Supplemental Data

Wolfe P, Murphy J, McGinley J, Zhu Z, Jiang W, Gottschall EB, Thompson HJ. (2004) Using nuclear morphometry to discriminate the tumorigenic potential of cells: a comparison of statistical methods. Cancer Epidemiol Biomarkers Prev. Jun;13(6):976-88

We used SYSTAT version 10 for graphics; SAS version 8.2, S-Plus 4.0, R[1], and iMiner 1.01 were used for all other analyses. In most cases, default parameters were used: stepwise logistic regression in SAS adds or removes variables at the 0.05 significance level, stepwise discriminant analysis adds or removes variables when their partial correlation coefficient is 0.01. The CART procedure in iMiner uses the RPART routine {Therneau, 1997 615 /id}, where entropy is the basis for splitting (stop when complexity <0.001), while S-Plus uses deviance (stop splitting when impurity is < 1%) and the number of observations at a node. We reduced the minimum number of leaves for a split (the default is 10) and specified the minimum deviance to be 0.001 in S-Plus. We did a search from 1 to 10 for the number of neighbors that minimized errors for the KNN application; misclassification within the learning set was minimized with sets of 5 neighbors; the default metric is the Mahalanobis distance computed using the full covariance matrix. We also tested the Euclidian distance and the Mahalanobis distance computed using only the diagonal of the covariance matrix in an effort to reduce the misclassification rate. Mahalanobis distance using the full covariance matrix was the best of the metrics we tried.

Multivariate Approaches to Separating the Groups of Cells
(shaded areas were not in the published results)

Method	Variables Selected from Full Dataset	Misclassification Rate*		Sensitivity	Specificity
		n-fold	20-fold		
LDA (SAS)	All	12.0%	-	86%	90%
QDA (SAS)	All	16.5%	-	85%	82%
KNN (SAS)	All k=5 metric=Mahalanobis Distance	8.0%	16.0% (7.7%)	83%	85%
Stepwise LDA (SAS)	Coefficient of Variation Pg DNA Sum Variance CellFeret Y Slope Valley	9.5%	10.0% (5.3%)	84%	96%
Stepwise Logistic (SAS)	Valley Coefficient of Variation Diagonal Moment Sum Optical Density	-	8.5% (6.2%)	89%	94%
CART (S-Plus)	Product Moment Difference Variance Sum Variance Contrast Peak Pg DNA Cell Area Avg Optical Density	-	11.0% (7.4%)	89%	89%
Neural Net (R)	All	-	20.5% (XX%)	85%	74%

Supplemental Data

Wolfe P, Murphy J, McGinley J, Zhu Z, Jiang W, Gottschall EB, Thompson HJ. (2004) Using nuclear morphometry to discriminate the tumorigenic potential of cells: a comparison of statistical methods. Cancer Epidemiol Biomarkers Prev. Jun;13(6):976-88

Method	Variables Selected from Full Dataset	Misclassification Rate*	Sensitivity	Specificity
Logistic (iMiner)**	**All**	5.5% (5.5%)	na	na
CART ** (iMiner)	**All**	3.5% (5.3%)	na	na
Neural Net ** (iMiner)	All	10.5% (7.2%)	na	na

*method: 20-fold cross-validation: the dataset was partitioned into 10 sets of 20 observations, 10 cancer and 10 normal. Each set of 20 was used as the test set for parameters estimated using the remaining 180 data points as the learning set. The misclassification rate for the test set, and its SD, were calculated from the misclassification rates across the 10 models.

** The iMiner program selects its own samples, randomly holding out 10% (10 observations from each cell line) to use as a test sample for the predictors based on the remaining 80% sample. This method is not the equivalent of 20-fold cross-validation, but is rather a bootstrapping technique for which 10 realizations does not give a good estimate of the misclassification rate. As with any mining package, we did not try to select the best variables in the initial pass through the data; rather, we let the program do the prediction based on all of the data. This and the difference in method of computation account for the misclassification rates being lower overall in the iMiner routines.

[1] R is an open source software environment for statistical computing and graphics, similar to the S system, which was developed at Bell Laboratories by John Chambers et al. A variety of sophisticated analysis packages can be downloaded form www.r-project.org.