

Application of Survival and Multivariate Methods to Gene Expression Data Combined from Two Sources

Linda Robb
GlaxoSmithKline
Statistical Sciences UK
Medicines Research Centre
Gunnels Wood Road, Stevenage
00 44 1438 764905
linda.c.warnock@gsk.com

Richard Stephens
GlaxoSmithKline
Transcriptome Analysis UK
Medicines Research Centre
Gunnels Wood Roa, Stevenage
0044 1438 768038
richard.j.stephens@gsk.com

JoAnn Coleman
GlaxoSmithKline
Statistical Sciences US
Upper Merion
1 610 270 5660
joann.n.coleman@gsk.com

ABSTRACT

Keywords

Principal Component Analysis, Cox Proportional Hazards Model, Survival, Meta p-value. False Discovery Rate

1. INTRODUCTION

The use of gene expression has the potential to have a large impact in the area of Oncology. If genes can be identified which are associated with prolonged life then these genes can be used as to aid decisions regarding patient treatment and care. The genes could also be used as targets for new cancer treatments in the area of drug research. Gene expression data and clinical information have been collected from two experiments to investigate the effect of gene expression on patients with lung cancer. Data collected from Harvard and Michigan have used different Affymetrix chip types and have different clinical parameters measured. This analysis combines the information across datasets and focuses on lung adenocarcinomas tumors to identify genes which have an association with survival.

2. PRE-PROCESSING OF THE DATA

The data for Harvard and Michigan were pre-processed and normalised independently. Two hundred and three .CEL files from Harvard and ninety six .CEL files from Michigan were processed in MAS5 and DChip software. The QC process involved identification of chip to chip variation using DChip algorithms. Any chip with 'probe set outlier %' >3 (DChip) were discarded and 3'/5' ratios >3 (MAS5) were also removed. Metrics were also collected from house-keeper genes - Bactin and Gapdh and also from background noise and overall chip intensity. These metrics were included in a PCA analysis with the aim of identifying poor quality chips. Using these quality control criteria thirteen chips were removed from the Harvard dataset and sixteen chips from the Michigan dataset. Through PCA analysis of the QC data an experimental bias in the Harvard data was discovered to be due to an outlying batch of IVT. This batch was also observed to create bias in the expression data.

The data from each source were normalised in DChip using the piece-wise linear normalisation algorithm on the Perfect Match (PM) data only. The data could not be combined at this stage due to the different chip types.

The Affymetrix website provides comparison spreadsheets which allows probe sets to be matched from different chip types. The HuGeneFL_to_U95_comp.xls spreadsheet¹⁰ was used to select probe_sets from the two datasets which had an old to new sequence relationship. This matching resulted in over 6000 probe sets being defined as common between the two datasets. This method of matching is more precise than using gene names.

3. METHODS

The analysis is performed separately on each dataset, 114 Harvard adenocarcinoma samples and 70 Michigan adenocarcinoma samples. Each sample was represented on circa 6000 genes which were identified as common between the two datasets. Principal Component Analysis (PCA) was used to explore the data prior to survival analysis. Survival analysis was performed on the clinical data to assess the effect of the clinical parameters on survival. Cox Proportional Hazards regression was performed on each gene including covariates which could affect survival such as tumor stage, age and sex in addition to gene expression. A forward selection process was used to select variables for inclusion in the model. Fisher's meta analysis approach⁴ was used to combine p-values from the two analyses and to define a new unique p-value for every gene. A false discovery rate adjustment was used to adjust the p-values for the genes.

An alternative analysis approach was to use the tumor stage as a surrogate marker for survival as patients with a stage I tumor were more likely to survive than patients with more advanced tumors. This approach simplified the analysis by allowing log gene expression to be used as a response with tumor stage as an explanatory variable. Fisher's meta analysis was also used with this approach.

The genes which had a significant effect on survival were looked at in more detail by using gene ontology.

4. RESULTS

4.1 Exploratory analysis of expression data

The quality control metrics identified the samples with IVT batch 3 from Harvard data as having low background and low average signal in relation to all the other samples. A PCA analysis shows that this bias also affects the gene expression from Harvard Adenocarcinoma samples.

Preliminary exploration of all the data from lung adenocarcinoma samples was performed using PCA. This unsupervised analysis identified data source (Harvard or Michigan) as the largest source of variation with the first principal component accounting for 89% of the variation with an eigen-value = 163 (Figure 1). The loadings plot (Figure 2) shows that there is a shift in the average expression from one dataset to the next across the majority of genes. Summary statistics show the geometric mean expression for Harvard to be 2.4 with a standard deviation of 0.50 and for Michigan to be 3.1 with a standard deviation of 0.38.

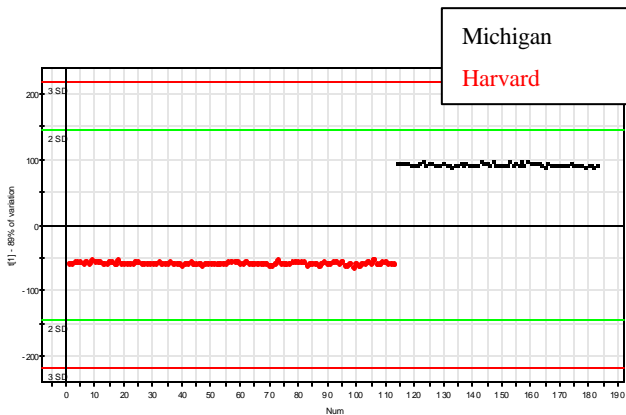


Figure 1: PCA scores plot showing the separation between Harvard and Michigan gene expression data before gene by gene normalisation. The first component is shown on the y-axis and accounts for 89% of the variation

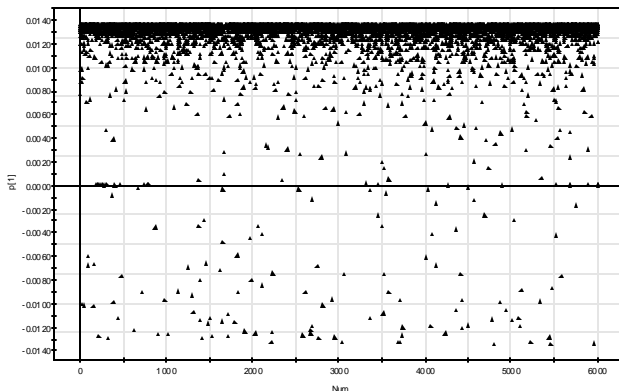
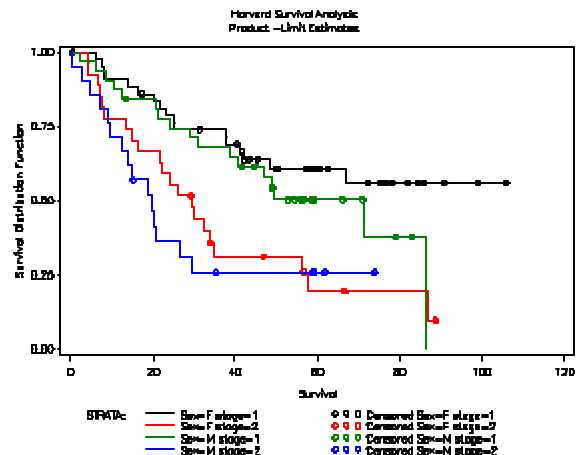
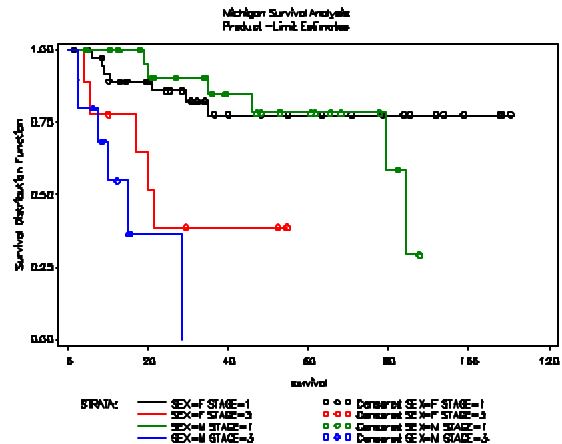


Figure 2: Loadings plot showing that the majority of genes are more highly expressed for one

For the purposes of visualisation this difference was adjusted for by doing a global normalisation which centered the expression of each gene from each data source around zero. After this gene by gene normalisation the first component accounted for 12% of variation (eigen-value = 22). There are several reasons for this separation which are discussed in section 5.

4.2 Exploratory analysis of clinical data

There are many factors which are more likely to affect survival than gene expression. Tumor stage was incomplete in the Harvard dataset and so classification rules⁶ were used to classify each sample as either stage I tumor or stage II+. The stage II+ category included patients from stage II to stage IV (metastasis). The factors of tumor stage, age and sex were investigated in a survival model. Figure 3a and 3b show that stage of tumor has a large effect on survival rates (Harvard $p < 0.0001$, Michigan $p < 0.0001$). There is slight evidence of a difference between sex (Harvard $p = 0.2015$, Michigan $p = 0.1623$) and slight evidence of an effect of age (Harvard $p = 0.382$, Michigan $p = 0.0904$).



Figures 3a and 3b: Kaplan-Meier plots of the effect of tumor stage and sex on survival. Stage I tumors have greater survival prospect than stage II+. Survival also appears to be better for the Michigan patients. This may be due to the fact that Harvard

4.3 Survival Analysis

Cox Proportional Hazards model was performed on every gene to determine the association between gene expression and survival. The model also included the covariates tumor stage, sex and age. The effect of tumor stage was more significant than most of the genes and so was usually entered first into the model via forward selection. Thus if gene was entered into the model after stage then the effect on survival should be interpreted as having an effect in addition to that of the tumor stage. An overall p-value for every gene was calculated using Fisher's meta analysis. The hazard parameter estimates for gene expression were used to assess whether the gene expression was positively or negatively correlated with hazard. Any genes which had a positive association in one dataset but negative in the other were discarded. 241 genes were selected with a meta p-value ≤ 0.05 (2 genes had an adjusted p-value ≤ 0.05).

A volcano plot³ (Figure 4) was used to visualise the data. Minus log₁₀ of the meta p-value was plotted against the parameter estimates from both Harvard and Michigan results. The exponential of the parameter estimate can be interpreted as the increase in hazard for every unit increase in log gene expression or rather the increase in hazard for every 10 fold change in gene expression.

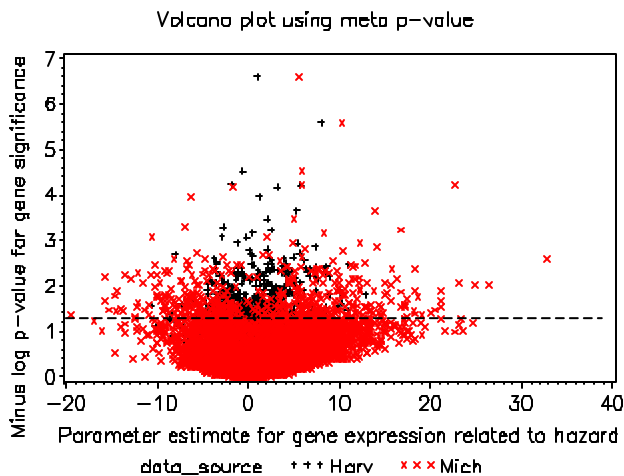


Figure 4: The gene expression hazard estimate is plotted against minus log₁₀ of the meta p-value. Genes with a minus log p-value greater than 1.3 are significant at the 5% level

4.4 Secondary Analysis

It is known from survival analysis on the clinical data that tumor stage has a large impact on survival as patients with a more advanced stage of tumor will live a shorter length of time. Tumor stage could be thought of as a surrogate marker for survival. The problem of identifying genes associated with survival is then simplified to finding genes which show significant differences between patients with different stages of tumor. This problem can be addressed through a basic Analysis of Variance with log gene expression as the response and tumor, sex and age as explanatory variables.

Fisher's meta analysis is again used to combine the p-values from the two datasets. A False Discovery Rate adjustment is made to the meta p-values. The following figure plots the difference in predicted geometric means between stage I and stage II+ tumors against minus log of the meta p-value.

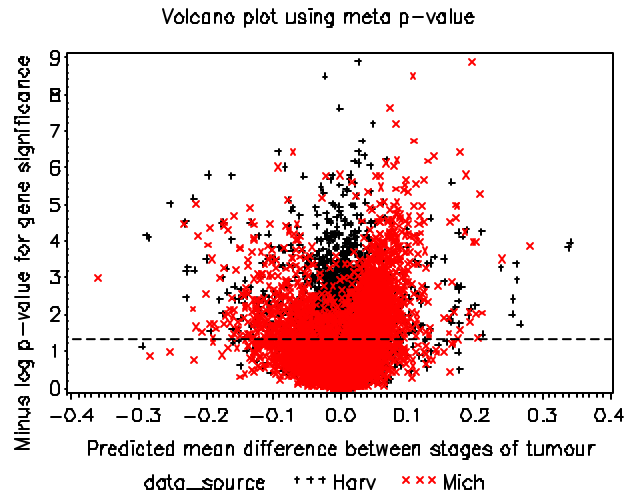


Figure 5: The difference in predicted geometric means is plotted against minus log₁₀ of the meta p-value. Genes with a minus log p-value greater than 1.3 are significant at the 5% level

It is interesting to note that there are very few genes with a two-fold (± 0.3) change or greater. A filter is placed on the genes so that only genes with a 1.5 fold-change and a meta p-value of ≤ 0.05 are taken forward for further work. This resulted in 43 genes being selected with $p \leq 0.05$ (33 genes had an adjusted p-value ≤ 0.05).

5. DISCUSSION

This analysis has identified a number of genes as being associated with survival, either through survival analysis or by using tumor stage as a surrogate for survival. The survival analysis identifies genes which are positively or negatively associated with survival. These genes in addition to tumor stage could be used to predict length of patient survival and hence help decisions on cancer treatment for the individual. The secondary analysis identifies

genes which show differences in gene expression between stage I and more advanced stages. These genes could be used as markers for new drug treatments. If it is known that a gene is active in more advanced tumors then drugs could be designed that target that gene so that under treatment a patient will never progress further than stage I tumor. The genes identified from the primary analysis have been quoted in other literature. The most significant gene is adrenomedullin which has been found before to be an important tumor survival factor in human carcinogenesis¹⁰.

A follow up experiment like real time PCR would help to validate these results. It is also important to validate the direction of the association to confirm whether the gene is positively or negatively associated with survival.

The task of combining data across chip types and different data sources is challenging. The PCA plot (figure 1) demonstrates this. There are a number of reasons for this clear separation. It could be due to differences in how the samples were treated and processed within the two sites. It could be due to the fact that the two datasets were normalised within DChip separately or it could be due to the differences between the probe sets across the Harvard and Michigan datasets which are targeting the same sequence.

The clinical information collected was sparse especially for the Michigan data. There are so many factors that could affect survival more than gene expression. Factors such as treatment, time of prognosis, data of operation, economic status, race and many others could all be important and yet have not been accounted for in the analysis.

6. REFERENCES

1. Beer, D.G., *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* **8**, 816-824 (2002)
2. Bhattacharjee, A., *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS* **98**, 13790-13795 (2001)
3. Wolfinger, R.D. *et al.*, *J. Computational Biol.*, **8** 625-638 (2001)
4. Fisher, R.A. *Statistical Methods for Research Workers* (4th Edition). London: Oliver and Boyd, 1932
5. Rhodes, D. R., Barrette, T.R., Rubin, M.A., Ghosh, D., Chinnaiyan, A.M. Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostrate Cancer. *Cancer Research* **62**, 4427-4433, (2002)
6. AJCC Cancer Staging Handbook, Sixth Edition, 191-203 (2002)
7. Hedges, L.V., Olkin, I. *Statistical Methods for Meta-Analysis*, Academic Press, 1985
8. Speed, T. *Statistical Analysis of Gene Expression Microarray Data*, Chapman and Hall/CRC, 2003
9. Allison, P.D., *Survival Analysis Using the SAS System: A Practical Guide*, Cary, NC: SAS Institute Inc., 1995
10. https://www.affymetrix.com/support/technical/comparison_s_preadsheets.affx
11. Adrenomedullin functions as an important tumor survival factor in human carcinogenesis. *Microsc Res Tech.* **57(2)**, 110-9 (2002)
12. Cheng Li and Wing Hung Wong Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proc. Natl. Acad. Sci.* **98**, 31-36.d, (2001a)