samples using different array assignments and handling processes. The first dataset was created using blocked randomization to assign arrays to samples and was handled by one technician in one run. The second dataset had an array assignment in the order of tumor sample collection and was handled by two technicians in multiple runs, which imitated the result under a typical laboratory setting. In this study, we inherit these datasets and refer to the first dataset as the “block randomized” or “benchmark” dataset and the latter as the “test” dataset in this paper. More details on data collection can be found in the study by Qin et al.\(^5\)

**Data preprocessing for the test dataset.** Data transformation, array normalization, and probe-set summarization are the steps in microarray data preprocessing.\(^6\) In this study, our choice of method for each of these three steps was log\(_2\) transformation, quantile normalization, and median summarization. There are a total of six combinations for ordering these three preprocessing steps. Since arrangement of logarithm and median results no difference mathematically, our choices of ordering trimmed to just four, which we call Orders A–D (Table 1). Orders A and B both applied quantile normalization to the probe level data. Their difference was whether quantile normalization was applied on the raw data or the logged data. Orders C and D both applied median summarization to replicate probes in each probe set before the normalization step, and then, quantile normalization was either applied to the raw or the logged probe-set level data. A comparison between the two ordering pairs, A–B versus C–D, can help answer the question whether quantile normalization should be applied at the probe level or the probe-set level. To serve as a basis in comparison, we also performed log transformation and median summarization – without quantile normalization – to our test dataset. In this study, we call this “Reference”. Comparing against Reference provides a context in understanding the magnitude of difference among Orders A–D.

**Data preprocessing for the benchmark dataset.** The use of uniform handling and blocked randomization has been shown to be effective at controlling confounding array effects.\(^6\) Therefore, the benchmark dataset was preprocessed only with log\(_2\) transformation and median summarization and without quantile normalization.

**Differential expression analysis for the benchmark dataset and the test dataset.** Two-sample \(t\)-statistic was applied to evaluate the level of statistical evidence of differential expression between the two tumor groups (ovarian cancer versus endometrial cancer) using the preprocessed data. We computed the \(t\)-statistic for each of the 3523 probe sets (or referred as “markers” interchangeably in this paper) on the array with a null hypothesis of no differential expression. We obtained two-sided \(P\)-values and used 0.01 as the cutoff. \(P\)-values smaller than 0.01 indicated differentially expressed, and \(P\)-values greater than 0.01 suggested no differential expression.

**Method comparison.** The obtained differential expression results for the test dataset were tabulated against the results from the benchmark dataset. True positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR) were computed for each preprocessed data.\(^7,8\) Their detailed definitions are provided in Table 2. A preprocessing ordering outperforms other orderings if it results in a higher TPR, lower FPR, and lower FDR.

**Simulation.** One characteristic of the Agilent miRNA array data is that they possess a very low level of variation between probe replicates.\(^3\) We speculate that this characteristic may influence the difference between applying quantile normalization to the probe level and to the probe-set level data. To test this hypothesis, we conducted a simulation study to further examine the effect of preprocessing step orderings when the between-probe variation was increased in the test dataset. The simulation followed the below steps.

1. Generate simulated test dataset: we generated random noise from a Gaussian distribution of zero mean and \(\sigma\) standard deviation and added them to the probe-level data of the test dataset. We considered four possible values for \(\sigma\): 0.2, 0.4, 0.8, and 2. These values of \(\sigma\) are multiples of the estimated standard deviation of probe replicates (0.4) observed in the empirical data.
2. Repeat and average: for each \(\sigma\) value, we created 100 simulated datasets and applied the preprocessing, differential expression analysis, and method comparison as described in the previous section. TPRs, FPRs, and FDRs were then averaged across the 100 simulation runs.

---

**Table 1.** Orderings of preprocessing steps applied to the test data.

| ORDERING | FIRST                  | SECOND                | THIRD     |
|----------|------------------------|-----------------------|-----------|
| A        | Quantile normalization | Log\(_2\)              | Median    |
| B        | Log\(_2\)              | Quantile normalization| Median    |
| C        | Median                 | Quantile normalization | Log\(_2\) |
| D        | Median                 | Log\(_2\)              | Quantile normalization |
| Reference| Log\(_2\)              | Median                | –         |

---

**Table 2.** Statistical measures for method comparison.

|                        | DECLARED SIGNIFICANT IN BENCHMARK | DECLARED INSIGNIFICANCE IN BENCHMARK |
|------------------------|----------------------------------|-------------------------------------|
| Declared significant in test data | True positive (TP)               | False positive (FP)                  |
| Declared insignificance in test data | False negative (FN)              | True negative (TN)                   |
**Results**

**Empirical evaluation.** Table 3A reports the TPRs, FPRs, and FDRs in percentage for each preprocessing ordering when applied to the test dataset and compared against the benchmark. In particular,

1. The result for Order A (quantile–log$_2$–median) identified 710 differentially expressed markers. Within these differentially expressed markers, 328 were identified as differentially expressed in the benchmark (TPR: 93.5%, FPR: 12.0%, FDR: 53.8%).

2. Order B (log$_2$–quantile–median) showed 708 differentially expressed markers and 328 were also differentially expressed in the benchmark (TPR: 93.5%, FPR: 12.0%, FDR: 53.7%).

3. Order C (median–quantile–log$_2$) indicated that 710 markers were different, among which 326 were also different in the benchmark (TPR: 92.9%, FPR: 12.1%, FDR: 54.1%).

4. Under Order D (median–log$_2$–quantile), 712 markers were differentially expressed and 326 of them were also differentially expressed in the benchmark (TPR: 92.9%, FPR: 12.2%, FDR: 54.2%).

5. Test dataset prepared with the Reference (log$_2$–median) indicated 1934 differentially expressed markers, and 185 of which were also differentially expressed in the benchmark (TPR: 52.7%, FPR: 55.1%, FDR: 90.4%).

The empirical evidence from the test dataset suggested that Orders B and A slightly outperformed Orders C and D. For Orders A and B, their TPRs were identical to each other and a little but not substantially higher than that for Orders C and D; their FPRs were a little lower than C’s and D’s. Moreover, Order B had the smallest FDR (53.7%) and hence was the most desirable of all. Order A closely followed with just less than 0.1% difference in FDR (53.8%), which corresponded to two more falsely discovered markers than B. Both Orders A and B applied quantile normalization to the probe level data, while Orders C and D applied it to the probe-set level. Order B differed from Order A in that it applied log transformation before quantile normalization.

To summarize, our result suggests that applying quantile normalization to the probe-level data is preferred than to the probe-set level data, and log transformation before quantile normalization is slightly more beneficial. Our results are robust to the P-value cutoff used to identify significant genes (Table 3B and Fig. 1).

**Simulation study.** We further assessed the robustness of our empirical observation to the level of between-probe variation. The result of this simulation study is displayed in Figure 2. It shows that

1. Comparing across the five levels of $\sigma$ under Reference, we see almost no difference in the simulated data from the original test dataset. Since Reference was quantile normalization free, the extra noise added in the simulation was simply taken out by median summarization.

2. As $\sigma$ increased, FDRs and FPRs decreased for each of the orderings A–D. As shown in our previous work, quantile normalization tended to underestimated standard deviation in the original test dataset, which consequently inflated t-statistics and produced smaller P-values and more positive markers. As extra variation was introduced to the test data, it became more robust to the underestimation, and hence, less prone to capturing false positive markers.

**Table 3A.** Results of differential expression analysis for the test dataset. A P-value cutoff of 0.01 was used to claim significant markers in both the benchmark dataset and the test dataset. At this cutoff, the number of significant markers was 351 in the benchmark dataset.

| ORDERING | NUMBER OF SIGNIFICANT MARKERS | TPR% | FPR% | FDR% |
|----------|-------------------------------|------|------|------|
| A        | 710                           | 93.5 | 12.0 | 53.8 |
| B        | 708                           | 93.5 | 12.0 | 53.7 |
| C        | 710                           | 92.9 | 12.1 | 54.1 |
| D        | 712                           | 92.9 | 12.2 | 54.2 |
| Reference| 1934                          | 52.7 | 55.1 | 90.4 |

**Table 3B.** Results of differential expression analysis for the test dataset. A P-value cutoff of 0.001 was used to claim significant markers in both the benchmark dataset and the test dataset. At this cutoff, the number of significant markers was 186 in the benchmark dataset.

| ORDERING | NUMBER OF SIGNIFICANT MARKERS | TPR% | FPR% | FDR% |
|----------|-------------------------------|------|------|------|
| A        | 443                           | 94.6 | 8.00 | 60.3 |
| B        | 441                           | 94.6 | 7.94 | 60.1 |
| C        | 441                           | 94.6 | 7.94 | 60.1 |
| D        | 442                           | 94.6 | 7.97 | 60.2 |
| Reference| 281                           | 54.8 | 5.36 | 63.7 |
Figure 1. Receiver Operating Characteristics (ROC) curves comparing the differential expression P-values for the test dataset versus that for the benchmark dataset, treating the latter as a gold standard.

Figure 2. Results of the simulation study. Dots represent the means and error bars the standard deviations for each summary statistics (TPR, FPR, and FDR) across the 100 simulation datasets for each simulation setting. X axis indicates the value for \( \sigma \) (the standard deviation of the zero mean Gaussian distribution from which the extra level of noise was generated and added to the probe-level data of the test dataset).
As the value of $\sigma$ increased from 0 to 2, Orders A and B remained to have higher TPRs and lower FPRs and FDRs than Orders C and D; however, the difference shrunk. This was primarily due to a drop in TPR for Orders A and B and an increase in TPR for Orders C and D (e.g., $\sigma = 0$ to $\sigma = 2$: Order A – TPR: 93.5–93.1%, Order B – TPR: 93.5–93.1%, Order C – TPR: 92.9–93.0%, Order D – TPR: 92.9–93.0%, Reference – TPR: 52.7–52.7%).

In summary, the difference of applying quantile normalization to probe-level data versus probe-set level data depended on the level of between-probe variation as we have hypothesized; there is evidence suggesting that smaller between-probe variation can lead to greater advantage for applying quantile normalization at the probe level.

**Conclusion and Discussion**

Our results showed that the ordering of the three data preprocessing steps had a very small effect on the downstream analysis of differential expression in the context of Agilent miRNA array data. Nevertheless, the ordering of log transformation, quantile normalization on probe-level data, and median summarization slightly outperformed the other three orderings.

Our conclusion eases the anxiety over the uncertain effect that the orderings could have on data analysis of Agilent miRNA arrays. It can be potentially generalized to other types of microarray data where the three preprocessing steps — log transformation, quantile normalization, and median summarization — are needed and the between-probe variability is small.

**Author Contributions**

Conceived and designed the experiments: LXQ. Analyzed the data: HCH, QZ. Wrote the first draft of the manuscript: HCH. Contributed to the writing of the manuscript: LXQ, HCH. Agree with manuscript results and conclusions: LXQ, HCH, QZ. Jointly developed the structure and arguments for the paper: LXQ, HCH. Made critical revisions and approved final version: LXQ. All authors reviewed and approved of the final manuscript.

**REFERENCES**

1. Bolstad B, Irizarry R, Astrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185–93.
2. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*. 2006;7:55–65.
3. Irizarry R, Hobbs B, Collin F. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
4. López-Romero P. Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics*. 2011;12:64.
5. Qin LX, Zhou Q, Bogomolnyi F, et al. Blocking and randomization to improve molecular biomarker discovery. *Clin Cancer Res*. 2014;20(13):3371–8.
6. Qin LX, Zhou Q. MicroRNA array normalization: an evaluation using a randomized dataset as the benchmark. *PLoS One*. 2014;9:e98879.
7. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press; 2003.
8. Storey J. The positive false discovery rate: a Bayesian interpretation and the Q-value. *Ann Stat*. 2003;31(6):2013–35.