

Making Sense of Human Lung Carcinomas Gene Expression Data: Integration and Analysis of Two Affymetrix Platform Experiments

Xiwu Lin, Daniel C. Park, Sergio Eslava, Kwan R. Lee, Raymond L.H. Lam, Lei A. Zhu

GlaxoSmithKline Pharmaceuticals R&D
Biomedical Data Sciences
1250 S. Collegeville road
Collegeville, PA 19426

ABSTRACT

CAMDA03 data sets for the 2003 challenge focused on lung cancers. Four microarray platform data sets were released for integration and combined analysis. We integrated the first two platforms (Harvard and Michigan). There are two types of data to use for integration: processed data and probe level data in the form of CEL files. The processed data could have been generated by the old version of Affymetrix MAS software (version 4.0) and had many problems including negative values and large variability among the low expressed genes. For accuracy, we generated new MAS 5.0 expression measures for each of the two Affymetrix platforms (before integration) from probe level CEL data using Bioconductor. Our MAS 5.0 based expression measures made clearer separation of cancer types and we believe that some of the previous findings based on MAS 4.0 expression measures can be misleading.

There are several more challenges for the proper integration of the two data sets on two different types of Affymetrix gene chips. One issue is the different number of genes (clones) in the two data sets. We have used two different approaches: one approach was using array comparison spreadsheet designed by Affymetrix and the other approach was simply using gene names for merging the data sets. Either way of combining the data sets resulted in approximately the same information as can be seen in the principal component analysis (PCA). Another issue is making the distribution of the expression measures comparable across the two data sets. In other words, normalization issues. We have chosen quantile normalization of expression measures across the platforms for each gene to achieve the goal.

The final data set combined this way yielded interesting results. Before normalization, two sets of data were completely disjoint as can be seen from the PCA plot. After normalization, they cover the approximately same space of projection. Moreover normal lung samples of combined data had clearer separation from the rest of the data. In order to further demonstrate the validity of

integration we used the Harvard data as a training set and Michigan data as test set in the discrimination of normal lung samples from the adenocarcinomas. All three prediction models showed good accuracy. Neural networks, for example, predicted the Michigan data with 99% accuracy with only one false positive. We believe the integrated data set is ready for analysis to answer many important biological questions. For example, we have performed survival tree analysis on the combined data to predict mortality and associated genes. Those resultant survival tree and associated genes might be of considerable biological interest and need further consideration.

keywords: Gene expression data, integration and analysis, Affymetrix MAS, principal component analysis, partial least squares, frailty Cox proportional hazard model, survival tree

1. DATA PROCESSING AND INTEGRATION

1.1 Processed Data vs. Raw Data

Processed data from Harvard and Michigan was available but it could have been generated by version 4 of Affymetrix' MAS or others. It is well known that MAS 4.0 has many shortcomings compared to a newer version, MAS 5.0. Those shortcomings include negative expression measures and large variability for low expressed genes. We have compared the existing Harvard data (processed) with our MAS 5.0 generated data using PCA (principal component analysis) projection. Our MAS 5.0 data had better separation of disease groups compared to existing processed data (Figure 1). More clear separation of normal lungs and adenocarcinomas from the rest can be seen for MAS 5.0. It appears that generating new expression measures from raw CEL files considerably enhanced the accuracy of the data.

Figure 1. Harvard data generated by MAS 4.0 (top) and MAS 5.0 (bottom).

Total number of samples 203 Tumors (186): AD, CO, SM, SQ

- AD** = lung adenocarcinomas (127)
and other adenocarcinomas (12)
- CO** = pulmonary carcinoids (20)
- SM** = SCLC cases (6)
- SQ** = squamous cell carcinoma (21)
- NL** = normal lung (17)

Although probe set names were unique at this point, different probe sets could in some cases correspond to the same gene. When this is true, we averaged expression levels across the probe sets. The two data sets were then merged by gene name, and we selected those genes common to both the Harvard and Michigan data sets.

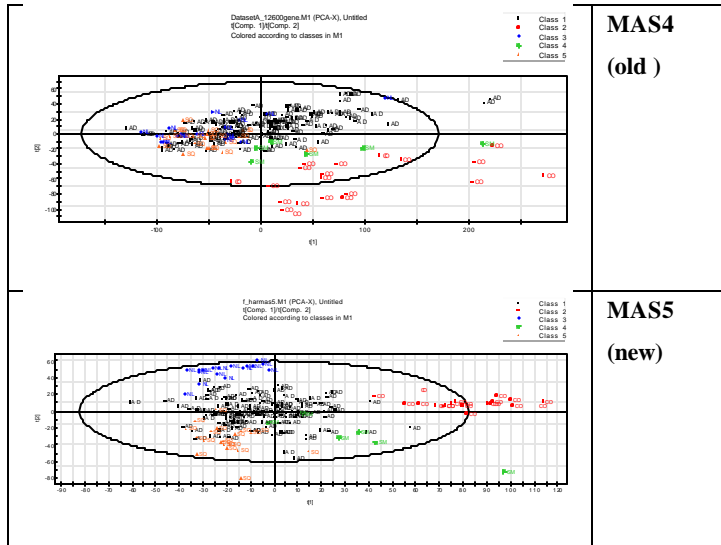
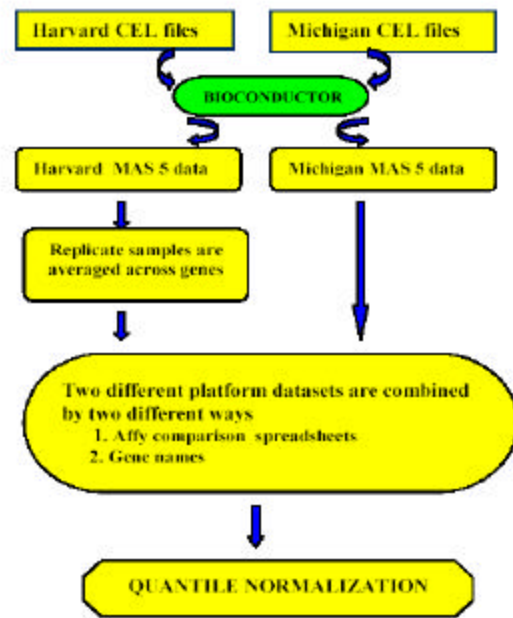


Figure 2. Flow chart for preprocessing of the data



1.2 Data Integration

CEL files, containing probe level data, were available for the Harvard and Michigan data sets. Thus we started with these raw data to integrate the two different platforms and merged them into a larger data set. Figure 2 shows the flow chart of processing from CEL files to final integrated data.

MAS5 expression level data were created by using BioConductor packages (www.bioconductor.org) for summarizing probe level data from each of the Harvard and Michigan data sets. Fifty-one sample pairs in the Harvard data were replicates. For such cases, we averaged the data across the replicates, and as a result, 254 samples were reduced to 203 samples. However none of the 96 samples in Michigan data set were replicates.

To integrate the MAS5 data from Harvard and Michigan, we used two different merging methods. Note that one-to-one matching was not possible because they used different chips. The first approach was matching the probe set ID from Harvard data (U95a) with the probe set ID from Michigan data (HuGene FL) by the Array comparison spreadsheets (ACS) (www.Affymetrics.com). We then selected those probe sets that were common to both the Harvard and Michigan data.

The second approach we use is to integrate the data sets using gene names. The housekeeping genes are not used in this approach. The gene names are obtained from Affymetrix' website.

1.3 Quantile Normalization of Combined Data from Different Platforms

In the final part of the integration, we have used a quantile normalization (Q-normalization) for making the different platforms data comparable. We obtained the same distribution of intensities across platforms for each gene by Q-normalization. The algorithm for Q-normalization is given in the following box.

Algorithm:

- a. Denote $X=(X^1, \dots, X^k)$ of dimension $p \times N$ where X^1, \dots, X^k are data from platform 1 to platform k respectively and $N=N_1+\dots+N_k$;
- b. Rank each row of X^1, \dots, X^k to give $X^1_{rank}, \dots, X^k_{rank}$
- c. Calculate $p^m(I,J)=(X^m_{rank}(I,J)-1)/(N_m-1)$ where $I=1,\dots,p$ and $J=1,\dots,N_m$ for each platform $m=1,\dots,k$
- d. The normalized value of $X^m(I,J)$ is the average of

1.4. Integrated Data

There are 203 samples (127 lung adenocarcinomas and 12 other adenocarcinomas, 20 pulmonary carcinoids , 6 SCLC cases, 21 squamous cell carcinoma and 17 normal samples) and 12600 probe sets in Harvard data set. For Michigan data, there are 96 samples (86 lung adenocarcinomas and 10 normal samples) and 7129 probe sets. The integrated dataset combined by using ACS had 6041 probe sets compared to dataset combined by using gene names, which had 4837 genes. As mentioned in preprocessing step, different probe sets could in some cases correspond to the same gene.

We performed PCA to see the difference between two integrated data sets. Both ways of combining the datasets resulted in approximately the same information. Hence, the integrated data set using gene names will be used for the following analysis.

2. NORMALIZATION ISSUE

Without proper normalization, the integrated Harvard and Michigan data were shown to be completely separated (Figure 3), which means the distributions of the two data sets are not comparable. Figure 4 shows that Q-normalization made the Harvard and Michigan samples cover the approximately the same projected space by PCA. Figure 5 shows the scatter plot of the Harvard vs. Michigan data before (top two plots) and after (bottom two plots) Q-normalization. Each point in the scatter plot corresponds to a gene. The left two plots are for normal samples while the right two plots are for cancer samples. The scatter plots confirm that the Q-normalization has made the two data sets comparable with approximately same distribution.

Figure 3. PCA plot of combined data before Q-normalization
(red : Harvard data black : Michigan data)

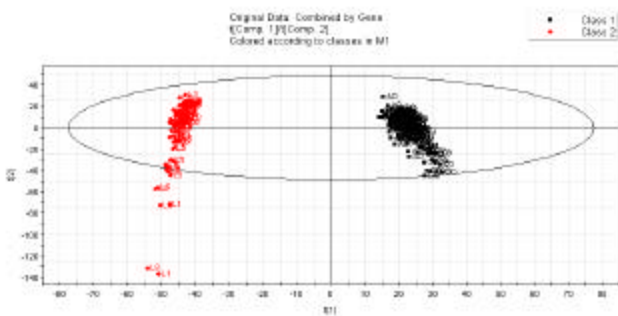


Figure 4. PCA plot of combined data after Q-normalization
(red : Harvard data black : Michigan data)

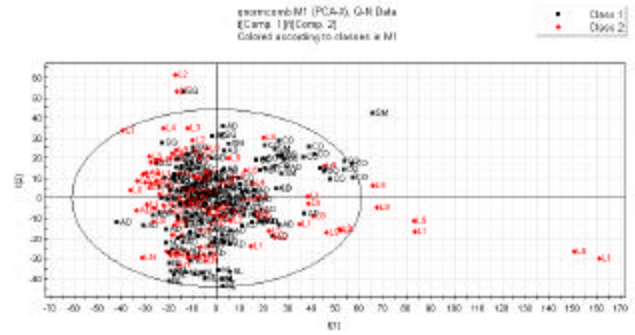
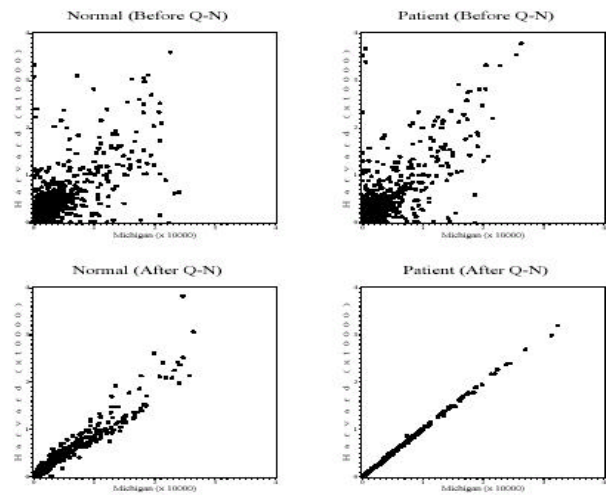


Figure 5. Harvard data vs. Michigan data
by normal (left) and patient (right) group
also by before (top) and after (bottom) Q-normalization



3. DISCRIMINATION OF NORMAL LUNGS FROM THE CARCINOMAS

One interesting property of the final data (MAS 5.0, integrated and normalized) is the clear separation of normal lung samples from the rest of the carcinomas. The PCA scores plot in the Figure 6 below shows the separation of normal lungs (red) from the adenocarcinomas samples (black). One projection method of supervised learning is partial least squares (PLS) and its related discriminant analysis (PLS-DA). Figure 7 shows the projection of PLS-DA results. Again normal lung samples have clearly separated themselves from the rest. Using PLS-DA, we can select genes that are responsible for the discrimination of the two classes. List of selected genes was not included because of space limitation.

Figure 6. PCA plot after Q-normalization
Class 1 = normal, Class 2 = carcinomas

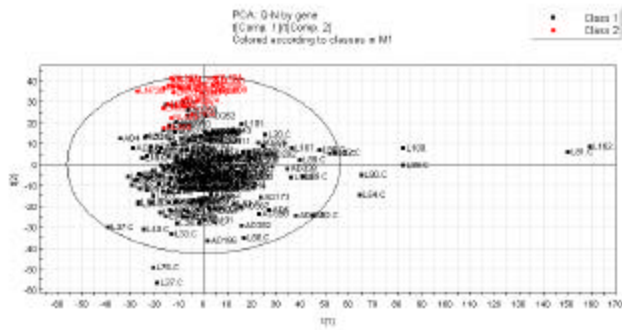
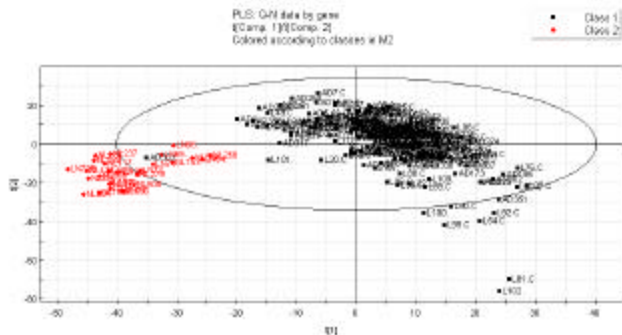


Figure 7. PLS plot after Q-normalization

Class 1 = normal, Class 2 = carcinomas



4. PREDICTION OF ONE PLATFORM FROM ANOTHER

Further validation of integrated data can be done by building predictive models using one of the four lung cancer dataset and then validating those models with one or more of the remaining data sets. Specifically our objective was to build a predictive model to classify lung tissue samples into two groups; adenocarcinomas (AD) and normal lung (NL) using data from the Harvard dataset and then validate this model by classifying new cases from the Michigan dataset.

We took one of the merged data sets generated in a previous step which contains 4837 common gene names from the Harvard and Michigan studies. We deleted all the observations from tumors other than AD and then created a new binary column (our dependent variable) named AD indicating whether a tissue sample histologically corresponds to adenocarcinoma (1 = AD, 0 = NL). Then we split the dataset into a training data containing 156 samples (139 AD and 17 NL) from the Harvard study and a test data containing 96 samples (86 AD and 10 NL) from the Michigan study.

High dimensionality of our dataset (4837 independent variables) makes it very difficult to handle for most classification algorithms. Therefore we have applied a feature selection algorithm based on CHAID (Chi-squared Automatic Interaction Detector) to the training data to initially select the 50 best predictors. The training data was then used again with these selected features to build predictive models using the three well-known classification tools: CART, C5 and neural networks (NN). We have applied these three prediction models to classify the 96 samples in the test data. The performance of these models is summarized in Table 1.

All the models showed very high sensitivity (98.84 -100%) but variable specificity (80 – 90%) when classifying new cases. The best performing model was NN with 100% sensitivity and 90% specificity, for an overall accuracy of 98.96%. Both classification tree models (CART and C5) obtained similar results with 98.84% sensitivity and 80% specificity for an overall accuracy of 96.88%.

Table 1: Summary of performance statistics for the 3 predictive classification models on the test data

	C5	CART	NN
Sensitivity	98.84%	98.84%	100.00%
Specificity	80.00%	80.00%	90.00%
PPV	97.70%	97.70%	98.85%
NPV	88.89%	88.89%	100.00%
Accuracy	96.88%	96.88%	98.96%

5. SURVIVAL ANALYSIS

Here we consider the time to death as the dependent variable for prediction and use the final model to identify those genes associated with high risk of mortality. We have used total of 211 patients (125 from Harvard data and 86 from Michigan data) that have both lung adenocarcinoma cancer and survival information.

Since the samples came from two totally independent studies, some study specific factors (known or unknown) might contribute to the risk of mortality. Consequently we need to consider the effect of different studies in the model when we examine the gene effects. For each gene, we use frailty (mixed effects) Cox proportional hazard model with gene as fixed effect and study effect (Harvard vs. Michigan) as random. We list the genes with significant FDR (False Discovery Rate) adjusted P-value (at 0.05 level) in Table 2.

Table 2: Significant mortality gene list from frailty (mixed effect) Cox Model

GENE	Coefficient	RAW P-Value	FDR adjusted P-Value
KIAA0211	-0.0025069	0.0000011	0.0054398
CTSL	0.0002727	0.0000368	0.0312787
KRT18	0.0001400	0.0000490	0.0312787
LHX1	0.0019983	0.0000362	0.0312787
PGK1	0.0001655	0.0000426	0.0312787
PRKCBP1	0.0034964	0.0000275	0.0312787
STX1A	0.0009447	0.0000517	0.0312787
VEGFC	0.0026009	0.0000309	0.0312787
P4HA1	0.0010053	0.0000646	0.0347284
INHA	0.0009011	0.0001035	0.0484484
RALA	0.0025610	0.0001102	0.0484484

Next, we used tree method for survival data (RPART function in R/Spplus) with the significant genes (at raw p-value 0.05 level) as predictors (480 genes) from the Cox proportional hazard model above. The number of samples and the predicted relative event rate (RR, time-adjusted and relative to the whole data set) defined by each node are shown in the Figure 8. Based on the recursive partitioning result, we grouped the samples into two risk groups: high risk (pink) and low risk (green). The Kaplan-Meier plot by risk group is shown in Figure 9. We can see that the mortality behavior is quite different between these two groups.

Figure 8: Survival tree for relative event rate

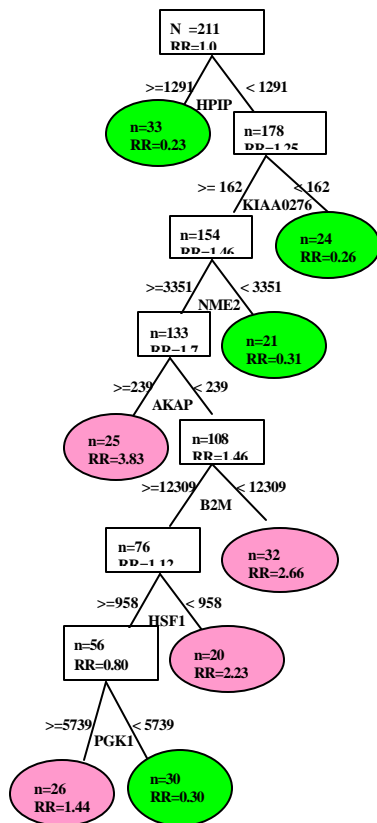
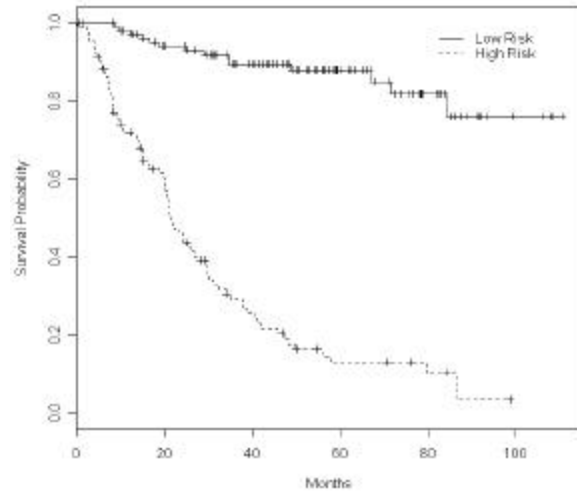


Figure 9: K-M plot by high risk and low risk group defined



6. DISCUSSION AND CONCLUSION

Combining data sets from several studies would provide more samples and give more statistical power in the analysis. However, due to difference in the design of probe sets for different Affymetrix chips, useful information may be lost when we use the combined data sets to do the analysis. For example, only 6041 probe sets from Harvard data were kept in the integrated data. About half of the total probe sets in Harvard data was not used. One possible way to maximize the information would be to integrate the results from combined data with the result from individual data. We are planning to look at individual data in the future and compare the results with the published information.

We have used some statistical and data mining methods for survival data to analyze the integrated data set. The result showed the possibility of using integrated gene expressions data to classify the adenocarcinoma samples into different mortality risk groups. Investigation of the results obtained from this survival analysis may be of some biological interest and need further consideration.

7. ACKNOWLEDGEMENTS

Our thanks to Phil Burstein and Alan Menius for their encouragement and support for doing research on the integration and analysis of biological data.

REFERENCES

- [1] Array comparison spreadsheets
www.affymetrix.com/support/technical/manual/
- [2] Beer, David *et al.*: Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine* vol 8. no 8. (August 2002), 816-824
- [3] Bhattacharjee, Arindam *et al.*: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas subclasses. *PNAS*, vol 98. no 28. (November 2001), 13790-13795
- [4] Bioconductor: open source software for bioinformatics
www.bioconductor.org/
- [5] Bolstad, Ben. Probe level quantile normalization of high density oligonucleotide array data. Technical report of university of Berkeley, division of Biostatistics , 2001
- [6] Therneau TM and Atkinson EJ : An Introduction to Recursive Partitioning using the RPART routines, Technical Report #61, Division of Biostatistics, Mayo Clinic , 1997
- [7] Terry M. Therneau and Patricia M. Grambsch (2000), *Modeling Survival Data*, Springer, New York