The Role of Proxy Genes in Predictive Models: An Application to Early Detection of Prostate Cancer

Jay Magidson[1], Karl Wassmann[2]
[1]Statistical Innovations Inc., 375 Concord Ave., Ste. 007, Belmont, MA 02478
[2]Source MDx, 199 Wells Ave., Ste. 209, Newton, MA 02459

Abstract

The most important predictor in a regression model may be a suppressor variable which does not predict the outcome variable directly but improves the overall prediction by enhancing the effects of other predictors in the model. The most important gene in a 9-gene model for early detection of prostate cancer is the gene SP1, whose mean is not significantly different between cancer and medically defined normal subjects. We suggest that SP1 predicts the pre-cancer expression of genes that it regulates, including some that do have direct effects. We refer to such suppressor variables as 'proxy genes', and its associated genes that have direct effects as 'prime genes'.

We introduce the basic ideas of proxy genes using a simple 2-gene prime/proxy model, and then present the 9-gene + PSA model developed by Correlated Component Regression (CCR). CCR is a structured approach for developing a reliable predictive model based on prime and proxy genes from a potentially large pool of gene candidates to be included in the model (Magidson, 2010a, 2010b). Simulation results suggest that when one or more suppressor variables are among the potential predictors, CCR improves over alternative methods for analyzing high dimensional data such as popular penalty approaches and PLS regression.

Keywords: gene expression, suppressor variable, logistic regression, proxy genes

1. Background

In our exploratory analyses towards the development of diagnostic blood tests for different kinds of cancer we have evaluated several million logistic regression models and encountered a consistent pattern among the best 2-gene models. This same pattern has also been found to occur among the best 2-gene Cox models for predicting survival time. Regardless of the type of cancer,

1. One of the genes ('prime gene') has a direct effect, providing highly significant discrimination between the cancer vs. normal groups (or between the longer and shorter surviving risk score groups) when used separately in a 1-gene model.
2. The other gene (which we refer to as a 'proxy gene') is not at all significant in a 1-gene model but when included in the 2-gene Prime/Proxy gene model, it significantly improves prediction over the model containing only the prime gene.
3. In the Prime/Proxy 2-gene model, the coefficients have different signs -- one gene has a significant positive coefficient and the other a significant negative coefficient.

Reprinted From:
Magidson, J., and K. Wassmann. (2010). "The Role of Proxy Genes in Predictive Models: An Application to Early Detection of Prostate Cancer", 2010 JSM Proceedings of American Statistical Association, Biometrics Section, pp. 2739-2753.

4. The Prime/Proxy gene pair has moderate to high positive correlation (> 0.5). That is,
   – subjects with high expression on the Prime gene also tend to have high expression on the Proxy gene for both cancer and normal subjects
   – subjects with low expression on the Prime gene also tend to have low expression on the Proxy gene for both cancer and normal subjects

From a statistical perspective, the proxy gene acts as a 'suppressor' variable (Horst, 1941, Lynn, 2003), enhancing the effect of the prime gene despite having no direct predictive power of its own. This general pattern has been observed previously among 'synergistic' pairs of genes (Hanczar, et. al., 2007). However, to the best of our knowledge, the suppressor effect has not previously been identified as such in applications involving biological conditions. The correlation between the prime and proxy genes is hypothesized to result from natural interaction of underlying biological mechanisms including transcriptional control, cell communication and extracellular signaling.

2. Prime and Proxy Genes

As an example of a prime/proxy relationship, Table 1 summarizes the results of a 2-gene logistic regression model discriminating between N = 128 prostate cancer (CaP) subjects and N = 94 normal males based on training data. Despite the fact that the gene SP1 is non-significant in a 1-gene model (validation p-value =.73), it makes a significant contribution in the 2-gene model containing the prime gene CD97.

Table 1: SP1 is Example of "Pure" Proxy Gene, for Prime Gene CD97 in the 2-gene Model

Genes in 2-gene model		Entropy	Coefficient		Wald p-value		1-gene model p-value	
Sample	N	R-sq	CD97	SP1	CD97	SP1	CD97	SP1
Training	**222**	0.37	5.39	-5.07	4.8E-13	1.4E-11	1.5E-6	**0.42**
Validation	**152**	0.28	3.63	-2.97	7.0E-09	2.2E-06	6.5E-06	**0.73**

Let 'b' denote the regression coefficient of SP1 in a linear regression of CD97, and define 'Logit' be the log-odds of Cancer vs. Normal. Rearranging terms in the logistic regression model below so that a + b * SP1 corresponds to the linear regression prediction of CD97, eq. (1) shows that SP1 functions as a proxy gene to enhance the effect of the Prime gene CD97 by subtracting out irrelevant variation from CD97. It is a *pure* proxy if $(\beta_2 / \beta_1) = b$

$$Logit = \alpha + \beta_1 * [CD97 - (a + b * SP1)] \qquad (1)$$

$$= \alpha' + \beta_1 * CD97 - \beta_2 * SP1$$

$$\beta_2 = b * \beta_1$$

Based on our training sample data, the estimates for β_2 / β_1 and for b are 0.94 and 0.88 (with standard error .05) respectively. The corresponding quantities estimated in the

validation data are and 0.82 and 0.86 (with standard error .07). The closeness of .94 to .88 (and .82 to .86) supports the hypothesis that SP1 is a pure proxy gene for the prime gene CD97.

SP1 is a transcription factor that has been extensively studied and widely reported for its role in regulation of gene expression for many genes. Specifically, SP1 is known to regulate CD97. SP1 can also be modulated by several regulatory pathways which also may account for its high correlation that we have observed with many genes. The observed correlation between SP1 and CD97 in the training data is .73 and in the validation data .78.

One way to assess the enhancer effect of the proxy gene SP1 is to estimate the difference in the area under the ROC curve (AUC), with and without SP1 included as a predictor. Figure 1 and Tables 2A and 2B provide the results of comparing the ROC curve for the 2-gene model with the corresponding curve based on CD97 alone. From the training data, we estimate the difference in AUC to be .17, which is significantly different from 0 (p = 7.6E-5), and the corresponding estimate from the validation data is .10 (p = .05).

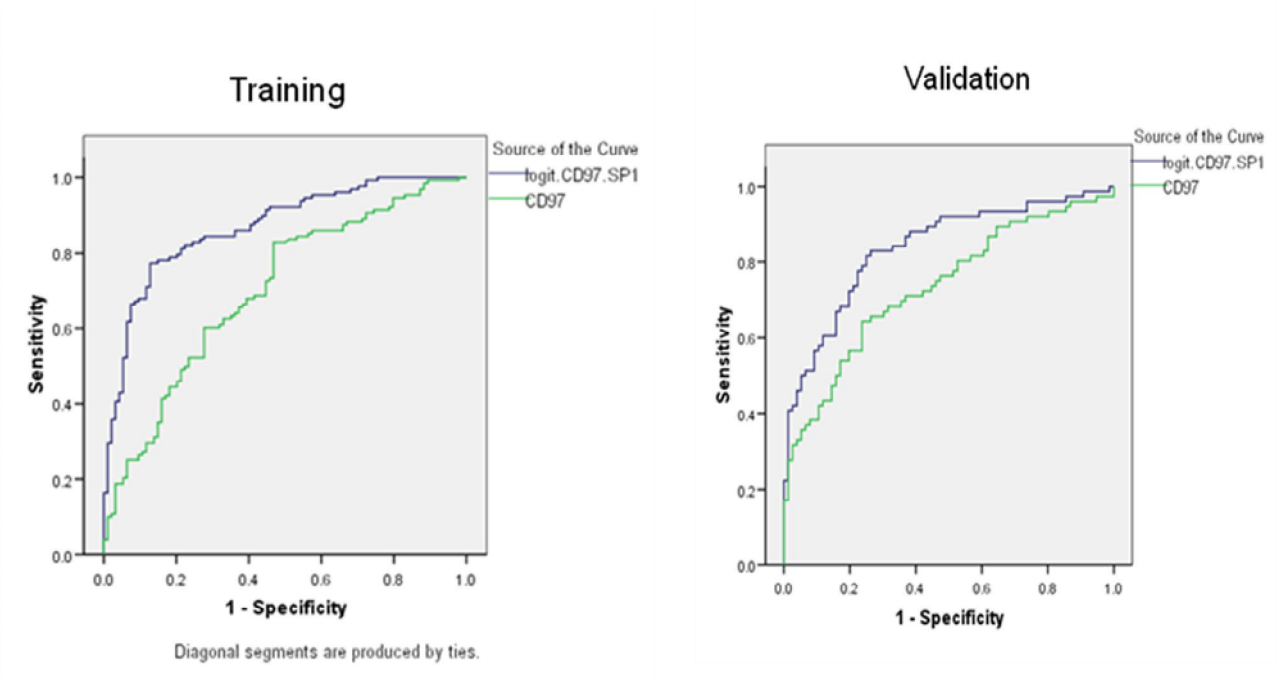

Figure 1: Enhancer Effect of Proxy Gene Illustrated by the difference in ROC Curves

Table 2A: comparisons of ROC results from Training Data:
CaP (N=128) vs. Normals (N=94)

2-gene model: CD97 + SP1	**AUC =**	**0.87**
1-gene model: CD97	**AUC =**	**0.70**
Improvement in AUC by Inclusion of SP1	**Δ AUC =**	**0.17**
AUC Difference p-value	**p-val =**	**7.6E-05**
Logit Model Unique Contribution for SP1	**p-val =**	**1.4E-11**
1-gene model: SP1	**AUC =**	**0.52**
AUC p-value	**p-val =**	**0.57**

Table 2B: comparisons of ROC results from Validation Data:
CaP (N=76) vs. Normals (N=76)

2-gene model: CD97 + SP1	**AUC =**	**0.84**
1-gene model: CD97	**AUC =**	**0.73**
Improvement in AUC by Inclusion of SP1	**Δ AUC =**	**0.10**
AUC Difference p-value	**p-val =**	**0.047**
Logit Model Unique Contribution for SP1	**p-val =**	**2.2E-06**
1-gene model: SP1	**AUC =**	**0.50**
AUC p-value	**p-val =**	**0.93**

Figure 2 displays scatterplots of CD97 and SP1 (in delta ct units) from the training and validation data, with separate 68% concentration ellipsoids superimposed for the cancer and normal subjects.

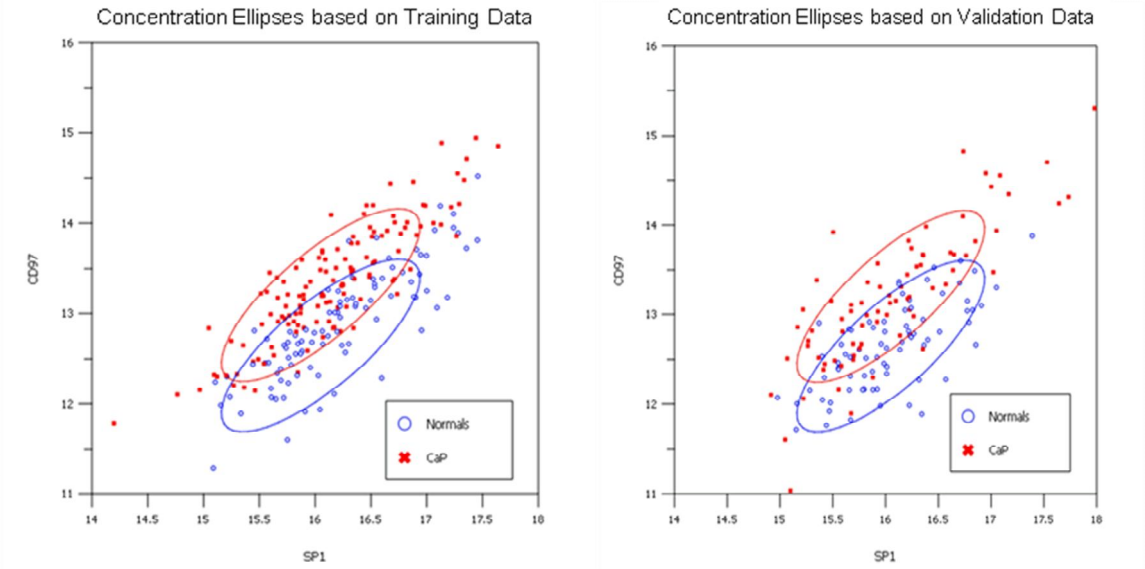

Figure 2: Example of Prime/Proxy Gene Pair Providing Good Separation of CaP vs. Normals as Confirmed by Validation Data

From a visual perspective, the validation data supports the hypothesis developed from the training data that the concentration ellipses provide good separation of CaP and Normals. Most normals are contained within the blue ellipse while most CaP subjects are contained within the red ellipse. The scatterplots also show the high positive correlation between CD97 and SP1.

Note that the CaP Subjects have an elevated Δct level for CD97 as compared to Normals – i.e., in both the training and validation plots, the red ellipse lies above the blue ellipse. (The Δct level is inversely related to the gene expression – the higher the Δct level, the lower the expression.)

Since CaP and Normals do not differ on SP1, they can be equated (i.e., they are exchangeable) on SP1 and differences between the CaP and Normal subjects are isolated to CD97.

Recall eq(1). The quantity a + b * SP1 can be interpreted as the prediction of the Δct level for CD97 based on given SP1 that would have been attained at an earlier time point, and for the CaP subjects, that time point is prior to the occurrence of cancer. At that prior time point, both the normal and CaP subjects can be viewed as being from a common 'normal' population as evidenced by the common concentration ellipsoid. In this way, subtracting out the prediction of the earlier CD97 measurement, inclusion of the proxy variable converts the model from prediction based on CD97 to prediction based on the change in CD97. Thus, the prime/proxy model provides a look back in time and CD97 – [a + b*SP1] reflects the predicted change in CD97 that has occurred since that time point. Hence a prime/proxy model can also be viewed as a way of capturing some of the power of longitudinal analysis based on cross-sectional data.

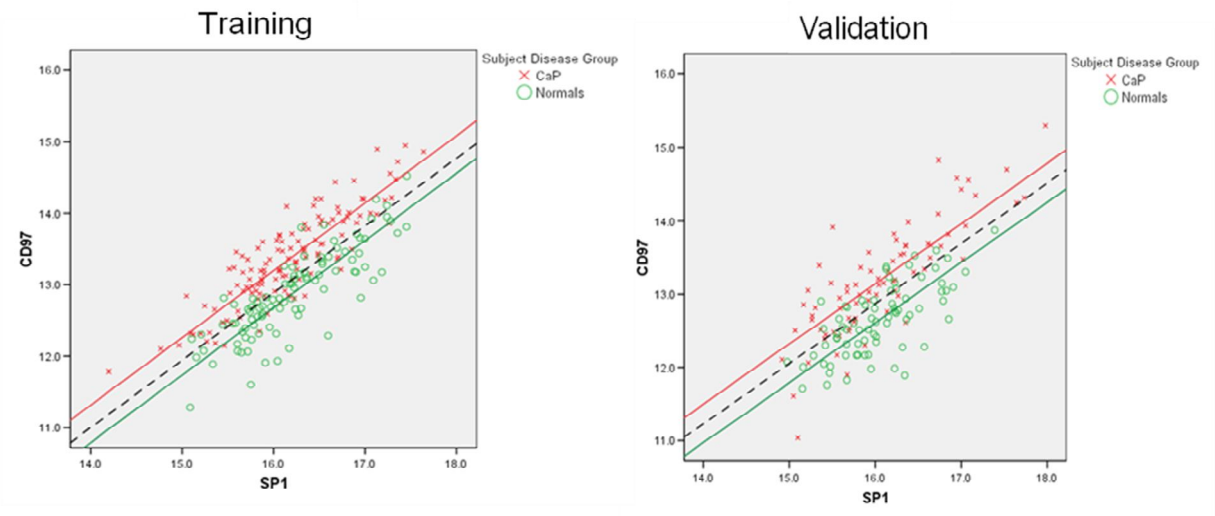

Figure 3: Regression Line for Cancer and Normal Subjects Under the Assumption of a Common Slope

Figure 3 shows the regression line for Cancer and Normal subjects under the assumption of a common slope:

- Regression line for CaP subjects (red line): CD97 = -1.30 + .91 SP1 Regression line for Normal subjects (green line): CD97 = -1.82 + .91 SP1
- Assuming that the Normals are representative of the gene expression for the CaP subjects when they were in a normal state, change in the measurement of CD97 expected as a subject goes from a normal to CaP state can be estimated by the vertical distance between the red and green lines = .53.

Since the measurements are obtained from whole blood, the signal being measured comes from a mix of blood cells. In the case of cancer, there may be a higher mix of natural killer cells, which may account for the change as opposed to down-regulation (lower expression of CD97). Either way, the prime/proxy model captures the effect.

	Training data	Validation data
* Predicted CD97 for CaP subjects:	CD97 = -1.30 + .91 SP1	CD97 = -0.40 + .85 SP1
Minus predicted CD97 for normals:	- (CD97 = -1.82 + .91 SP1)	- (CD97 = -0.92 + .85 SP1)
	= 0.52	= 0.53

Figure 4 revises the scatterplots given in Figure 2 by downshifting the values on CD97 by 0.53 for the cancer subjects. Both the Normal and CaP Subjects are now contained by a single common normal ellipse, suggesting a single homogeneous normal population.

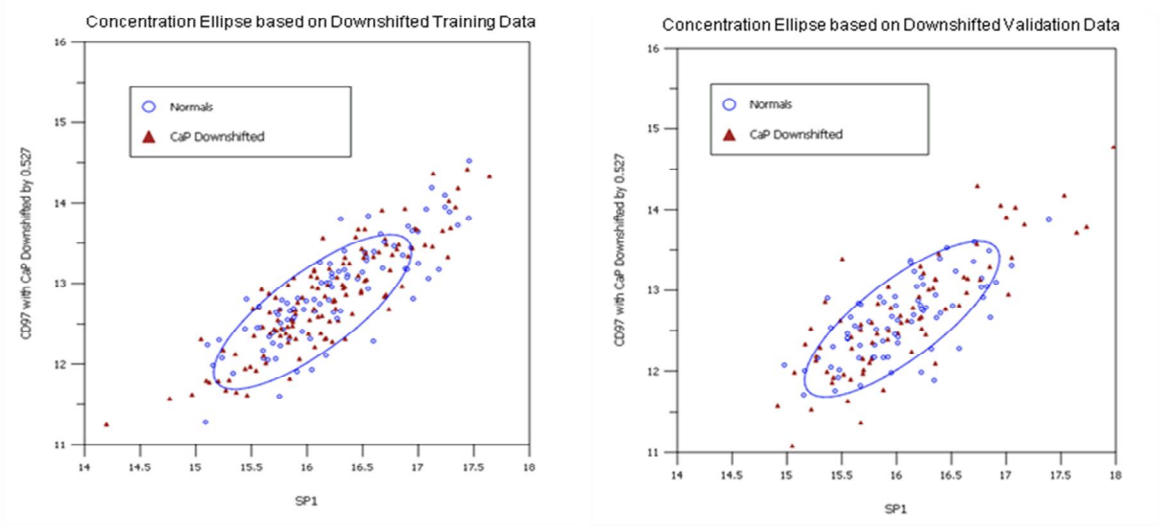

Figure 4: Revision of Figure 2 Scatterplots with CaP Subjects Downshifted by .53* ct on CD97

3. Variable Selection and Pre-screening

Due to problems with high and ultra-high dimensional data associated with hundreds or thousands of correlated candidate genes, a method for selecting a subset of these genes is required. Unfortunately, most methods select only predictors that are significantly related to the dependent variable, and hence exclude proxy genes. When suppressor variables

exist among the candidate predictors, this misguided strategy results in a substantial reduction in predictive power. Fan, et. al (2009) recognized this problem with his SIS variable selection scheme and proposed an alternative (ISIS) which is intended to retain suppressor variables.

Magidson (2010b) proposed an alternative selection method based on his Correlated Component Regression (CCR) model and showed that it outperformed ISIS using the simulated data provided by Fan. Additional simulations demonstrate that when one or more suppressor variables are included among the candidate predictors, with or without variable selection, CCR predicts better than the commonly used penalty regression techniques (Magidson and Yuan, 2010).

Here, we present results from the Magidson CCR-LDA step-down algorithm as applied to 22 genes (plus the protein PSA) all of which were available on our training and validation data. The resulting CCR model which was developed on the training data consisted of 9-genes + PSA. For comparison, results from a stepwise linear discriminant analysis, which selected 5 predictors -- 4 genes + PSA, are also described and compared with the CCR model on the validation data.

A brief introduction to CCR is given in the next section. For further details on CCR, including the step-down algorithm, see Magidson (2010a, 2010b).

4. Overview of Correlated Component Regression (CCR)

When the number of genes (# predictor variables P) approaches or exceeds the sample size (# subjects N), traditional regression modeling becomes problematic and near or perfect prediction is always attainable by *overfitting*. In this case, results are unstable and fail to validate when applied to new subjects. In CCR, a regression model consisting of K < P components is developed by applying the Naïve Bayes rule K times. Each component is an exact linear combination of the P predictors. The first component captures the effects of the genes which have direct effects (e.g., 'prime genes'), and the 2^{nd} component, correlated with the 1^{st}, captures the effects of the proxy genes that improve prediction by removing extraneous variation from one or more prime genes. Each additional component significantly improves prediction further.

When K=P, the resulting CCR model is equivalent to the traditional regression model. In the case of a dichotomous dependent variable, different variants of CCR correspond to the logistic regression model (CCR-Logistic) and linear discriminant analysis (CCR-LDA). Here we present results from a 3-component CCR-LDA model developed using the 9 genes plus PSA which were selected by the CCR step-down algorithm. Since P = 10 predictors, the 10-component CCR model is equivalent to the traditional LDA model. Taking K < P reduces the dimensionality of the model from P to K which serves as a regularization similar to PLS regression and also to penalty regression approaches.

The optional CCR step-down algorithm can be used to reduce the number of predictors in the model, where at each step the least important predictor is removed from the model. The algorithm stops when elimination of the least important predictor produces a significant drop in predictive performance, as assessed say by M-fold crossvalidation.

Prime genes are identified as those having significant loadings on component #1, and proxy genes as those having significant loadings on component #2, and non-significant loadings on component #1.

5. Results for Prostate Cancer Data: 9-Gene + PSA Model

Ross, et. al. (2010) describes the results from the 9-gene +PSA 3-component model obtained using CCR-LDA, the linear discriminant analysis variant of CCR. The proxy gene, SP1, is the most significant gene in the model despite having no direct predictive power of its own. SP1 improves prediction by transforming the effect of the prime gene(s) CD97, RP51077B9.4 and IQGAP1 to the more predictive effects associated with changes in the prime genes, in the matter suggested earlier in the 2-gene model. Table 3A presents the coefficients in the final model

Table 3A: Coefficient Estimates for 3-component CCR-LDA model

Gene	Coefficient
ABL1	1.50
BRCA1	-1.72
CD97	2.12
IL18	-1.38
IQGAP1	1.04
MAP2K1	2.32
MYC	-1.80
RP51077B9.4	2.58
SP1	-3.29
plnPSA	3.42

Table 3B: p-values for component loadings

	p-value		
k=	1	2	3
ABL1	5.2E-06	4.8E-06	0.46
BRCA1	0.0009	5.4E-09	0.40
CD97	2.3E-07	0.0001	0.048
IL18	5.2E-06	0.0003	0.91
IQGAP1	0.03	7.8E-07	0.15
MAP2K1	9.0E-05	6.8E-07	0.14
MYC	0.08	6.9E-10	0.03
RP51077B9.4	4.7E-09	0.0002	0.46
SP1	0.42	2.0E-12	0.007
plnPSA	3.9E-40	0.008	0.01

Table 3B presents p-values of the loadings on each component for each predictor in the 3-component model. The p-values assess the contribution of each gene on each component k of the 3-component model. Prime genes (shaded blue) have significant ANOVA-like p-values in a 1-gene model (i.e. significant p-values associated with the loading on the first component), and proxy genes (shaded green) have non-significant 1-component loadings, but significant component #2 loadings. (Note: since the loadings for MAP2K1 on both components 1 and 2 are significant, and the loading on the 2nd component is more significant than on the 1st component, MAP2K1 possibly serves as both a Prime and a Proxy gene. Therefore, we left MAP2K1 unshaded).

Tables 4 and 5 present group means and likelihood ratio (LR) p-values for each of the 9 genes for the validation and training data respectively (for the training data these p-values are similar to the ANOVA-like p-values associated with the 1st component in the CCR-LDA model. The difference is that the LDA normal distribution assumption is not made in the derivation of the LR p-values).

Table 4: Group Means and p-values Based on Validation Data

Validation Data CaP (N=76) vs. Normals (N=76)				
		Mean		LR
1-gene Model	Prostate	Normals	Difference	p-val
ABL1	18.65	18.12	0.53	1.3E-09
RP51077B9.4	16.61	16.26	0.35	9.5E-09
CD97	13.20	12.65	0.56	1.8E-07
MAP2K1	15.99	15.61	0.38	1.9E-06
IL18	21.72	22.04	-0.32	2.1E-03
IQGAP1	14.23	13.99	0.24	3.5E-02
BRCA1	21.56	21.73	-0.18	0.12
MYC	18.28	18.16	0.12	0.28
SP1	16.07	16.04	0.03	0.73

The 9-Gene Model Consists of 5 Prime Genes that are Significant Predictors in a 1-Gene Model as Assessed by the p-value in the 1-component model

Table 5: Group Means and p-values Based on Training Data

Training Data CaP (N=128) vs. Normals (N=94)				
		Mean		LR
1-gene Model	Prostate	Normals	Difference	p-val
RP51077B9.4	16.70	16.37	0.32	3.1E-09
CD97	13.34	12.88	0.46	1.5E-07
ABL1	18.64	18.32	0.32	4.2E-06
IL18	21.64	22.08	-0.44	1E-05
MAP2K1	16.05	15.81	0.24	7.6E-05
BRCA1	21.66	21.95	-0.29	0.0009
IQGAP1	14.34	14.15	0.19	0.03
MYC	18.24	18.39	-0.16	0.07
SP1	16.15	16.22	-0.06	0.42

Table 6 shows that the Proxy Genes Have Moderate to Large Correlation with some of the Prime Genes. Correlations in Table 6 above greater that 0.5 are shaded in yellow.

Table 6: Pooled Within-Groups Correlations

	plnPSA	CD97	ABL1	BRCA1	IL18	IQGAP1	MAP2K1	MYC	RP51077B9.4	SP1
plnPSA	1.00	.06	.15	.04	.10	.07	.17	.14	.14	.06
CD97	.06	1.00	.66	.68	.21	.85	.71	.44	.71	.86
ABL1	.15	.66	1.00	.55	.29	.65	.86	.72	.60	.73
BRCA1	.04	.68	.55	1.00	.48	.79	.66	.47	.56	.79
IL18	.10	.21	.29	.48	1.00	.35	.49	.32	.27	.36
IQGAP1	.07	.85	.65	.79	.35	1.00	.77	.55	.63	.91
MAP2K1	.17	.71	.86	.66	.49	.77	1.00	.71	.65	.81
MYC	.14	.44	.72	.47	.32	.55	.71	1.00	.52	.64
RP51077B9.4	.14	.71	.60	.56	.27	.63	.65	.52	1.00	.70
SP1	.06	.86	.73	.79	.36	.91	.81	.64	.70	1.00

Prime genes shaded in blue and proxy in green.

Table 7 shows how the coefficients in the K-component CCR-LDA model change as K increases from 1 to P = 10. For example, in the 1-component model the coefficient for SP1 is -.14 and non-significant (p=.42; recall table 3B). It increases to -.75 in the 2-component model (after taking into account component 1) and then further to -3.29 in the 3-component model. This rapid increase is typical for an important suppressor variable (Table 6 shows that the correlation between SP1 and CD97 is .86). In contrast, PSA (measured in logarithmic units by 'plnPSA') remains fairly constant as the number of components K increases, typical for predictors that have fairly low correlations with the other predictors in the model. Table 6 shows that the correlations between PSA and the other model predictors are all less than .2.

As mentioned earlier, for K = P, the model coefficient estimates are equivalent to the traditional regression estimates, here being the logistic regression coefficients attainable from LDA. As such, standard errors are confidence intervals for those coefficients are readily available. Table 8 below provides the Z statistic (coefficient/standard error) as well as the 95% confidence interval as estimated from 1000 bootstrap samples.

The correlation between the coefficients can be used as a measure of similarity between the different numbers of components. The correlation between the coefficients associated with K=10 (LDA coefficients) and K=5 is .9987 and higher than .999 for K > 5. The correlation between the LDA coefficients and those for K=4 is .987, .960 for K=3, .735 for K=2 and .740 for K=1.

Table 7: Comparison of Coefficient Estimates in K-component Model as K goes from 1 to P

# Components =	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9	K = 10
plnPSA	2.95	4.27	3.42	3.53	3.55	3.49	3.48	3.46	3.45	3.43
ABL1	0.95	0.85	1.50	1.51	1.38	1.45	1.46	1.49	1.49	1.49
BRCA1	-0.54	-1.11	-1.72	-1.49	-1.34	-1.38	-1.51	-1.50	-1.49	-1.48
CD97	0.87	0.88	2.12	2.06	2.12	2.04	2.03	1.99	1.99	1.98
IL18	-0.69	-1.13	-1.38	-1.01	-1.09	-1.11	-1.04	-1.03	-1.02	-1.02
IQGAP1	0.35	0.12	1.04	1.59	2.27	2.36	2.43	2.44	2.43	2.42
MAP2K1	0.94	0.72	2.32	3.00	2.50	2.47	2.33	2.29	2.26	2.25
MYC	-0.28	-0.78	-1.80	-1.83	-1.63	-1.51	-1.50	-1.51	-1.51	-1.50
RP51077B9.4	1.61	1.70	2.58	2.33	2.35	2.42	2.42	2.44	2.42	2.41
SP1	-0.14	-0.75	-3.29	-4.57	-5.14	-5.28	-5.19	-5.16	-5.13	-5.11

Table 8: Coefficient Estimates in P-component Model ("LDA") and Associated Statistics

	LDA		95% Conf. Interval	
	Coeff.	Z	Lower	Upper
pInPSA	3.43	7.6	2.6	4.4
ABL1	1.49	1.6	-0.2	3.5
BRCA1	-1.48	-2.2	-2.9	-0.2
CD97	1.98	2.6	0.5	3.5
IL18	-1.02	-2.5	-1.8	-0.2
IQGAP1	2.42	2.3	0.5	4.6
MAP2K1	2.25	1.4	-0.9	5.3
MYC	-1.50	-2.7	-2.7	-0.5
RP51077B9.4	2.41	2.4	0.5	4.6
SP1	-5.11	-3.7	-8.1	-2.6

6. Conclusions and Future Research

6.1 Summary and conclusion

Retrospective analysis of logistic regression and survival models for seven types of cancer revealed a persistent pattern in the strongest early detection and survival gene expression models including breast, cervical, colon, lung, melanoma, ovarian and prostate cancers. This pattern led us to the discovery of the importance of including suppressor variables in our models and motivated the development of the Correlated Component Regression (CCR) method for developing sparse models with small sample sizes and high dimensional data (many predictors). The benefits of the Prime/Proxy gene expression concept for developing predictive models are the following:

1. Higher Predictive Power of the Gene Expression Panels: Gene expression models with Prime and Proxy genes have higher predictive value than models with Prime genes alone.
2. Improved Method for Selecting Gene Candidates: Prime/Proxy gene pairs have moderate to high positive correlation (0.5+): This correlation is expected if the genes are related via common biological mechanisms. This understanding is a key insight for the selection of candidate markers.
3. More Likely to Validate in an Independent Dataset: Supported by simulation results (Magidson and Yuan, 2010), CCR models are more likely to include proxy genes and thus result in less "noise" due to the incorporation of proxy genes "suppressor" effects. They are more likely to validate than traditional gene expression models containing prime genes alone.

This paper summarized the results of CCR as applied to the early detection of Prostate Cancer (CaP). Early detection of CaP is important because it is currently both over diagnosed (resulting in a large number of unnecessary biopsies) and over-treated. One potential limitation of the 9-gene + PSA model is that it was developed from case-control data to distinguish prostate cancer subjects from healthy normals. As such, it may or may not generalize to the population of subjects who are currently scheduled to undergo a biopsy. If not, it may not be useful in improving the screening of those patients who should undergo a biopsy. To address this issue, we are currently engaged in a 1,000 man prospective clinical biomarker trial called PRostate Cancer Individual Signature

Evaluation (PReCISE) in men undergoing scheduled prostate biopsy. If the biopsy negative population turns out to be significantly different from that of healthy normal, we will develop another model that works for the relevant population. See Appendix for further information.

References

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional variable selection: beyond the lienar model. Journal of Machine Learning Research, 10, 2013-2038.

Hanczar, Zucker, J., Henegar, C. and L. Saitta (2007). Feature Construction from Synergic Pairs to Improve Microarray-based Classification. Bioinformatics. Vol. 23, No. 21 2007, pages 2866-2872.

Horst, P (1941). The role of predictor variables which are independent of the criterion. Social Science Research Bulletin, 48, 431-436.

Lynn, H (2003). Suppression and Confounding in Action. The American Statistician, Vol.57, 2003.

Magidson, J. (2010a). A Fast Parsimonious Maximum Likelihood Approach for Prediction Outcome Variables from a Large Number of Predictors. Paper presented at COMPSTAT 2010.

Magidson, J. (2010b). Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features. Proceedings of the American Statistical Association.

Magidson, J. and Y. Yuan (2010). Comparison of Results of Various Methods for Sparse Discriminant Analysis. Unpublished report #CCR2010.1, Belmont MA: Statistical Innovations.

Ross, R. W., P. W. Kantoff, D. Bankaitis-Davis, S. Seng, L. Siconolfi, K. Storm, H. Xiong, L. Katz, J. Stein, J. Englund, J. Magidson, K. Wassmann, W. Oh, 2010, "Development of a Whole-Blood RNA Transcript-Based Test to Improve the Diagnosis of Prostate Cancer", submitted for publication.

Stamey TA, Donaldson AN, Yemoto CE, McNeal JE, Sozen S, Gill H (1998) Histological and clinical findings in 896 consecutive prostates treated only with radical retropubic prostatectomy: epidemiologic significance of annual changes. J Urol 160:2412–2417

Appendix: PReCISE Trial

The primary objective of the PReCISE trial is to validate that Prostate MDx Early Detection will predict prostate biopsy outcome as positive or negative for prostate cancer with high negative and positive predictive value. As a secondary objective, the trial will seek to validate that the Prostate MDx test can differentiate low-grade prostate cancer (Gleason score of 6 or less) from intermediate/high grade prostate cancer (Gleason score of 7, 8, 9 or 10). The Company expects that these tests will serve to reduce the number of unnecessary biopsies by improving upon the false-positive rates of other screening tests/criteria currently in use. The inclusion and exclusion criteria for the PReCISE Clinical Study are based on a protocol established by the National Cancer Institute's Early Detection Research Network program.

The PReCISE Clinical Study is a seventeen-site clinical study currently being conducted with eleven leading medical centers, led by the Dana-Farber Cancer Center/Brigham and Women's Hospital, the sites include Mt. Sinai Medical Center, Beth Israel Deaconess Medical Center, Duke University Medical Center and the affiliated VA Medical Center in Durham, NC, Johns Hopkins Medical Center, Northwestern University Medical Center, University of Chicago Medical Center, University of Michigan Medical Center, University of Washington Medical Center/Fred Hutchinson Cancer Research Center, Wayne State University/ Karmanos Cancer Institute, and five private urology practices including Bay State Clinical Trials, San Bernardino Urological Associates, University Urology Associates, New York City, Urology Clinics of North Texas, Urology Center of Colorado and Urology San Antonio Research.

This trial is expected to take 12 months to complete data accrual and analysis which are expected to be available by the end of 2010.

Need for an Improved Diagnostic Test for Prostate Cancer

Prostate cancer is the leading cause of cancer in men, with over 200,000 new cases projected in 2010 in the United States and is the second leading cause of cancer deaths in men after lung cancer, with approximately 29,000 deaths per year. Prostate cancer generally grows slowly and is both over-diagnosed and over-treated. The American Cancer Society estimates that one in six men will be diagnosed with prostate cancer in their lifetime, but only one in 35 men with prostate cancer will die of the disease (www.cancer.org).

The current screening method for early detection of prostate cancer is the quantification of circulating Prostate-Specific Antigen (PSA). PSA elevation, however, may result from a variety of causes including prostate cancer, benign prostatic hyperplasia (BPH), acute urinary tract infection, and bacterial prostatitis. Inadequate specificity of the PSA test leads to false-positive results which then prompt unnecessary but costly and invasive prostate biopsies in patients whose PSA elevation is not caused by an underlying malignant process in the prostate. Each year in the United States, PSA screening leads to more than 1.2 million men undergoing prostate biopsies, while only 200,000 new cases are actually diagnosed each year. This means that approximately one million men

undergoing biopsy could potentially have been spared the procedure if a non-invasive blood test such as the Source MDx® Prostate MDx™ had been available.

In the U.S., the reported economic costs associated with PSA screening for prostate cancer and the associated medical costs from the high PSA false positive rate are estimated at over $3 billion. An individual prostate biopsy procedure generally costs several thousand dollars. In addition to the significant costs to the health care system, biopsy-associated pain occurs in up to 20% of cases, as well as other co-morbidities including hematuria, urine retention, infection and arteriovenus fistula in up to 15% of cases (Stamey et al., 1998). There is therefore a medical need for a simple, non or minimally invasive screening test in patients with elevated PSA that can further discriminate likely prostate cancer patients that will benefit from exploratory biopsy from unlikely prostate cancer patients for whom the procedure is unnecessary and associated with costs and the risk of morbidity. Such a test could reduce the overall economic burden and co-morbidities related to prostate cancer screening.