



A study of inter-lab and inter-platform agreement of DNA microarray data

Huixia Wang¹, Xuming He¹, Mark Band², Carole Wilson², and Lei Liu^{2*}

¹Department of Statistics, University of Illinois at Urbana-Champaign, 101 Illini Hall, 725 South Wright Street, Champaign, Illinois 61820

²W. M. Keck Center for Comparative and Functional Genomics, University of Illinois at Urbana-Champaign, 1201 W. Gregory Drive, Urbana, Illinois 61801

Received line

ABSTRACT

Motivation:

As the gene expression profile data from DNA microarray accumulate rapidly, there is a natural need for comparing data across labs and platforms. Unlike DNA sequence comparisons, comparisons of microarray data can be quite challenging due to data complexity and variability. Different labs may adopt different technology platforms. What degree of agreement can we expect from different labs and different platforms? To address this question, we conducted a study of inter-lab and inter-platform agreement of microarray data across three platforms and three labs. The statistical measures of consistency and agreement are the Pearson correlation, intra-class correlation and kappa coefficients.

Results:

The three platforms compared were Affymetrix GeneChip, custom cDNA arrays, and custom oligo arrays. Using the within-platform variability as a benchmark, we found that these technology platforms exhibited an acceptable level of agreement. RNA samples used in the experiments, however, could be more variable than the technology platforms in the study of microarray.

Availability:

The results and the data sets generated for this paper are available at the following site:

http://titan.biotech.uiuc.edu/cross_platform/

Contact: leiliu@uiuc.edu

Keywords:

Microarray; cross platform comparison; intraclass correlation; kappa statistics

*To whom correspondence should be sent.

INTRODUCTION

Diversity of microarray data poses some unique and interesting questions on cross-experiment comparisons and the analysis tools needed for such comparisons. Since the invention of the microarray technology in 1995 (Schena et al.), statistical methods and data mining techniques specific for microarray data have

mushroomed (e.g., Quackenbush, 2001), many of which have been packaged into commercial software such as GeneSpring and Spotfire. Such tools are useful for handling individual experiments, including quality control, significance testing, and clustering. However, researchers have questioned whether studies across different labs and technology platforms will have an acceptable level of agreement.

Possible incompatibility of results between similar microarray experiments is a major challenge that needs to be addressed, even though the data produced within a single experiment may be consistent and easy to analyze. Different labs produce microarray data in different ways using different technology platforms, such as Affymetrix GeneChip, spotted cDNA array, and spotted oligo array. Affymetrix GeneChip uses one fluorescent dye while the spotted array uses two fluorescent dyes in the experiments. Direct comparison of raw data obtained from different technologies may not be meaningful. Instead, the final form of the data is often presented as relative expression levels, mostly ratios of intensities, after some statistical treatments including normalization, filtering and significance testing. Experiments using different technologies require different protocols for analyzing the raw data to derive the ratios. Scientists have published microarray data in a variety of formats including raw intensities and ratios of intensities. Does it make a difference which technology platform is chosen? Can we make use of the studies from different platforms and labs?

To answer these questions and provide some guidance for platform comparisons, we report on a comparative study of three different platforms. The experiment is a simple two-tissue comparison between mouse liver and spleen. We used previously published data sets from two different sources as well as new data sets produced in house. This study provides a basis for further development of methodologies for comparing microarray data across different experiments and for the integration of microarray data from different labs.

DATA COLLECTION

As summarized in Table 1, a total of five data sets were collected from either a public source or in house. The samples for the experiments were normal mouse liver and spleen RNA, which were purchased from Clontech (Catalog No. 64042-1 liver; Catalog No. 64044-1 spleen) except for the data set GNF_Affy generated by Su et al. (2002) at the Genomics Institute of the Novartis Research Foundation. Detailed sample descriptions for the GNF_Affy data can be found at <http://expression.gnf.org>. Two data sets were downloaded from the NCBI Gene Expression Omnibus (<http://www.ncbi.nih.gov/geo/>), which were generated by Choi et al. at California Institute of Technology (Cal Tech) using oligo and cDNA arrays respectively. Two other data sets were generated at the Functional Genomics Unit at the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois using an in-house printed cDNA mouse array and Affymetrix mouse expression set 430A. Another data set was downloaded from <http://expression.gnf.org>, which was generated using Affymetrix Murine Genome set U74Av2.

Table 1. Summary of data collection

Data Set	Array	Genes	Sample	Replicates	Data type	Platform	Lab	Data Source
Keck_cDNA	CI 15K cDNA in house	15K	Clontech	4	Raw intensity	cDNA	Keck	In house
Keck_AV	Affymetrix 430A	23K	Clontech	2	AV(Bioconductor)	Affymetrix	Keck	In house
Keck_LW	Affymetrix 430A	23K	Clontech	2	Li and Wong	Affymetrix	Keck	In house
CIT_cDNA	Riken16K cDNA by Agilent	16K	Clontech	3	Raw intensity	cDNA	Cal Tech	NCBI GEO
CIT_Oligo	Riken16K Oligo by Agilent	16K	Clontech	3	Raw intensity	Oligo	Cal Tech	NCBI GEO
GNF_Affy	Affymetrix U74Av2	12K	In house	2	AV(MAS4.0)	Affymetrix	GNF	expression.gnf.org

METHODS

Data processing:

For spotted arrays (Keck_cDNA, CIT_cDNA, and CIT_Oligo), we used raw intensity data from both Cy5 and Cy3. We filtered out the non-expressive data points using median plus three times Median Absolute Deviation (MAD, Huber, 1981, p. 107) of the control genes as a criterion. We then performed global lowess normalization on each slide. For the Keck_cDNA data, we also performed paired-slide normalization following the method in Yang et al. (2001) because of the dye swap in the experiment.

For the in-house Affymetrix data, we used two methods to generate ratios. One is based on the Bioconductor software using the average difference (AD) between Perfect Match (PM) and Mis-Match (MM) probe pairs; the other is the model-based expression indexes developed by Li and Wong (2001). We also performed global lowess normalization on each slide. The GNF Affymetrix data were available to us only in the format of average differences.

Gene matching across arrays:

There are five different data sets in this study. The origins of the genes vary in those datasets. In order to

study inter-lab agreement, we have to first identify the same genes represented in different arrays. Based on the annotation of each data set, we found that we could maximize the number of cross-matched genes using the mouse UniGene IDs. Based on the common UniGene IDs, we found 517 common genes across all five different data sets. But in the pairwise comparisons, the number of common genes ranged from 1,463 to 4,782 (see Table 2). All comparisons between data sets were made from the matched genes.

Statistical procedures for inter-platform and inter-lab comparisons

In the analyses, the ratio is defined as normalized intensity from liver samples versus that from spleen samples.

Agreement of two-fold changes using kappa coefficients

An intuitive measurement of agreement is to count the percentage of genes falling in the same categories (two-fold up-regulated, no change, and two-fold down-regulated). However, this percentage can be high even if the data obtained from different platforms are not so

compatible. Usually the ratios for the great majority of genes show changes significantly less than 2, and the percentage of agreement can be high just due to chance. To adjust for this excess agreement expected by chance, we prefer to use the kappa coefficient, which is a popular measure of inter-rater agreement in many other areas of science. The kappa coefficient was first proposed by Cohen (1960) for analyzing dichotomous responses, and was extended later to more than two categories of responses. We used this method to check three categories (two fold up-regulated, no change, and two fold down-regulated), and computed the kappa coefficients between two data sets from 3 by 3 frequency tables. For a study of q categories, the kappa coefficient is calculated by:

$$\text{kappa} = \frac{P_a - P_e}{1 - P_e}, \text{ where } P_a = \frac{1}{n} \sum_{k=1}^q n_{kk} \text{ is the overall}$$

agreement probability, and $P_e = \sum_{k=1}^q \frac{n_{+k}}{n} \cdot \frac{n_{k+}}{n}$ is the measure of the likelihood of agreement by chance.

Correlation and intra-class correlation of the ratios

We used two measures of correlation to compare the ratios from different data sets: Pearson correlation and intra-class correlation. Intra-class Correlation Coefficient (ICC) measures the inter-rater reliability relative to the total variability of the ratios. Here, a rater could be a replicate or a technology platform. ICC is the variance of different ratios between Unigene IDs, σ_b^2 , divided by the total variance σ_T^2 . A high ICC close to 1 means that the inter-rater ratios vary little relative to the overall variability in the data. In computing the ICC for the replicates, $\sigma_T^2 = \sigma_b^2 + \sigma_e^2$, where σ_e^2 is the variance within Unigene IDs. If we consider lab as a random effect in the overall comparison, the total variance σ_T^2 will equal $\sigma_b^2 + \sigma_c^2 + \sigma_e^2$, where σ_c^2 is the variance between labs.

RESULTS

Consistency of replicates

One indication of data reliability is the consistency of replicates in a particular data set. We used kappa coefficients as well as the two correlations discussed above on the replicates within each data set. Those measures set a benchmark against which the reliability of different platforms can be assessed; see Figure 1. The Keck_cDNA has four replicates from double spots of each gene on the array and from the dye swap. Therefore, there are six pairwise comparisons. The CIT_cDNA and CIT_Oligo have three replicates each; therefore, there are three pairwise comparisons. All Affymetrix data sets have two replicates and only one comparison. From Figure 1, we see that the replicates were quite consistent within a technology. With the exception of GNF_Affy, the replicates in all the data sets showed pairwise correlations of 0.8 or higher, intra-class correlations of 0.7 or higher, and kappa coefficients of 0.5 or higher. The data from the Cal Tech (CIT_cDNA and CIT_Oligo) showed the highest agreement among the replicates, and the data from GNF showed a low level of agreement between replicates. These results suggest a lab effect in microarray data experiments.

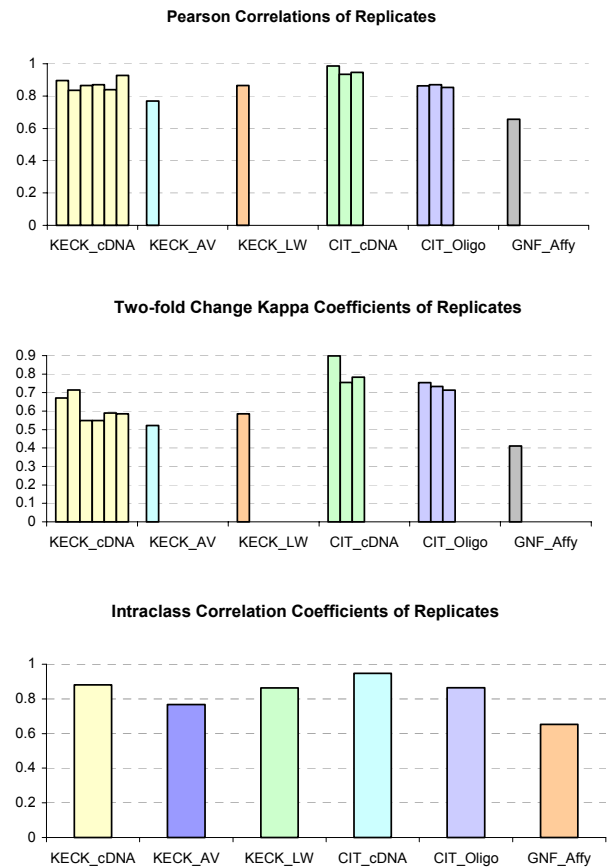
For the Keck_cDNA data, we can see that two comparisons gave slightly higher agreement than the other four comparisons. It is, we believe, due to the double spots of each gene on the array. The two comparisons with higher agreement are the comparisons between the replicates within the slides.

Pairwise comparisons among data sets

Using the matched genes by common UniGene IDs, we compared different data sets in our study. Table 2 shows the Pearson correlation and ICC on ratios, and the kappa coefficients for two-fold changes. Except for the comparisons with the GNF_Affy, the kappa coefficients are 0.4 – 0.5, which is only slightly below 0.6 for the replicates within the Keck_cDNA data.

Based on ICC and kappa, we found that the CIT_cDNA and CIT_Oligo, CIT_cDNA and Keck_AV or LW,

Figure 1. Consistency of replicates



CIT_Oligo and Keck_AV or LW have higher agreement. The GNF_Affy dataset has lower agreement with the other 4 data sets. The same conclusion can be drawn from the sensitivity check of the comparisons among all of the five data sets. We did this by leaving out one data set at a time and recording the changes of ICC as shown in Table 3. Excluding GNF resulted in the largest increase in ICC.

The overall comparison of all five data sets gives ICC=0.643 as comparing to the ICC around 0.8 for replicates within each data set. These results indicate that the agreement of different technologies is decent. Since we only used a subset of data for the study, one may argue that the variation of data may be different depending on the subset. By comparing the box plot of the full data set and that of the subset we used for the overall ICC analysis, we found no significant difference between the variations. Figure 2 shows the box plots for the full and subset data of CIT_cDNA.

Table 2. Pearson correlation, kappa coefficient, and ICC for pairwise comparisons

Comparisons	NO. of Matched Unigene IDs	Pearson Correlation	Kappa Coefficient	ICC
Keck_AV vs. GNF_Affy	3211	0.580	0.294	0.592
Keck_AV vs. Keck_cDNA	4107	0.675	0.406	0.657
Keck_AV vs. CIT_cDNA	3313	0.713	0.456	0.737
Keck_AV vs. CIT_Oligo	4246	0.760	0.530	0.737
GNF_Affy vs. Keck_cDNA	1947	0.534	0.270	0.614
GNF_Affy vs. CIT_cDNA	1463	0.529	0.267	0.653
GNF_Affy vs. CIT_Oligo	1987	0.564	0.294	0.619
Keck_cDNA vs. CIT_cDNA	2730	0.637	0.390	0.704
Keck_cDNA vs. CIT_Oligo	2986	0.622	0.403	0.671
CIT_cDNA vs. CIT_Oligo	3170	0.731	0.490	0.771
Keck_LW vs. GNF_Affy	3633	0.594	0.300	0.606
Keck_LW vs. Keck_cDNA	4513	0.680	0.424	0.681
Keck_LW vs. CIT_cDNA	3775	0.780	0.451	0.763
Keck_LW vs. CIT_Oligo	4782	0.731	0.492	0.740

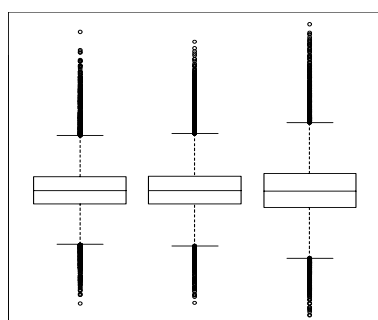
Table 3. Sensitivity checks of the comparison among all of the data sets

	ICC
All five data sets	0.643
Leave out GNF	0.706
Leave out Keck_LW	0.633
Leave out Keck_cDNA	0.648
Leave out CIT_cDNA	0.622
Leave out CIT_Oligo	0.651

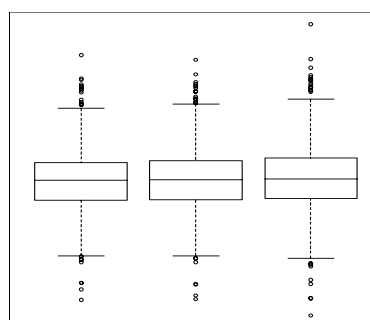
Taking a deeper look at the comparisons between different technologies within the same lab, Keck, we found from Table 4 that Keck AV detected a lot more two-fold changes than Keck cDNA. This is consistent with Figure 3, which showed that the Affymetrix data had a much higher variability or scale than the cDNA

Figure 2. Box plots of the CIT_cDNA data

A. Box plots of the full data set of CIT_cDNA, a total 12471 Unigenes



B. Box plots of a subset of CIT_cDNA, overlapped with the other 4 datasets, a total 517 Unigenes

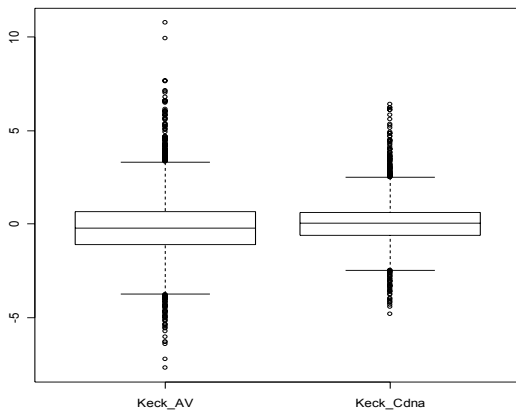


data. This suggests that the filtering process used for cDNA might lead to lower efficiency in detecting changes, but more studies are needed to support this claim.

DISCUSSION

When we matched genes from different arrays using the mouse UniGene IDs, we found that there were multiple gene IDs in an array corresponding to one UniGene ID. Those genes were considered as "duplicate" genes, which made the cross matching of genes more complicated. A common approach is to average the expressions of those "duplicate" genes, however we considered these "duplicate" genes as replicates in the technology and lab comparisons. One observation we should make is that the variability among these "duplicate" genes can be large. For

Figure 3. Box plots of Keck_AV and Keck_cDNA



example, in the third array of CIT cDNA, there are total of 11301 genes (upon filtering), corresponding to 8318 UniGene IDs. Among the 8318 UniGene IDs, 1708 of them had “duplicate” genes. About one third of those “duplicate” genes had standard deviations greater than 0.5 for the ratios. Note that matching of genes was performed only for the comparison among data sets, and that we did not use the UniGene IDs in measuring the consistency within the same data set. This means that the agreement measures we obtained from different data sets were expected to be slightly lower than those from the replicates, even if the actual agreement was the same within or between data sets.

Table 4. Frequency table for Keck_AV and Keck_cDNA, kappa coefficient=0.406

Keck_AV	Keck_cDNA			
Frequency	2 fold down	No change	2 fold up	Total
2 fold down	446	660	22	1128
No change	157	1902	149	2208
2 fold up	18	340	413	771
Total	621	2902	584	4107

From the present study, we showed that the GNF_Affy data set has the lowest agreement with other data sets. This difference is compounded with the facts that the consistency of replicates within the data set is the lowest and the sample used to generate the data was different from those used by other labs. We believe that the technology platform plays a minor role in the disagreement, but the variation introduced by sample differences is one of the major factors. It has been shown that the expression level can vary significantly between genetically identical mice (Pritchard et al. 2001). Variation among different individuals can be a significant factor for sample differences. The results also indicate that data generated from different labs may have different quality even among the replicates, and thus quality control is important.

We recognize that the present study has limitations. The results were generated from a very limited number of data sets. Using UniGene ids for gene matching across data sets can also be a shortcoming as indicated above. We could not provide convincing support about the sample effect, since only one sample was different in one lab from the rest. An expanded experiment with a more careful design to produce more data sets from multiple labs is desirable for future studies.

CONCLUSION

In this paper, we aim to address the issues in comparing microarray data across different platforms and different labs. We demonstrated that the consistency of replicates in each experiment varies from lab to lab. With high consistency of replicates, different technologies seem to show good agreement either within one lab or between two labs. The source of RNA samples may also make a difference in microarray data, however in our present study we do not show conclusive results pertaining to possible sample or lab effects, because we did not have data collected from two different samples within one lab.

ACKNOWLEDGEMENT

This study was supported in part by the NIH Grant No. 2 P30 AR41940-10. We would like to acknowledge Al Bari for his work on printing the mouse cDNA array at the Keck Center.

REFERENCES

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37-46.

Gwet, K. (2002) Computing Inter-Rater Reliability With the SAS System *Statistical Methods For Inter-Rater Reliability Assessment*, No. 3, October 2002.

Huber, P.J. (1981) Robust Statistics. John Wiley & Sons.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* 98, 31-36.

- Pritchard, C.C., Hsu, L., Delrow, J., and Nelson, P.J. (2001) Project normal: Defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci.* 98: 13266-13271.
- Quackenbush, J., (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, 2(6):418-427.
- Schena, M., Shalon, D., Davis, R.W., and Brown, P.O., (1995) Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science* 270:467-470.
- Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., Patapoutian, A., Hampton, G.M., Schultz, P.G., and Hogenesch, J.B. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* 99(7):4465-4470.
- Yang, Y.H., Dudoit, S., Luu, P., and Speed, T.P. (2001) Normalization for cDNA microarray Data. *Microarrays: Optical Technologies and Informatics*. SPIE BIOS 2001, San Jose, CA.