

TOWARDS A GENETIC SIGNATURE OF LYMPH NODE POSITIVE BREAST CANCER

Paul Hergenroeder^{1,2}, David Peters³, Dan Handley³, David Dabbs⁴, James Lyons-Weiler¹, Soumyaroop Bhattacharya¹, Adam Brufsky^{1,2,4}
¹UPCI(University of Pittsburgh Cancer Institute) – ²Division of Medical Oncology,³University of Pittsburgh Graduate School of Public Health,
⁴Magee-Womens' Hospital – UPMC(University of Pittsburgh Medical Center)

ABSTRACT

Gene expression profiling has the potential to classify primary breast tumors into prognostic groups. A 70-gene prognosis profile (NEJM 2002; 347: 1999) was recently developed by a supervised classification of gene expression data from young women with lymph node negative breast cancer, and was used to stratify a larger group of 295 women into good and poor prognosis groups independent of lymph node status. Lymph node status, however, remains the best clinical predictor of patient outcome. We believe that a prognostic profile of breast cancer based on a gene expression profile unique to lymph node positive breast cancer would have clinical utility. In addition, prior work with gene prognosis profiling, including the above study, has been performed on a set of primary tumors heterogeneous for hormone receptor and Her2 Neu status. To address these issues, we have developed a gene expression profile of ER (-), PR (-), Her2 (-), lymph node positive breast cancer. RNA was extracted from 5 LN(+) and 5 LN (-) primary breast tumors and gene expression was analyzed in duplicate by Affymetrix U133A microarrays containing 22,283 probes. **Relative gene expression analysis produced a preliminary expression profile of 18 up-regulated genes (>1.5 fold, p< 0.01 by t-test) and 40 down-regulated genes (>1.5 fold, p<0.01) unique to lymph node positive breast cancer. Up-regulated genes and ESTs include those related to IFNGR1, RAD23A, and a MAFF related protein. Down-regulated genes and ESTs include those related to Enolase 3, TFDP2, ARD1 and Cyclin E2. Data from a larger validation sample set will be presented. This profile differs significantly from poor prognostic profiles found in hormone receptor mixed sample sets. ** Supported by grants from the Komen Foundation and the US Army Department of Defense.

Results displayed in poster differ from those of abstract

BACKGROUND

Published reports have identified predictive gene expression classifiers that predict lymph node spread¹ and distant relapse^{1,2,3} by correlation of primary tumors to discovered predictive gene expression classifiers. These reports have used training sets of samples heterogeneous for ER/PR expression. They have also used distant relapse rates to assign primary tumors to good versus bad prognosis groups, despite marked variation in patient treatment between the time of excision of the tested sample and noted relapse

HYPOTHESIS

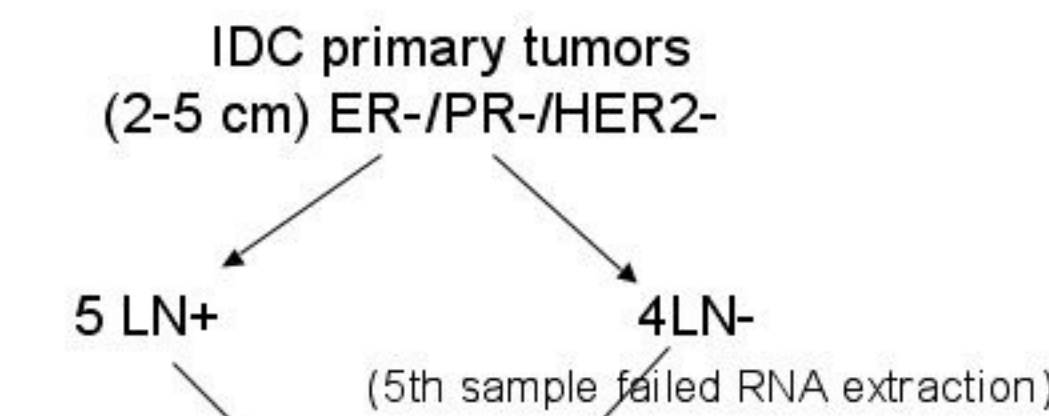
➤A predictive classifier of genetic dysregulation indicative of the potential for lymph node spread exists in an IDC primary tumor

➤Restricting the phenotype variation i.e. ER/PR/HER2neu expression status, of test samples will allow for detection of a predictive gene expression classifier with a smaller number of samples

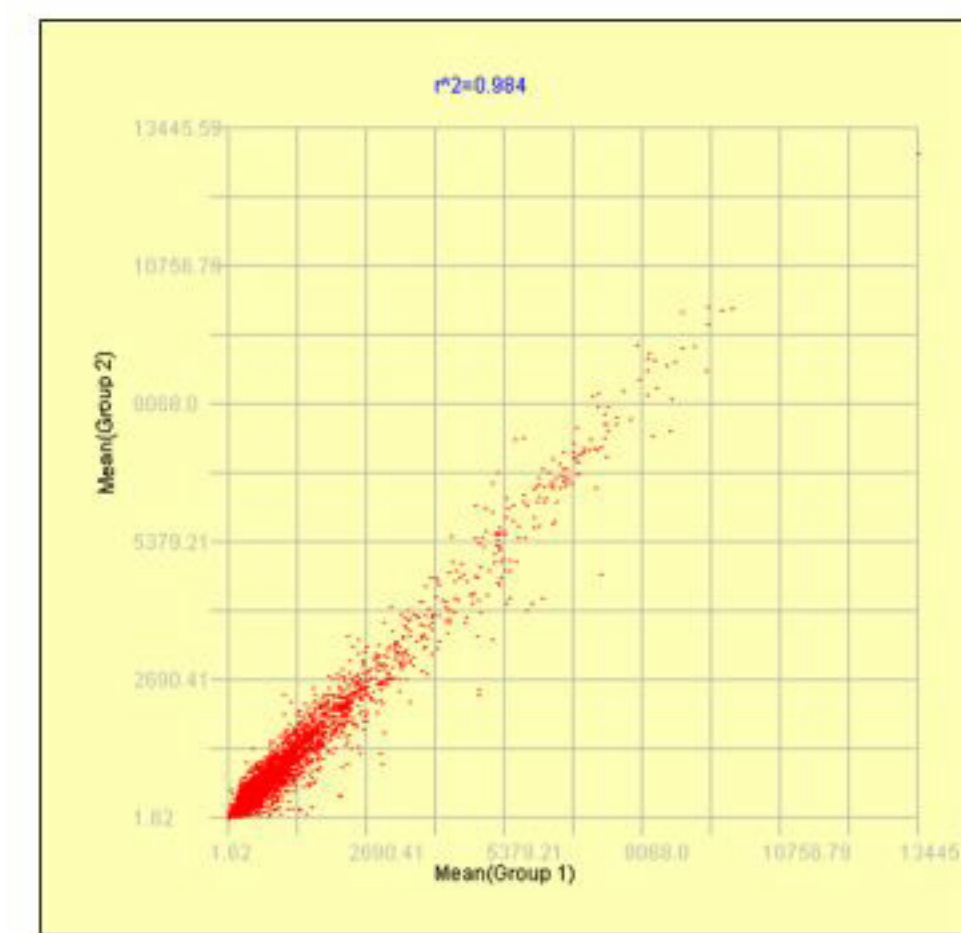
MICROARRAY SIGNAL ANALYSIS

- Perfect Match(PM) signal averaged across duplicate hybridization/scanning
- Assessment of data correlation coefficient
- Data analyzed using MDSS^{4,6} (maximum difference subset) analysis algorithm using pooled variance t-test statistic
- Structured agglomerative hierarchical clustering using Euclidian distance
- Clustering validation by leave one out validation(LOOV) technique
- Bipartition validation by parametric bootstrapping technique using 100 iterations

STUDY DESIGN



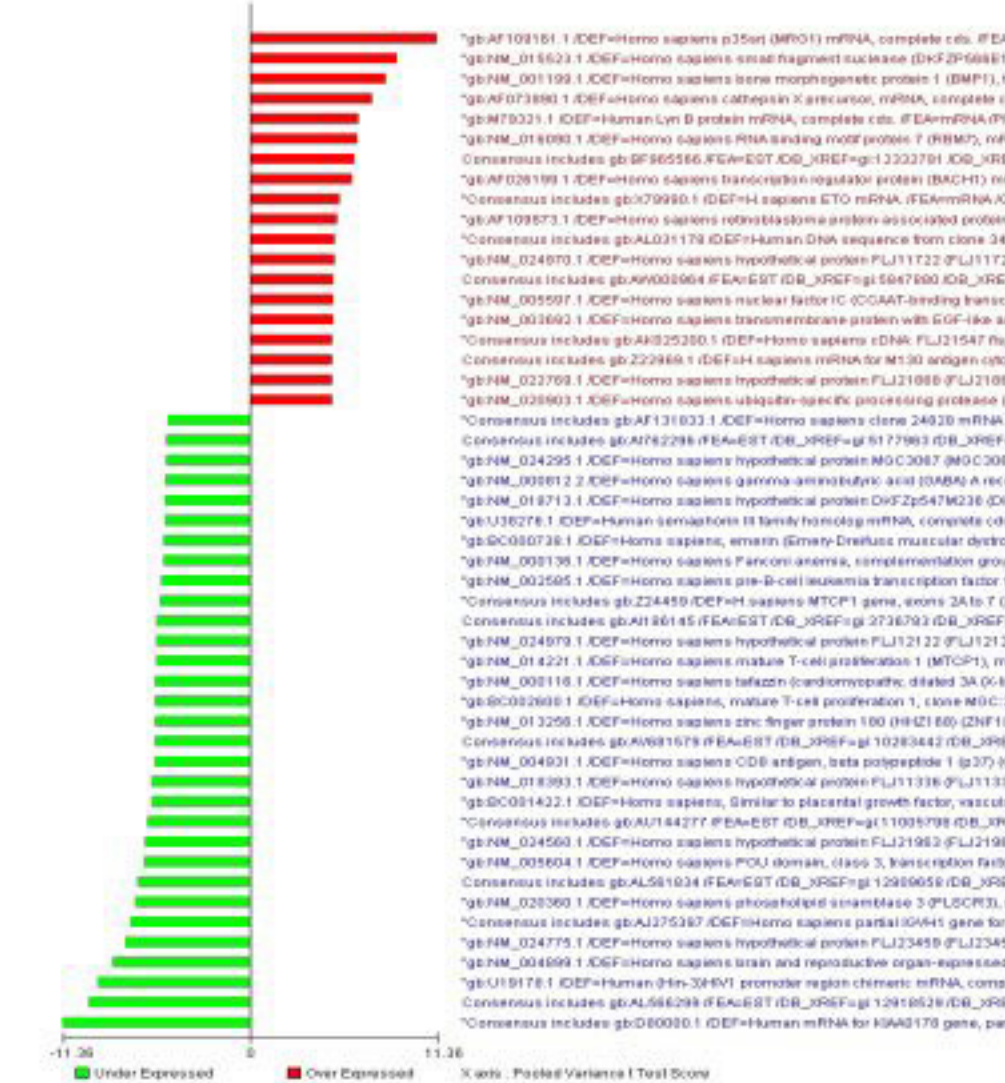
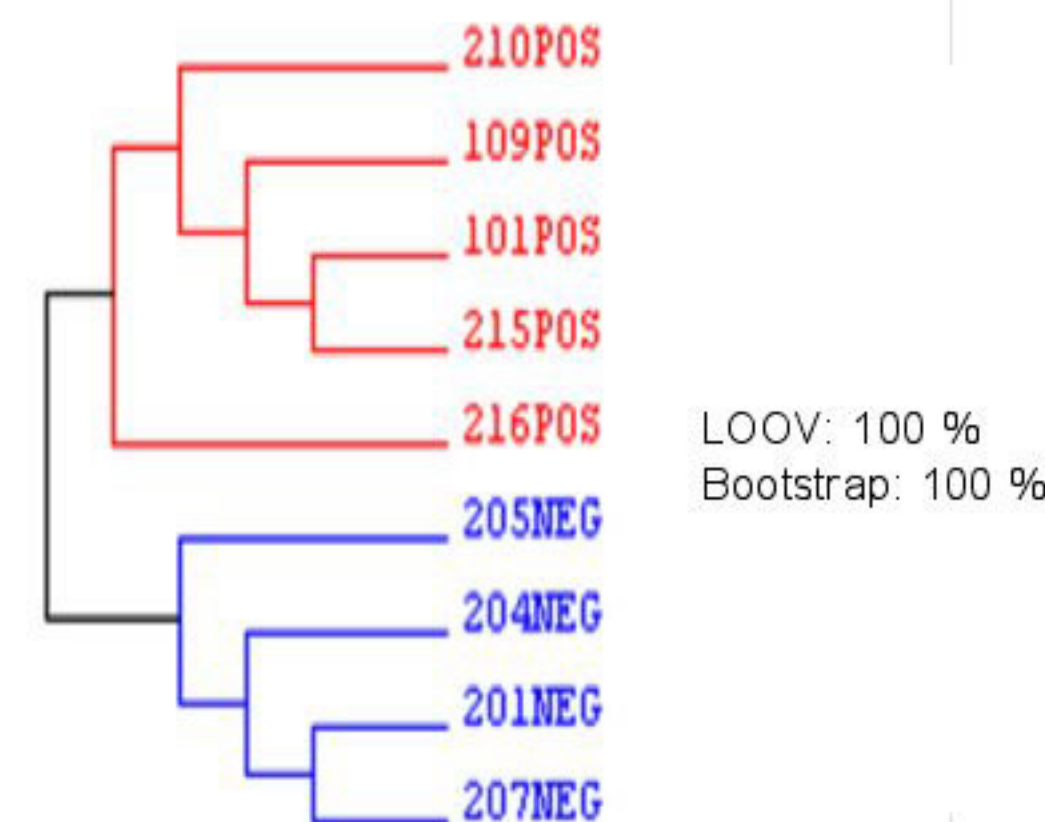
RESULTS (figures)



R² = 0.984
cov = 0.032

Pooled Variance t-test

•646 retained genes (α = 0.05)



Histogram of 50 top dysregulated genes between LN+ and LN- in classifier of 261 genes generated by applying MDSS^{4,6} algorithm to above 646 gene classifier

RESULTS (text)

The data set demonstrated a high correlation coefficient between subgroups and low among-array COV(coefficient of variation). Pooled variance t-test ($\alpha=0.05$) analysis of average PM values across duplicates yielded a lymph node positive predictive classifier of 646 genes. The classifier demonstrates strong internal validation with a LOOV score of 100% and nonparametric bootstrap analysis score of 100%. **In the bootstrap analysis, sample labels are randomized and the entire procedure is repeated. A 100% score strongly supports the lack of overtraining in group membership assignment.** MDSS^{4,6} (maximum difference subset) analysis algorithm using more restrictive t-test alpha thresholds was performed. Using ($\alpha=0.001$), generating a classifier of 22 genes, the predictive the LOOV score of 100% persisted

DISCUSSION

These results suggest the discovery of an internally validated gene expression classifier correlating with associated LN status of 2-5cm IDC ER-/PR-/HER2- primary tumors. This predictive classifier was generated from gene expression profiles of only nine samples; 5 LN+ and 4 LN-.

In an earlier report on the exploration for a predictive classifier of lymph node spread⁵, a 100 gene classifier generated by gene expression profiles of 10 LN+ and 22 LN-, IDC primary tumors, revealed marked uncertainty on LOOV cross validation. These primary tumors were heterogeneous for ER/PR status.

These results are of an internally validated predictive classifier of a test or training set and are exploratory until confirmed by predictive correlation with a larger validation set.

The discovery of this strongly internally validated classifier by this small number of samples is partly due to duplicate hybridization and scanning, offsetting background variability. However we believe it is to a greater degree due to limiting the biological heterogeneity of the samples studied.

Incremental discoveries of elements of predictive classifiers of tumorigenesis may be facilitated by minimizing phenotypic differences between studied sample groups.

Conclusions

➤Exploratory results suggest that IDC primary tumors can be correctly classified as to their association with lymph node spread by genetic expression analysis.

➤Restricting the phenotype variation, i.e. ER/PR/HER2neu expression status of test samples allows for discovery of a predictive gene expression classifier using a small number of samples.

NEXT STEPS

➤Testing this discovered classifier for its ability to predict known lymph node status in a larger validation set of samples.

➤ Comparison of the microarray expression data with SAGE (Serial Analysis Gene Expression) libraries being processed on the same samples.

REFERENCES

- ¹Huang E, Cheng SH, Pittman J, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003;361:1590-96.
- ²van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene- Expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-36.
- ³van de Vijver MJ, He YD, van't Veer LJ, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. NEJM 2002;347:1999-09
- ⁴Lyons-Weiler J, Patel S, Bhattacharya S. A Classification-Based Machine Learning Approach for the Analysis of Genome-Wide Expression Data. Genome Research 2003; 13:503-12
[analysis tools available at: <http://bioinformatics.upmc.edu/GEDA.htm/>]
- ⁵West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS 2001;98:11462-68
- ⁶MDSS Algorithm(p510 of reference 4)
The goal of MDSS is to find genes expression differences that are not only significant, but also carry information useful to the classification applications. Steps in the general form:
1. Calculate a test(e.g., test statistic or S/N ratio) for each gene(spot) in a comparison of sample groups A and B(e.g. LN+ vs. LN- primary tumors)
2. Rank the genes in descending order according to the magnitude of that measure
3. Find the largest threshold value of that measure for which a classification in which the sample groups are clearly delineated (T_{θ}). For example, if the test employed in Step 1 is a t-test, find the largest significance level (α) where the classifier succeeds in discriminating between the two sample groups. This gene set is called the 'initial MDS'
4. Jackknife out individual samples and store the list of genes that are significant beyond threshold T_{θ} . Adjust the degrees of freedom due to the exclusion of one sample as needed. Each time a sample is removed, an individual MDSS list is created.
5. Identify the subset of genes common to all individual MDSSs. This gene set is comprised of genes that are not only significantly different between the two; they also pass, as a set, the criterion of predictive utility. This gene set is called the 'overall MDSS'
6. Verify that the overall MDSS returns the expected classification (e.g., a hierarchical cluster diagram with a bipartition between sample groups A and B). If this does not occur, adjust (increase or decrease) the threshold value set at Step 3 and begin again.