

# On The Optimal Number Of Gene Expression Markers

## For Tissue of Origin Cancer Diagnostics

Ljubomir J. Buturovic

Pathwork Diagnostics, Inc., Sunnyvale, CA



### BACKGROUND

#### Optimal Number of Markers for Multiplex Genomic Diagnostic Tests

Multiplex genomic tests based on gene expression combine multiple markers using computer algorithms to generate clinically useful information.

The number of markers used in the gene expression tests has implications for the choice of gene expression platform.

A large number of markers may favor microarrays, smaller numbers may favor other platforms such as quantitative PCR.

Theory [1] provides lower bound on the number of markers (at least  $c-1$ , where  $c$  is the number of test categories). Also, it suggests that the required number of markers increases with the number of categories.

We undertook to determine the optimal number of markers for a particular diagnostic problem: cancer classification. An example of a test designed to solve the problem is the Pathwork Tissue of Origin Test (TOO).

### OBJECTIVES

#### Determine optimal number of markers for the Pathwork™ Tissue of Origin (TOO) test

The Pathwork Tissue of Origin Test is designed to identify the tissue of origin of metastatic and poorly differentiated cancers by comparing the expression signature of the test biopsy sample with expression patterns of 15 common cancer types.

TOO uses a machine-learning algorithm (model) to identify the tissue of origin. The models considered here take expression signatures over the selected markers as input, and produce predicted tissue identity as output.

The goal was to determine the minimum number of markers required for the optimal performance of the TOO test.

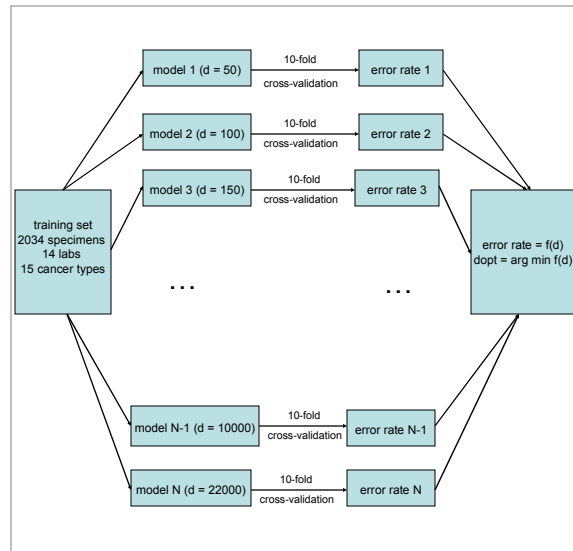
### METHODS

#### Build and compare optimal TOO models with different number of markers

Tissue of Origin models were built using 2034 training samples (gene expression profiles) and varying the number of markers. The data originated from 14 laboratories and 15 cancer types, on Affymetrix GeneChip Human Genome U133A and U133 Plus 2.0 platforms.

For a given number of markers,  $d$ , we chose the best  $d$  markers, and built the optimal model using gene expression profiles of the best markers.  $d$  varied from 50 to 10,000 in the steps of 50, with an additional model with 22,000 markers.

For each model, 10-fold cross-validation was used to estimate the TOO performance. The performance criterion was percent agreement between predicted Tissue of Origin and clinical truth (or, alternatively, the prediction error rate). The combination of machine learning algorithm and cross-validation was designed to reduce overfitting.



### RESULTS

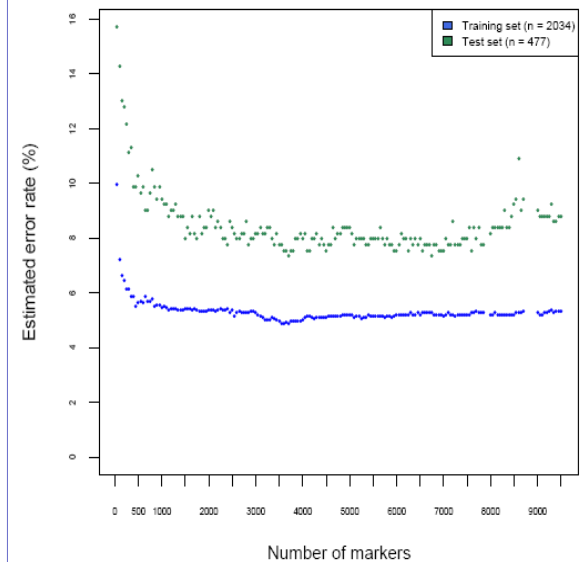
#### Clinically useful accuracy achieved with 1550 markers

We achieved 5.4% error rate with 1550 markers.

The performance gain over the optimal TOO 100-marker classifier (7.2%) was clinically and statistically significant (two-sample chi-square = 4.7,  $P = .03$ ).

Observed slight performance decrease with increasing number of markers used in the test. The trend was maintained when all (approx. 22,000) markers on the U133A chip were used in the model. The same pattern was observed when the models were applied on independent test set of 477 specimens.

#### Tissue of origin error rate vs. number of markers used in the model



### CONCLUSIONS

Cancer classification diagnostic tests based on gene expression may require relatively high number of markers – over a thousand – to approach clinically useful performance.

This conclusion favors DNA microarrays over quantitative PCR for cancer classification problems.

Pathwork is actively investigating whether this finding holds for other types of diagnostic problems (drug resistance, prognosis).

### REFERENCES

- [1] Lj. J. Buturovic, On the Minimal Dimension of Sufficient Statistics. IEEE Trans. Inform. Theory, vol. IT-38, pp. 182-186, January 1992.
- [2] F. A. Monzon, C. I. Dumur, M. Lyons-Weiler, L. Buturovic, Q. Tran, S. H. Becker, T. Rigl, G. G. Anderson, Clinical Validation of a Gene Expression Microarray-based Tissue of Origin Test Applied to Primary and Metastatic Tumors. Association for Molecular Pathology Annual Meeting, Orlando, 2006. Abstract ST #26.
- [3] Dumur et al., Interlaboratory Performance of a Microarray-Based Gene Expression Test to Determine Tissue of Origin in Poorly Differentiated and Undifferentiated Cancers. Submitted to Journal of Molecular Diagnostics, 2007.