Two case studies in Genomics

WORKSHOP ON HUMAN GENOMIC VARIATION AND DISEASE: Norway, August 2008

Robert Tibshirani
Stanford University

The Tibshirani "lab"

- 4 grad students; work closely with Trevor Hastie and his students
- statistical and data mining techniques for genomic and proteomic data
- writing and supporting software packages for some of these tools - Excel Add-ins and R language

Examples of our work

- SAM- Significance analysis of microarrays
- PAM- Prediction analysis of microarrays- for sample classification
- CGH-miner- for fiding "hot spots" in CGH data
- Superpc- Supervised principal components- survival analysis from genomic/proteomic data
- Currently- gene set analysis, complementary clustering

Outline

- two recent controversies in Cancer genomics: a gene expression study of lymphoma and a GWA of cancer
- some general suggestions for improving the state of statistical analyses in these areas

A recent experience

- Dave et al published a high-profile study in NEJM, reporting that they had found two sets of genes whose expression were highly predictive of survival in patients with Follicular Lyphoma.
- the paper got a lot of attention at the recent ASH meeting, because the genes in the clusters were largely expressed in non-tumor cells, suggesting that the host-response was the important factor
- One of my medical collaborators- Ron Levy, asked me to look over their paper- he wanted to apply their model to the Stanford FL patient population.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 18, 2004 VOL. 351 NO. 21

Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells

Sandeep S. Dave, M.D., George Wright, Ph.D., Bruce Tan, M.D., Andreas Rosenwald, M.D., Randy D. Gascoyne, M.D., Wing C. Chan, M.D., Richard I. Fisher, M.D., Rita M. Braziel, M.D., Lisa M. Rimsza, M.D., Thomas M. Grogan, M.D., Thomas P. Miller, M.D., Michael LeBlanc, Ph.D., Timothy C. Greiner, M.D., Dennis D. Weisenburger, M.D., James C. Lynch, Ph.D., Julie Vose, M.D., James O. Armitage, M.D., Erlend B. Smeland, M.D., Ph.D., Stein Kvaloy, M.D., Ph.D., Harald Holte, M.D., Ph.D., Jan Delabie, M.D., Ph.D., Joseph M. Connors, M.D., Peter M. Lansdorp, M.D., Ph.D., Qin Ouyang, Ph.D., T. Andrew Lister, M.D., Andrew J. Davies, M.D., Andrew J. Norton, M.D., H. Konrad Muller-Hermelink, M.D., German Ott, M.D., Elias Campo, M.D., Emilio Montserrat, M.D., Wyndham H. Wilson, M.D., Ph.D., Elaine S. Jaffe, M.D., Richard Simon, Ph.D., Liming Yang, Ph.D., John Powell, M.S., Hong Zhao, M.S., Neta Goldschmidt, M.D., Michael Chiorazzi, B.A., and Louis M. Staudt, M.D., Ph.D.

ABSTRACT

BACKGROUND

Patients with follicular lymphoma may survive for periods of less than 1 year to more than 20 years after diagnosis. We used gene-expression profiles of tumor-biopsy specimens obtained at diagnosis to develop a molecular predictor of the length of survival.

METHODS

Gene-expression profiling was performed on 191 biopsy specimens obtained from patients with untreated follicular lymphoma. Supervised methods were used to discover expression patterns associated with the length of survival in a training set of 95 specimens. A molecular predictor of survival was constructed from these genes and validated in an independent test set of 96 specimens.

RESULTS

Individual genes that predicted the length of survival were grouped into gene-expression signatures on the basis of their expression in the training set, and two such signatures were used to construct a survival predictor. The two signatures allowed patients with specimens in the test set to be divided into four quartiles with widely disparate median lengths of survival (13.6, 11.1, 10.8, and 3.9 years), independently of clinical prognostic variables. Flow cytometry showed that these signatures reflected gene expression by nonmalignant tumor-infiltrating immune cells.

CONCLUSIONS

The length of survival among patients with follicular lymphoma correlates with the molecular features of nonmalignant immune cells present in the tumor at diagnosis.

From National Cancer Institute (S.S.D., G.W., B.T., A.R., W.H.W., E.S.J., R.S., H.Z., N.G., M.C., L.M.S.); Center for Information Technology (L.Y., J.P.); and National Heart, Lung, and Blood Institute (S.S.D.) — all in Bethesda, Md.; British Columbia Cancer Center, Vancouver, Canada (R.D.G., J.M.C., P.M.L., Q.O.); University of Nebraska Medical Center, Omaha (W.C.C., T.C.G., D.D.W., J.C.L., J.V., J.O.A.); Southwest Oncology Group, San Antonio, Tex. (R.I.F., T.M.G., T.P.M., M.L.); University of Rochester School of Medicine, Rochester, N.Y. (R.I.F.); Oregon Health and Science University, Portland (R.M.B.); University of Arizona Cancer Center, Tucson (L.M.R., T.M.G., T.P.M.); Fred Hutchinson Cancer Research Center, Seattle (M.L.); Norwegian Radium Hospital, Oslo (E.B.S., S.K., H.H., J.D.); Cancer Research UK, St. Bartholomew's Hospital, London (T.A.L., A.J.D., A.J.N.); University of Würzburg, Würzburg, Germany (A.R., H.K.M.-H., G.O.); and University of Barcelona, Barcelona, Spain (E.C., E.M.). Address reprint requests to Dr. Staudt at the National Cancer Institute, Bldg. 10, Rm. 4N114, NIH, Bethesda, MD 20892, or at lstaudt@mail.nih.gov.

N Engl J Med 2004;351:2159-69.

Summary of their findings

- They started with the expression of approximately 49,000 genes measured on 189 patient samples, derived from DNA microarrays. A survival time (possibly censored) was available for each patient
- they randomly split the data into a training set of 89 patients and a test set of 90 patients
- using a multi-step procedure (described below), they extracted two clusters of genes, called IR1 (immune response 1) and IR2 (immune response 2).
- They averaged the gene expression of the genes in each cluster, to create two "super-genes".

... continued

- They then fit these super-genes together in a Cox model for survival, and applied it to the training and test sets. The p-value in the training set was $< 10e-7$ and 0.003 in the test set. IR1 correlates with good prognosis; IR2 with poor prognosis
- In the remainder of the paper they interpret the genes in their model

What happened next...

- I downloaded the data
- Applied some familiar statistical tools- eg SAM (Significance analysis of microarrays), less familiar ones- supervised principal components, and also gave the data to Brad Efron. Our initial finding- no significant correlation between gene expression and survival.
- I spent 2-3 weeks emailing back and forth with their statistician (George Wright) and programming in R, to recreate their analysis
- Confession- it was fun being a "forensic statistician"

Details of their analysis

1) Divide the data randomly into training and test sets of approximately equal numbers of patients. Apply the following recipe [steps 2–6] to the training set.

2) Choose all genes with univariate Cox score > 1.5 in absolute value. This reduced the number of genes from roughly 49,000 to roughly 3,000, with about a 50-50 split between good prognosis genes (negative scores) and poor prognosis genes (positive scores).

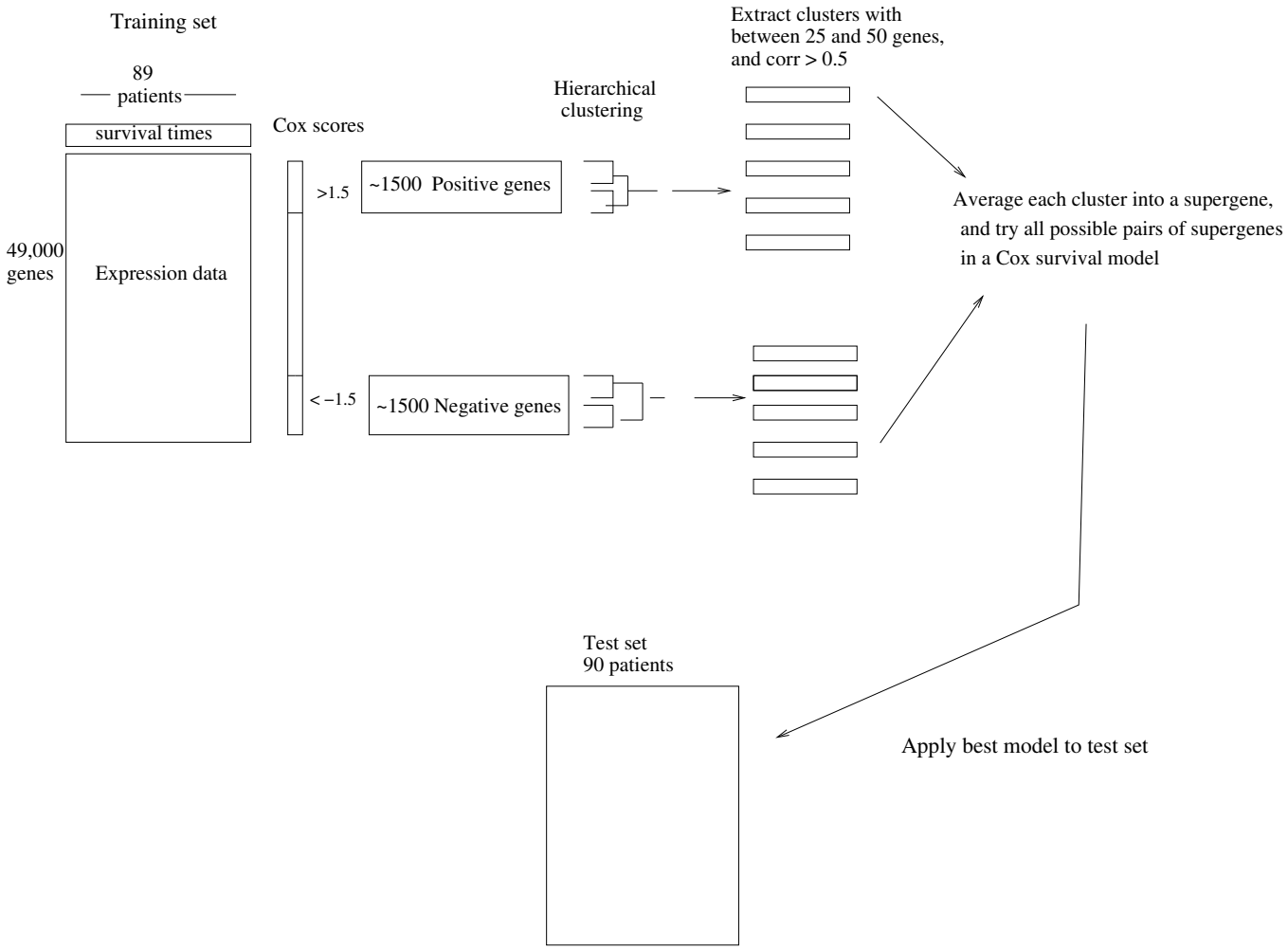
3) Do separate hierarchical clusterings (correlation metric, average linkage) of the good and poor prognosis genes.

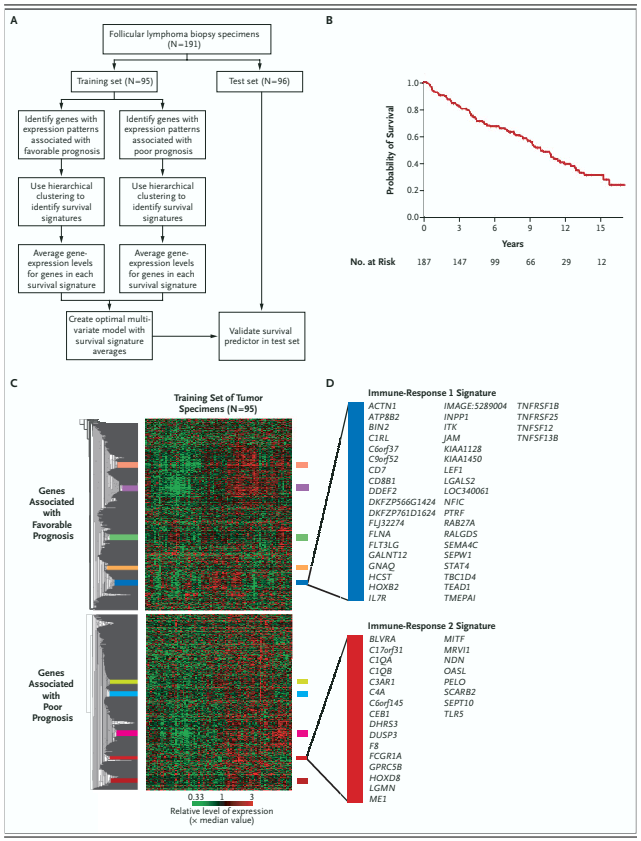
Details continued...

4) Find all clusters in the dendrograms (clustering trees) containing between 25 and 50 genes, with internal correlation at least 0.5. Represent each cluster by the average expression of all genes in the cluster– a "supergene" Try every pair of supergenes as predictors in Cox models for predicting survival.

5) Choose the most significant pair from this process. The authors call the resulting pair of clusters IR1 (good prognosis) and IR2 (poor prognosis).

6) Finally use the model (IR1, IR2) in a Cox model to predict survival in the test set.


Training set
89
patients
survival times
49,000
genes
Expression data
Cox scores
>1.5
~1500 Positive genes
< −1.5
~1500 Negative genes
Hierarchical
clustering
Extract clusters with
between 25 and 50 genes,
and corr > 0.5
Average each cluster into a supergene,
and try all possible pairs of supergenes
in a Cox survival model
Test set
90 patients
Apply best model to test set



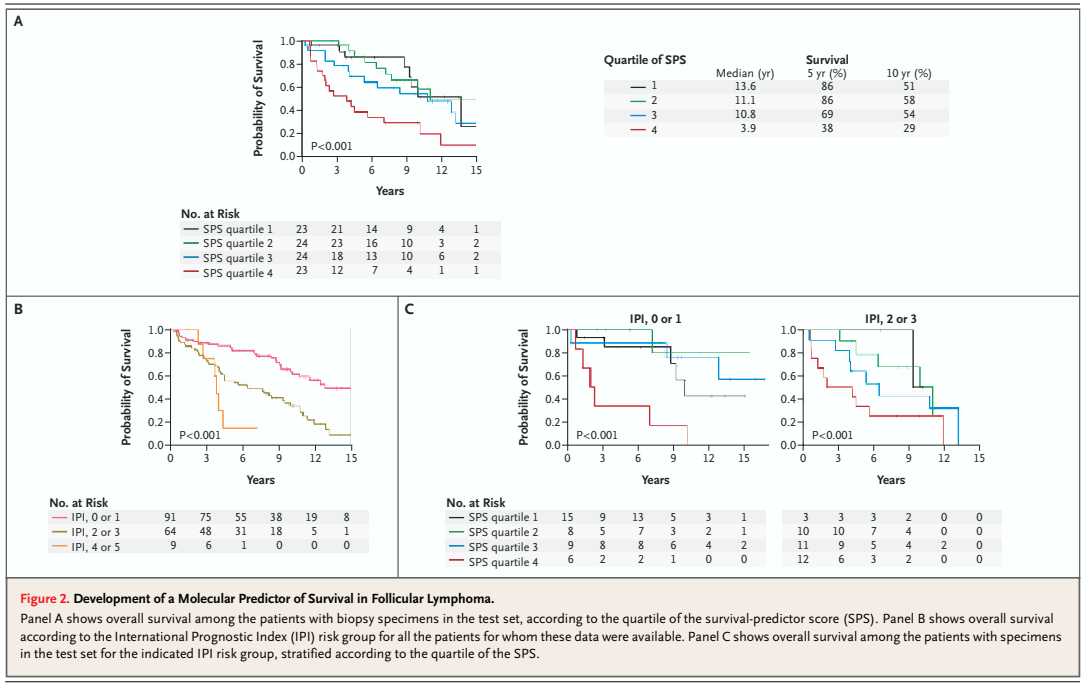
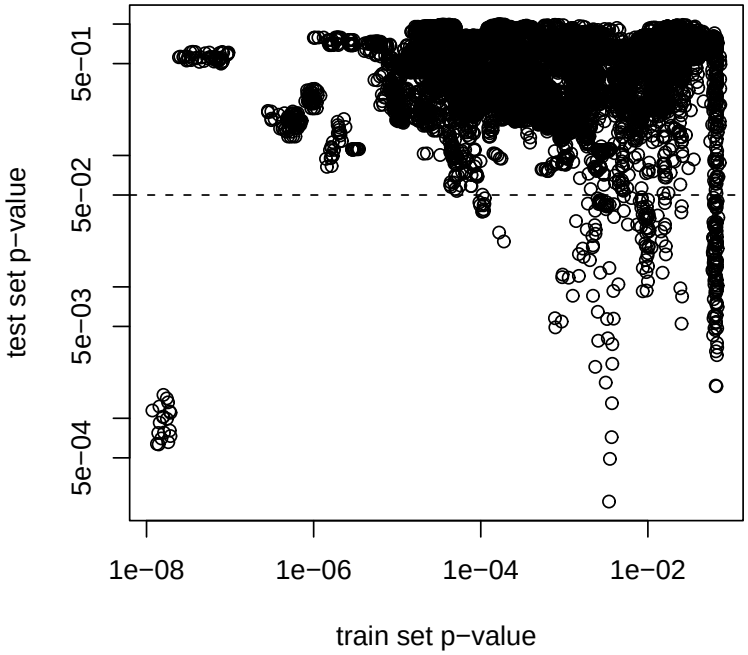


Figure 2. Development of a Molecular Predictor of Survival in Follicular Lymphoma.

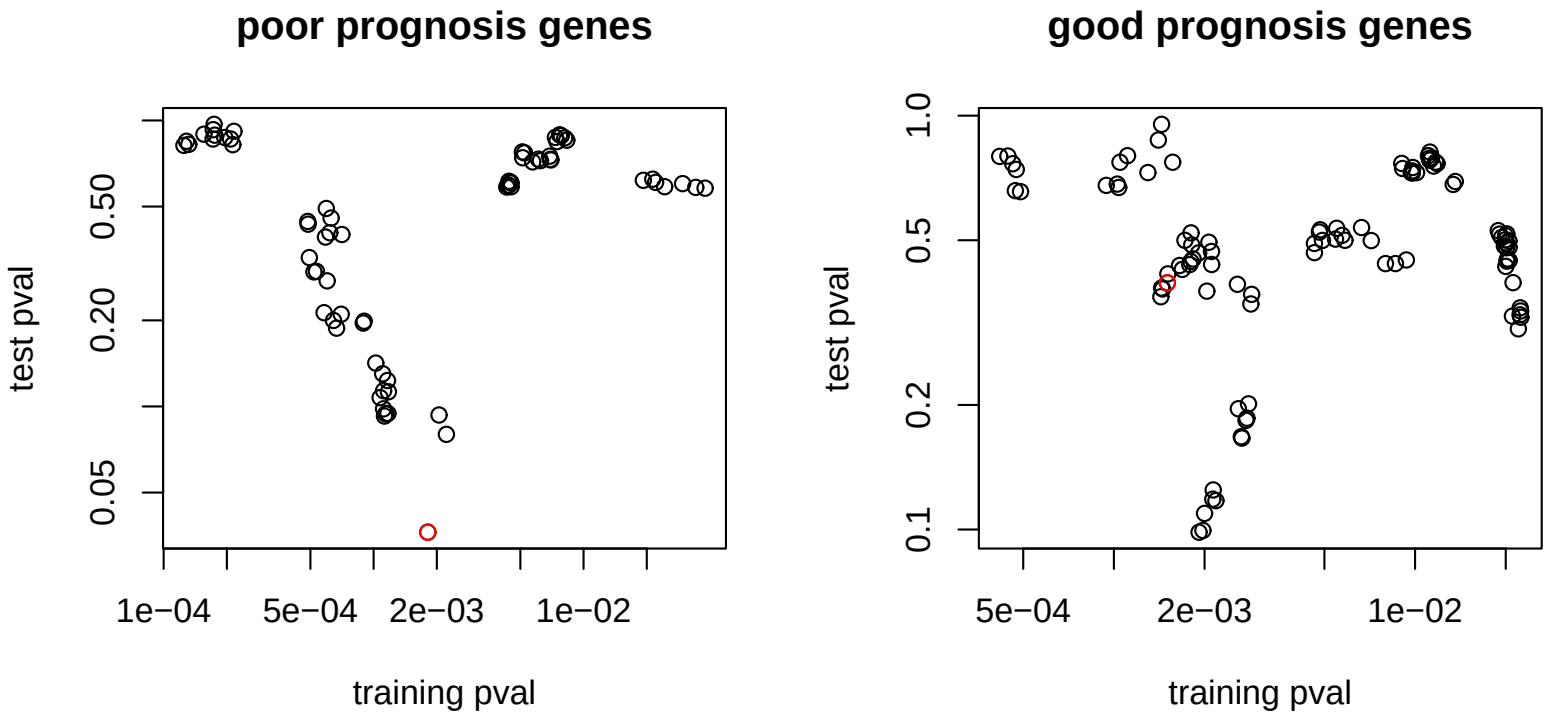
Panel A shows overall survival among the patients with biopsy specimens in the test set, according to the quartile of the survival-predictor score (SPS). Panel B shows overall survival according to the International Prognostic Index (IPI) risk group for all the patients for whom these data were available. Panel C shows overall survival among the patients with specimens in the test set for the indicated IPI risk group, stratified according to the quartile of the SPS.

p-values of all cluster pairs

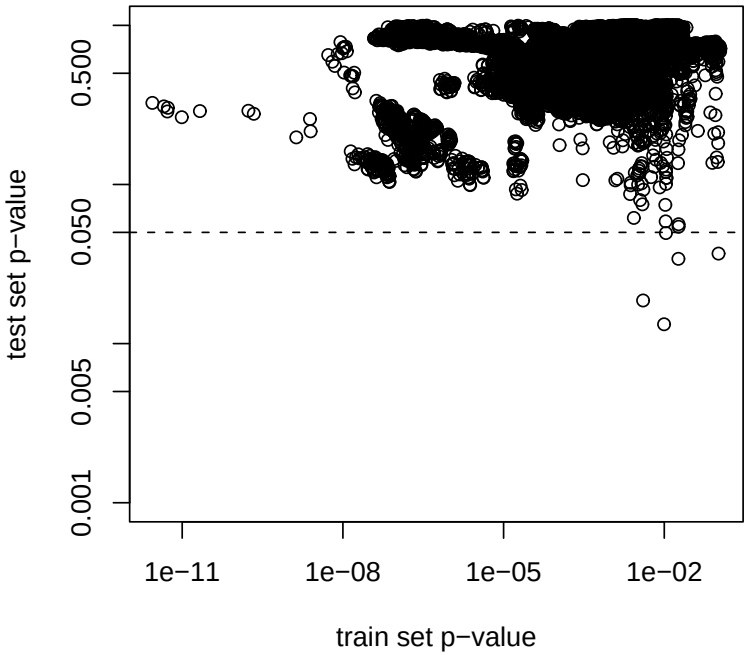


The total number of points (cluster pairs) with test set p-values less than 0.05 (239) is far fewer than we'd expect to see by chance (735)

Univariate p-values

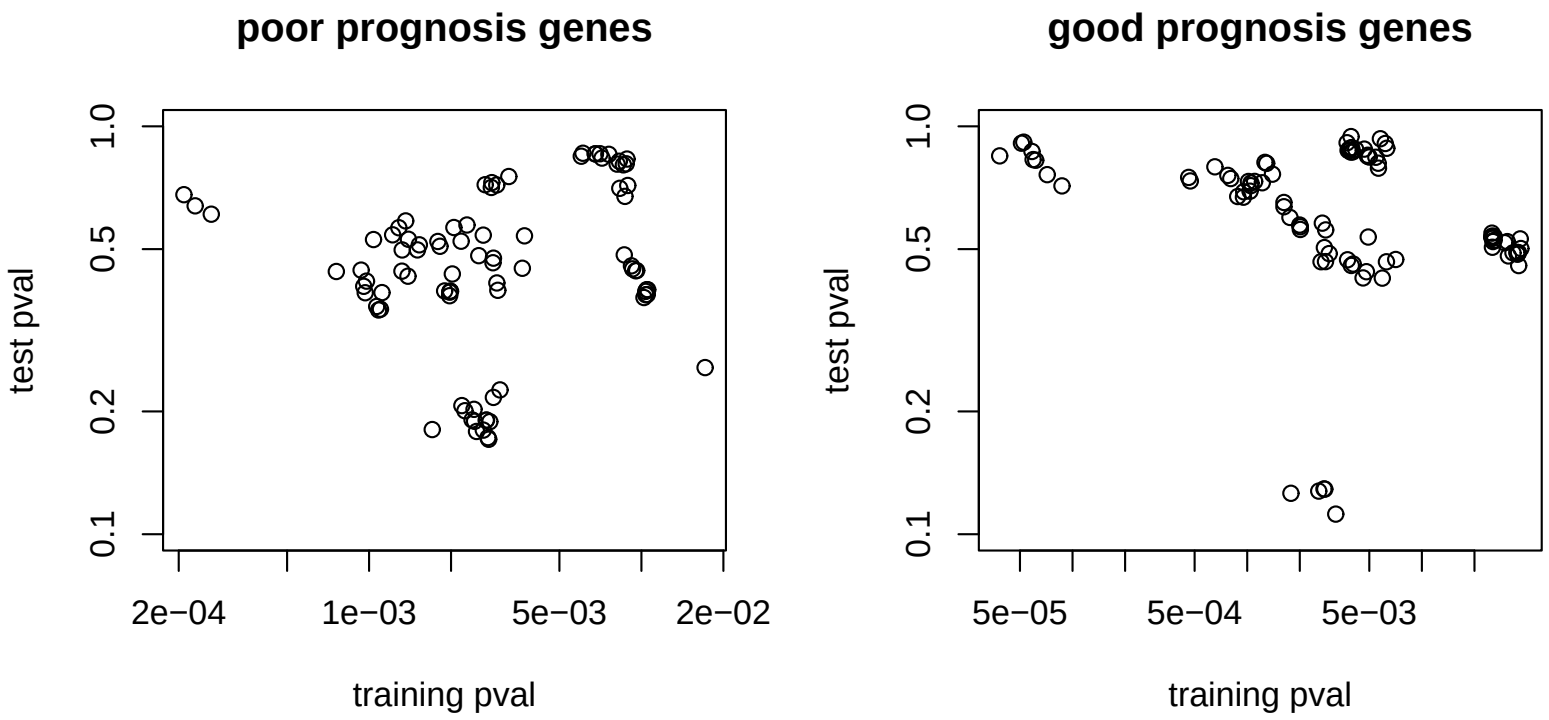


Swapping train and test sets

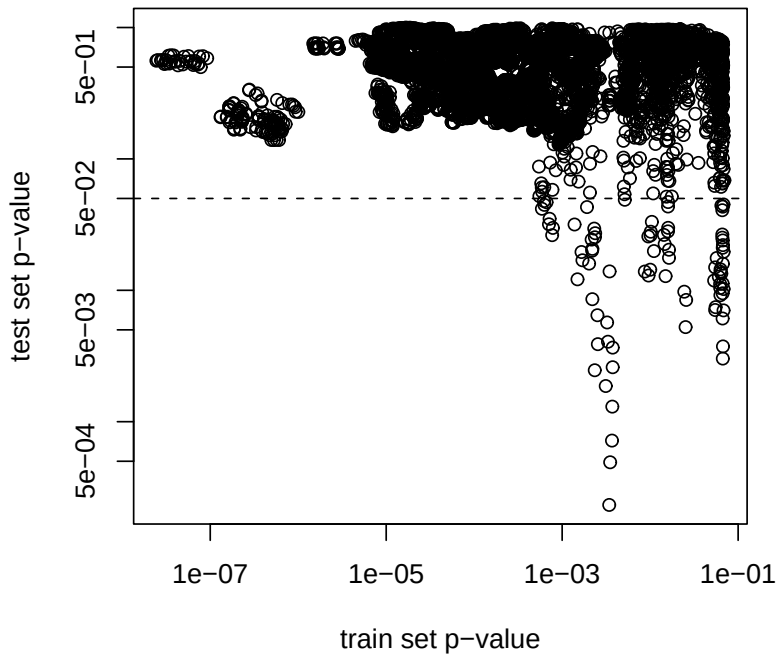


There are only 85 pairs out of 11628 that are significant in the test set at the 0.05 level, while we would expect 11628.05=581 pairs just by chance.*

Swapping train and test sets, ctd...



Cluster size ranges (30,60) rather than (25,50)



The aftermath

- I published a short letter to NEJM in March 2005; full details of my re-analysis appear on my website
- The authors published a rebuttal in the same issue.

  "Nothing in Tibshirani's analysis calls into dispute the fact that we discovered and validated a strong association between gene expression in follicular lymphoma and overall survival. "

  Their arguments:
  - (1) we followed standard statistical procedures, found a small p-value on the test set, therefore our finding is correct;
  - (2) our method found an interaction, which SAM can't find
  - (3) we get small p-values if we apply our model to random halves of the data (????!!!!!!)

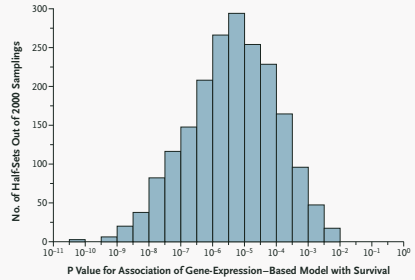


Figure 1. Results of Samplings of Half-Sets and Association between Gene Expression and Survival.

Our predictor could not have been discovered with the SAM method, which relies solely on univariate associations with survival. Rather, our predictor derives its strength from the synergistic combination of two gene-expression signatures in a multivariate model. Tibshirani confuses the ability of our method to discover a survival association with the fact that we actually found one that validated the association. When he exchanged the training set for the test set, he was unable to rediscover our gene-expression predictor because some genes in our predictor fell below the P value threshold for association with survival in the test set. This does not negate the fact that our model is highly associated with survival in the test set.

Hong et al. have made three errors. First, in our sorted subpopulations, the CD19– fraction contained, on average, 12.6 percent contamination with follicular-lymphoma cells, not 25 percent, as they claim. To believe that the higher expression of the immune-response signatures in the CD19– fraction is due to this 12.6 percent contamination requires that the lymphoma cells in the CD19– fraction have expression of the immune-response signatures that was more than eight times as high as that in their counterparts in the CD19+ fraction. Second, Hong et al. incorrectly discount the immune-response 1 signature, which contributes significantly to the survival model in the test set ($P<0.001$). Third, many of the immune-response signature genes are selectively expressed in T cells, monocytes, or dendritic cells, or in more than one of these, but not in B cells, making the contention of Hong et al. even more implausible.

Louis M. Staudt, M.D., Ph.D.
George Wright, Ph.D.
Sandeep Dave, M.D.
National Cancer Institute
Bethesda, MD 20892
lstaudt@mail.nih.gov

1. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309-14.

When Doctors Go to War

TO THE EDITOR: Like Bloche and Marks in their Perspective article on doctors in combat (Jan. 6 issue),[1] the American Medical Association (AMA) applauds the outstanding work of military physicians in treating wounded soldiers under extremely challenging circumstances.[2] Physicians are defined by their common calling to prevent harm and treat people who are ill or injured and by their universal commitment to uphold recognized principles of medical ethics whenever patients rely on their

General comments

- Their finding is *fragile*. I don't believe that it is real or reproducible
- This experience uncovers a problem that is of general importance to our field:
  - with many predictors, it is too easy to overfit the data and find spurious results
  - we can inadvertently mislead the reader, and mislead ourselves. I have been guilty of this too

Some recommendations

- encourage authors to publish not only the raw data, but a script of their analysis
- encourage authors to use "canned" methods/packages, with built-in cross-validation to validate the model search process - see "supervised principal components", Bair et al 2005