UroToday International Journal®

www.urotodayinternationaljournal.com
Volume 2 - August 2009

A Four-Gene Expression Signature for Prostate Cancer Cells Consisting of UAP1, PDLIM5, IMPDH2, and HSPD1

Isabelle Guyon,[1] Herbert A. Fritsche,[2] Paul Choppa,[3] Li-Ying Yang,[2] Stephen D. Barnhill[1]
[1]Health Discovery Corporation, Savannah, Georgia; [2]University of Texas, M.D. Anderson Cancer Center, Houston, Texas; [3]Clarient Inc., Aliso Viejo, California
Submitted May 19, 2009 - Accepted for Publication June 30, 2009

ABSTRACT

INTRODUCTION: The objective of the study was to develop a gene expression test that is highly associated with the presence of prostate cancer for use as an adjunct to the pathology examination of tissue.

METHODS: A gene expression database (U133A Affymetrix) was produced from 87 preparations of laser microdissected cells obtained from cancer (G3 and G4) and noncancer prostate tissues. The database was analyzed using univariate feature ranking and recursive feature elimination algorithms (support vector machine) to identify overexpressed genes that were associated with prostate cancer. RT-PCR assays were developed for the unique 4-gene set that was found to be reflective of prostate cancer. The gene expression data were used to construct a mathematical equation to classify tissues as cancer vs noncancer. The RT-PCR tests and the calculated gene expression score were validated in an independently collected set of formalin-fixed and fresh-frozen prostate tissues.

RESULTS: Analysis of the U133A gene expression database identified a group of 63 genes that were overexpressed in cancer and also gave an AUC (area under the curve) of > 0.84 for separating cancer vs noncancer. The gene discovery was validated with a database of 164 independently collected tissues reported in the Oncomine database. The 63 gene set was reduced to a subset of 4 complementary genes (UAP1, PDLIM5, IMPDH2, and HSPD1), using univariate feature ranking and recursive feature elimination (RFE) algorithms that gave an AUC = 0.94 for discrimination between cancer and noncancer prostate cells. Quantitative RT-PCR (reverse transcriptase polymerase chain reaction) assays were developed and validated. A mathematical formula based on the gene expression values of the 4 genes along with a housekeeping gene was developed for the classification of cancer vs noncancer tissues. In a blinded validation study of 71 independent prostate tissue samples that included both fresh prostate tissues and formalin fixed tissues, the 4-gene test gave a sensitivity of 90% with a specificity of 97% (the 95% confidence interval was 86% - 100%).

CONCLUSION: The 4-gene RT-PCR test can be used to detect Gleason grade 3 and grade 4 cancer cells in prostate tissue and may be useful as an adjunct test to the pathology examination of prostate tissue taken at biopsy or prostatectomy.

KEYWORDS: Prostate cancer detection; Support vector machine; Gene signature; UAP1, PDLIM5, IMPDH2, HSPD1, DNA microarray; RT-PCR assay

CORRESPONDENCE: Herbert A. Fritsche, PhD, Dept. Laboratory Medicine, MD Anderson Cancer Center, Houston, TX 77030 (hfritsche@mdanderson.org).

CITATION: *UroToday Int J* 2009 Aug;2(4). doi:10.3834/uij.1944-5784.2009.08.06

UroToday International Journal®

original study

A Four-Gene Expression Signature for Prostate Cancer Cells Consisting of UAP1, PDLIM5, IMPDH2, and HSPD1

INTRODUCTION

Prostate cancer is a deadly disease that warrants early detection. There are an estimated 200,000 new cases and 25,000 deaths from prostate cancer each year in the US alone [1]. Although most men diagnosed with prostate cancer will not die as a result of this disease, Gleason grades 3 and 4 cancer cells (graded on a scale of 1 to 5), which are identified at biopsy, are generally recognized as aggressive cancers that require treatment. The standard blood test used to identify men for prostate biopsy measures the concentration of prostate-specific antigen (PSA). This test primarily detects benign prostate hyperplasia (BPH), so only 20 – 30% of biopsies are found to be positive for cancer. For those men who are at high risk for developing prostate cancer, such as those who have a positive family history, a negative biopsy is usually followed by a second biopsy. Approximately 10% of repeat biopsies are found to be positive for cancer. Thus, there is potential value in a gene-based test to aid the pathologist in the visual inspection of the biopsy tissue and, perhaps, to identify those men with negative first biopsies that should undergo a second biopsy for cancer detection.

Tissue microarray studies have been instrumental in identifying new gene candidate markers for prostate cancer [2–6]. In this paper, the authors describe discovery using gene expression databases for both cancer and noncancer prostate cells obtained by laser microdissection from radical prostatectomy specimens. They evaluated the gene expression profiles of prostate cells, taking into account the zonal origin and histological tissue classification. They are also preparing a future paper that will provide a detailed analysis of the gene expression characteristics of prostate cells by tissue zone and cancer grade. Analysis of the gene expression data with support vector machine (SVM) algorithms [7] yielded accurate tissue classification according to histological categories. Additional analysis of those genes with the recursive feature elimination algorithm (RFE) [8] resulted in small gene sets that were predictive of grade 3 and grade 4 prostate cancers. Through this discovery approach, the authors have identified a diagnostic molecular signature consisting of only 4 genes and have designed a practical, cost-effective RT-PCR assay. This assay is highly accurate for detecting the presence of grade 3 or grade 4 cancer cells in prostate biopsy tissue.

METHODS

DNA Microarray Data

The authors performed gene discovery with a dataset of 87 prostate cell preparations representative of the 3 anatomic zones of the prostate (central, CZ; peripheral, PZ; transition, TZ) and histological Gleason grades 3 and 4 cancer cell preparations.

Table 1. Distribution of Discovery Tissues Used for Gene Expression Analysis. doi: 10.3834/uij.1944-5784.2009.08.06t1

Zone	Histology	Number
CZ	Normal	9
	Dysplasia	4
	Grade 4	1
PZ	Normal	13
	Dysplasia	13
	Grade 3	11
	Grade 4	18
TZ	BPH	10
	Grade 4	8
Total		**87**

The histology and Gleason grades are shown in Table 1. The gene expression data for these cell preparations was provided by Thomas Stamey, MD, Emeritus Professor at Stanford University Medical Center. The tissue cell types were laser microdissected from frozen sections of the prostates obtained from patients having undergone prostatectomy. The cells were analyzed with an Affymetrix U133A microarray (Affymetrix, Santa Clara, CA), which reveals the gene expression of over 20,000 genes using a previously published protocol [9]. An average of the differences in fluorescence between the perfect match and mismatch pairs was provided by the Affymetrix GeneChip® reading software. This software was used in all subsequent calculations as the gene expression coefficients.

For validation of the gene discovery, the authors used publicly available gene expression data reported in the Oncomine repository [10]. The published gene expression databases [11-14], which are listed in Table 2, used the U95A Affymetrix array consisting of approximately 12,500 genes. These gene expression data sets were not generated from laser micro dissected cells, did not contain zonal and histological annotations, and only defined the tissues as cancer or noncancer. In order to carry out comparative analysis using data from both the U133A and U95A arrays, the authors reduced the U 133A gene set to 6830 genes having identical or highly similar probes on both arrays. They restricted themselves to the task of discriminating cancer from noncancer tissues.

RT-PCR Methods and Data

According to methods described in the data analysis section, the authors identified a 4-gene signature suitable for separation of cancer from noncancer prostate cells. To validate this gene signature as a diagnostic test for prostate cancer,

Table 2. Oncomine Data Used for Validation of Gene Discovery. doi: 10.3834/uij.1944-5784.2009.08.06t2

Source [Reference]	Histology	Number
Febbo [11,12]	Normal	50
	Tumor	52
LaTulippe [13]	Normal	3
	Tumor	23
Welsh [14]	Normal	9
	Tumor	27
Total		164

Table 3. Tissues Used for RT-PCR Validation of the 4-Gene Signature. doi: 10.3834/uij.1944-5784.2009.08.06t3

Source	Histology	Number
Set 1	Normal	5
	BPH	5
	Tumor	11
Set 2	Normal	5
	BPH	4
	Tumor	12
Set 3	Normal	8
	BPH	10
	Tumor	9
Total		71

they measured the gene expression of UAP1, PDLIM5, IMPDH2, and HSPD1 using 71 additional prostate tissue samples with real-time RT-PCR assays developed and validated at Clarient Inc. (Aliso Viejo, CA). Table 3 lists the number and source of the validation samples. Validation sets 1 and 3 were paraffin-embedded formalin fixed tissues obtained from MD Anderson Cancer Center, Houston TX; set 2 was frozen tissues obtained from Hue City Hospital, Hue, Viet Nam.

Tissue preparation. The freshly collected frozen tissue was thawed and homogenized in lysis buffer following collection. The lysate was further processed using the Qiagen RNA Blood Mini extraction protocol (Qiagen, Valencia, CA). The RNA samples were DNase treated following the RNA isolation. The RNA quality was assessed by the RNA integrity number (RIN) using the Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). The paraffin-embedded formalin fixed tissues were sectioned at 5.0 μM on glass slides. The tissue sections were assessed for areas of interest by a pathologist using an H&E stained slide. The targeted areas were selectively removed from the unstained slide using a manual microdissection technique. The collected tissue was digested for 5 hours using Proteinase K and a digestion buffer optimized for RNA isolations. The lysate was further processed using a column-based Qiagen RNA extraction protocol. The samples were DNase treated following the isolation. The RNA yield was determined using a NanoDrop 1000 (NanoDrop, Wilmington, DE), and all samples were brought to a uniform final concentration.

Oligonucleotides. The primers and probes for IMPDH2 and PDLIM5 were obtained from Applied Biosystems TaqMan Gene Expression Assays (Applied Biosystems, Foster City, CA). The primers and probes for HSPD1 and UAP1 were designed using Primer Express v. 2.0 (Applied Biosystems). All primer sets and probes spanned an exon boundary and generated amplification products of similar sizes. The reaction efficiencies were evaluated for each set and determined to have comparable efficiencies. Various combinations of primers and probes were evaluated in multiplex reactions to find the best arrangement. Additionally, expression analysis was evaluated for each gene using prostate tissue to determine which genes had similar expression levels relative to each other. The most efficient and robust arrangement was found to be IMPDH2 and HSPD1 in one reaction and PDLIM5 and UAP1 in a second reaction. Table 4 lists the primer information and sequences.

Reference Genes. During the initial development of the assay, the authors evaluated a number of reference genes for use in the quantitative RT-PCR assays. They used prostate samples that were evaluated by a pathologist and determined to be either normal, benign prostatic hyperplasia, or cancer to evaluate the stability of various reference genes. Five genes were found to be acceptable for use as reference genes for quantitative gene expression analysis. All 5 reference genes were assayed for each sample. Quantitation of the target gene expression was assessed for each gene individually and relative to the geometric mean expression of the reference genes. Following evaluation of all 5 reference genes, beta-2-microglobulin (B2M) was found to have the most stable expression overall. B2M performed better than any individual gene and was comparable to the average of the 5 reference genes.

RT-PCR Assay. All RNA samples were assayed using one-step real-time RT-PCR (Applied Biosystems, Foster City, CA). A uniform quantity of input RNA was evaluated for each gene in duplicate reactions. Various concentrations of primers and probes were tested for each reaction to find the optimal reaction conditions. The most efficient and robust amplification was generated using 0.9 μM for each primer and 0.25 μM for each probe. The reactions were all found to have balanced amplification using

Table 4. Primer and Probe Information for the 4 Target Genes and 5 Reference Genes. doi: 10.3834/uij.1944-5784.2009.08.06t4

Primer/Probe Mix	Sequence Information	PCR Product Size (bp)
HSPD1 Forward	5′ AAC CTG TGA CCA CCC CTG AA 3′	64
HSPD1 Reverse	5′ TCT TTG TCT CCG TTT GCA GAA A 3′	
HSPD1 Probe	5′ VIC ATT GCA CAG GTT GCT AC NFQ 3′	
IMPDH2	ABI 20X (Gene Expression Assay Reagent Hs01021353_ml)	71
PDLIM5	ABI 20X (Gene Expression Assay Reagent Hs00935062_ml)	70
UAP1 Forward	5′ TTG CAT TCA GAA AGG AGC AGA CT 3′	68
UAP1 Reverse	5′ CAA CTG GTT CTG TAG GGT TCG TTT 3′	
UAP-1 Probe	5′ VIC TGG AGC AAA GGT GGT AGA NFQ 3′	
ABL	ABI 20X (Gene Expression Assay Reagent Hs99999002_mH)	105
ACTB	ABI 20X (Gene Expression Assay Reagent Hs03023943_gl)	96
B2M	ABI 20X (Gene Expression Assay Reagent Hs00187842_ml)	64
GAPDH	ABI 20X (Gene Expression Assay Reagent Hs00266705_gl)	74
GUSB	ABI 20X (Gene Expression Assay Reagent Hs99999908_ml)	81

the same primer and probe concentrations for each gene. All samples were determined to be free of contaminating DNA by running minus RT reactions for each sample. All samples were run on an ABI 7900HT using SDS v. 2.3 (Applied Biosystems, Foster City, CA). The relative expression data were determined for each target and reference gene using consistent settings for each run. Standard curves were prepared using Universal RNA (Stratagene, La Jolla, CA). The dilution series ranged from 100 ng to 10 pg of total RNA. Standard curves and calibration controls were run for each gene to generate relative quantitative values and assess amplification efficiency as well as run-to-run variation.

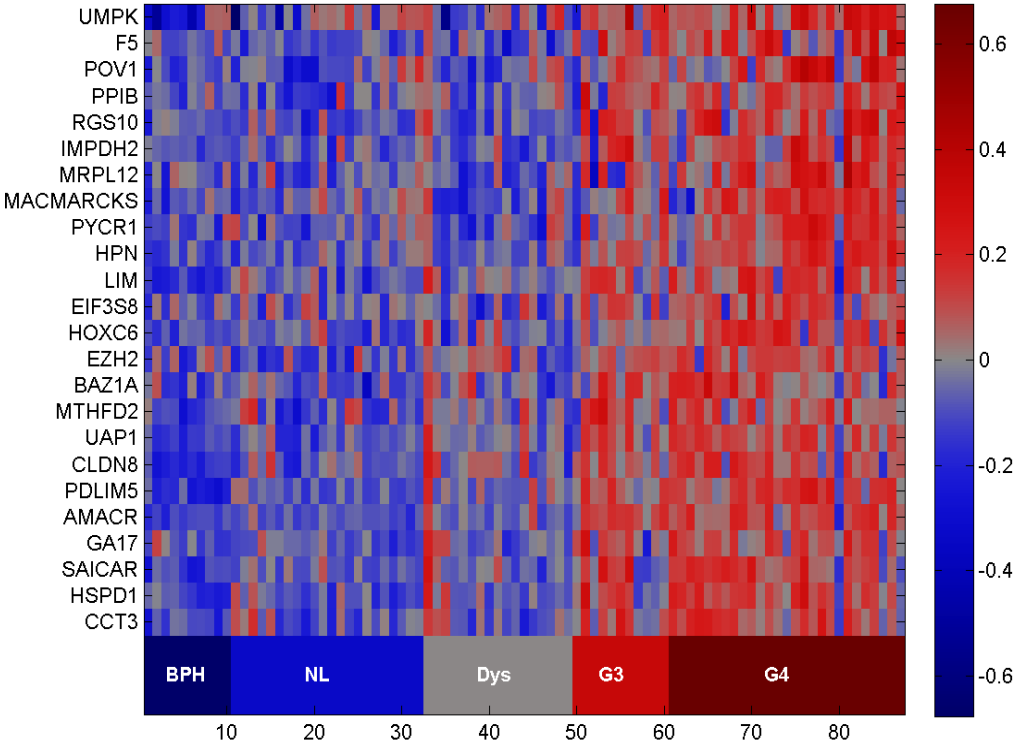
Data Analysis

The gene expression coefficients were processed by a suite of data analysis algorithms in Matlab® (The Mathworks). The mathematical problem was reduced to a two-class classification problem. In Table 1, *grade 3* and *grade 4* samples were labeled as *cancer* and all others as *noncancer*; in Table 2 and Table 3 *tumor* samples were labeled as *cancer* and all others as *noncancer*. A gene signature was defined using the discovery data (Table 1) and then validated with the microarray validation data (Table 2) and the RT-PCR test data (Table 3) in a three-step procedure. Prior to performing steps 1 and 2, the gene expression coefficients were preprocessed by: (a) log transformation of all gene expression coefficients; (b) standardization of all expression values for each sample within each microarray, accomplished by subtraction of the array mean and division by the standard deviation; (c) standardization of the expression values of each gene across all samples, performed in a similar manner; (d) repeat of *step b*; (e) repeat of *step c*; (f) take the tanh of the resulting values. The preprocessing *step a* equalizes the variances of the cancer and noncancer classes. *Step b* reduces the variance due to sample processing. *Step c* suppresses the effect of variation of abundance in mRNA between genes. Repeating *step b* and *step c* further reduces undesired scaling variability. *Step f* reduces the problem of outliers.

Step 1. Univariate gene ranking. Using discovery data, the gene expression coefficients were ranked on the basis of the area under the ROC curve (AUC) of individual genes to identify genes most characteristic of cancer (ie, separating best cancer samples from non-cancer samples). A single gene may be used for classification by setting a threshold on its expression value. Varying the threshold allowed the authors to monitor the tradeoff between sensitivity and specificity and obtain the ROC curve, which plots *sensitivity* vs *specificity* (sensitivity is defined as the rate of successful disease tissue classification; specificity is the rate of successful control tissue classification). The area under that curve (AUC) is a number between 0 and 1 providing a score, independent of the choice of the threshold, such that larger values indicate better classification power. Thus, ranking on the basis of the AUC allows us to assess the classification power of individual genes. The statistical significance of

UroToday International Journal®
original study
Isabelle Guyon, Herbert A. Fritsche, Paul Choppa, Li-Ying Yang, Stephen D. Barnhill
www.urotodayinternationaljournal.com

Figure 1. Heat map of the 19 genes preselected in step 1 compared with 5 other genes previously reported by others to be overexpressed in prostate cancer (HPN, LIM, HOXC6, EZH2 and AMACR). Each box represents a normalized gene expression coefficient; red means overexpressed; blue means underexpressed. doi: 10.3834/uij.1944-5784.2009.08.06f1



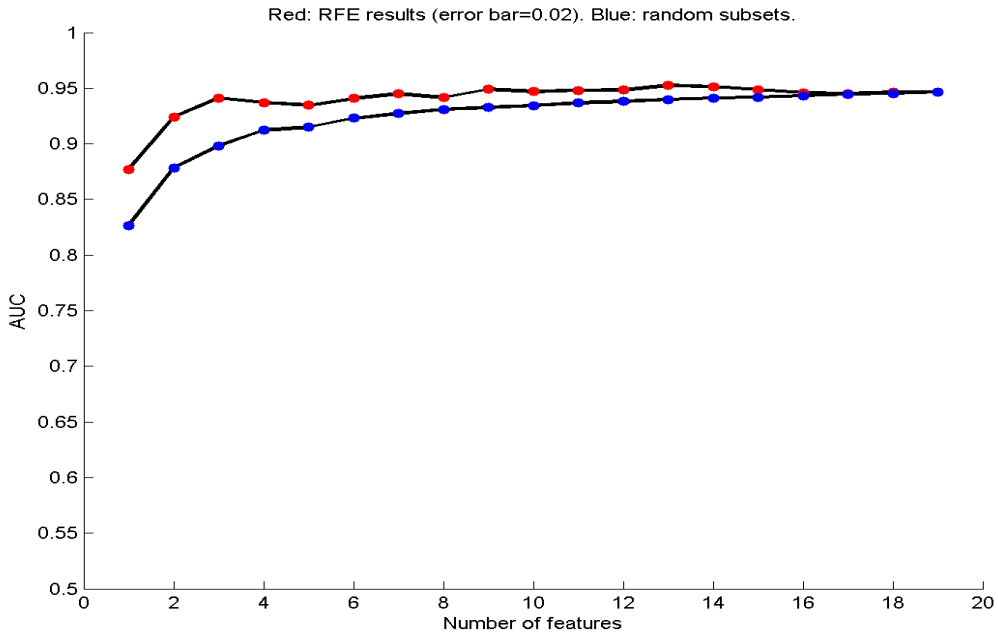
the genes selected with this criterion was assessed with the Wilcoxon-Mann-Whitney test, from which the authors obtained a *P* value. The fraction of insignificant genes in the top *r* ranked genes or *false discovery rate* was estimated with FDR ~ pvalue. n_0/r, where n_0 is the total number of genes under consideration [15]. Only genes overexpressed in cancer with FDR ≤ 10^{-5} were retained for further analysis.

Step 2. Multivariate analysis. Using the genes retained in step 1, a smaller subset of complementary genes was selected by multivariate analysis. Recursive feature elimination (RFE) [8] was carried out on discovery data using as selection criterion the magnitude of the weights of a regularized linear classifier, similar to a support vector machine (SVM) [7]. In this application, *features* or *variables* are gene expression coefficients. This procedure results in nested subsets of genes, each of which is associated with a multivariate classifier performing a linear combination of gene expression coefficients to obtain a *discriminant value*. A threshold is set on that value to decide whether a sample is cancer or noncancer. The predictive power of the gene subsets was then evaluated with the AUC criterion (similarly as in step 1), but computed for the multivariate discriminant value rather than for single gene expression coefficients. The evaluation was done using the independent microarray validation data (Oncomine repository, Table 2). A subset of genes with high predictive power was selected to comprise the diagnostic gene signature.

Step 3. RT-PCR validation. The RT-PCR data were used to evaluate the accuracy of the gene signature for tissue classification. The testing was done in a blinded manner. The tissues were classified using a simple average of the log expression values of the chosen genes normalized by B2M expression, without knowledge of the tissue categories. Confidence intervals (CI) for the sensitivity (at 90% specificity) and specificity (at 90% sensitivity) were computed using the adjusted Wald method [16].[1] After the release of the class labels, 10 times ten-fold cross-validation experiments were carried out to evaluate the potential benefit of retraining a classifier with the RT-PCR data,

[1] For a calculator, see http://www.measuringusability.com/wald.htm.

Figure 2. Performance on the Independent Validation Set as a Function of Number of Genes for 19 Genes Overexpressed in Cancer. doi: 10.3834/uij.1944-5784.2009.08.06f2



rather than using the simple average of the expression values as a prediction score. Finally, the prediction score was mapped to a probability using logistic regression [17]. The authors used the Matlab® statistics toolbox.

RESULTS

Development of the 4-Gene Signature

In step 1, using the discovery data from the 87 cell preparations (Table 1), the univariate AUC gene ranking method selected 63 genes with an AUC ≥ 0.84 and a false discovery rate of less than 10^{-5}. Of those, the authors retained $n_1 = 19$ genes that were overexpressed in cancer (Figure 1). Multivariate analysis was then carried out (step 2). The RFE method [8] produced nested subsets of genes of decreasing numbers and, for each subset of genes, a corresponding predictor for the classification of tissues into cancer vs noncancer. Performance was evaluated with the AUC using validation data. As validation data, the authors used the 164 samples from the Oncomine data (Table 2). In Figure 2, the authors plotted the performance of all the predictors obtained by the RFE method, as a function of the size of the gene subset (red markers). According to these results, an AUC of 0.94 ± 0.02 is reached with only 2 to 4 genes. Even though 2 genes appear to be sufficient to reach the best results analysis, unknown sources of variability may degrade performance when moving from microarray to RT-PCR data. Therefore, the authors decided to use 4 genes for the RT-PCR validation (Table 4). For comparison, the authors also plotted the average performance of 50 classifiers trained on subsets of genes of the same size drawn at random among the $n_1 = 19$ preselected genes (blue markers). The curve indicates that an AUC value of > 0.9 is achieved on average with subsets of 4 randomly selected genes. This increased the authors' confidence that 4 genes should suffice for the gene signature. Figure 3 shows the individual ROC curves of the four genes selected and the ROC curve for the classifier based on all 4 genes, estimated with the validation data.

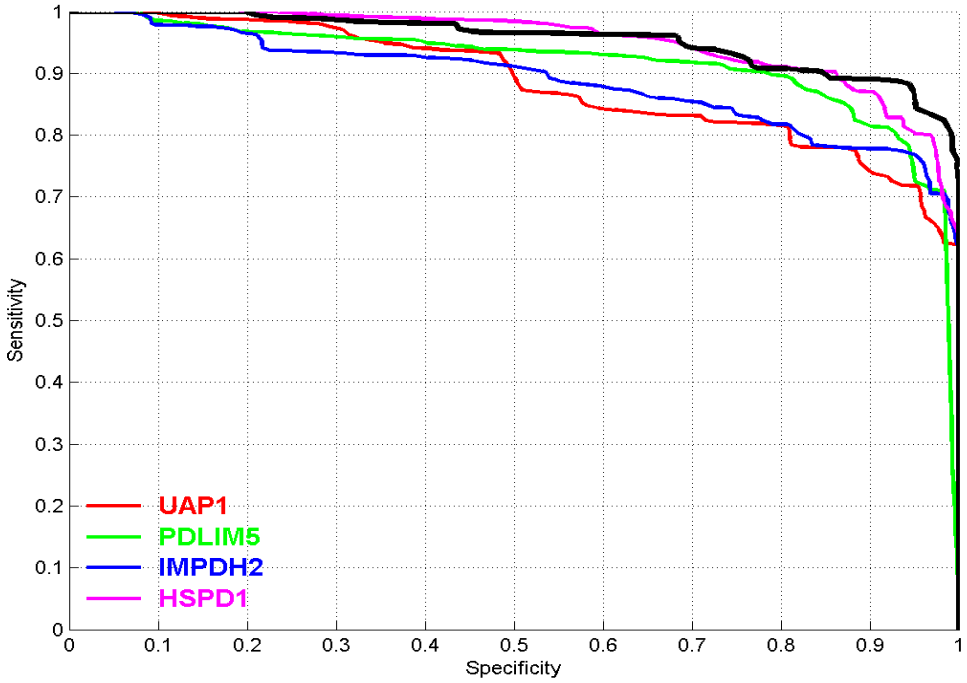
Development and Assessment of the 4-Gene RT-PCR Test

The 4-gene molecular assay was developed using the gene expression values of UAP1, PDLIM5, IMPDH2, and HSPD1 as measured by RT-PCR. Normalization of the gene expression data was accomplished with the expression of the gene B2M. It was selected from the 5 housekeeping genes evaluated because of its high signal and low variance. For prediction, the authors used a simple average of the normalized expression values of the 4 selected genes. The equation is:

$$S = \ln(HSPD1/B2M)+\ln(IMPDH2/B2M)+\ln(PDLIM5/B2M)+\ln(UAP1/B2M)+b$$

UroToday International Journal®
original study
Isabelle Guyon, Herbert A. Fritsche, Paul Choppa, Li-Ying Yang, Stephen D. Barnhill
www.urotodayinternationaljournal.com

Figure 3. **Affymetrix platform**. ROC curves on validation data (Oncomine) for the individual genes selected (Panel) and the four gene signature depicted in black. doi: 10.3834/uij.1944-5784.2009.08.06f3



The tissues were classified according to the sign of a prediction score, where a positive value indicates cancer and a negative value indicates noncancer tissue. In the course of the study, the authors received the data in consecutive phases. They performed blind tests by adjusting the bias value b of the prediction score on data received previously and making predictions on new data, not knowing in advance the identity of the tissues. Using phase 1 data to adjust b, followed by testing on phase 2 data, only 2 tissues were misclassified. Similarly, by adjusting the bias on the data of the two first phases and testing on the last one, only 2 tissues were misclassified.

After the identity of the tissues was revealed, the authors performed 10 times ten-fold cross-validation experiments to compare various classification techniques including Support Vector Machines (SVM) [7]. They found no statistically significant performance differences, so they decided to use the simplest model: the prediction score S performing a simple average of normalized log expression values.

Varying the bias b on the prediction score S allowed the authors to monitor the trade-off between the sensitivity (fraction of cancer tissue well classified) and the specificity (fraction of control tissues well classified). Figure 4 contains a plot of sensitivity vs specificity (ROC curve) for all RT-PCR samples (n=71) used as test examples. The diagnostic molecular signature achieved an area under the ROC curve AUC=0.97. The authors singled out 2 points of interest on the curve: specificity at 90% sensitivity and sensitivity at 90% specificity, for which they obtained 97% specificity (86%-100%, 95% CI) and 97% sensitivity (83%-100%, 95% CI), respectively.

Although the sign of the prediction score S allowed the authors to classify samples into cancer vs noncancer, its magnitude further informed them of the confidence with which this classification was performed. For ease of interpretation of S as a confidence, it can be mapped to a score between 0 and 1 providing an estimate of the probability that the sample is cancer:

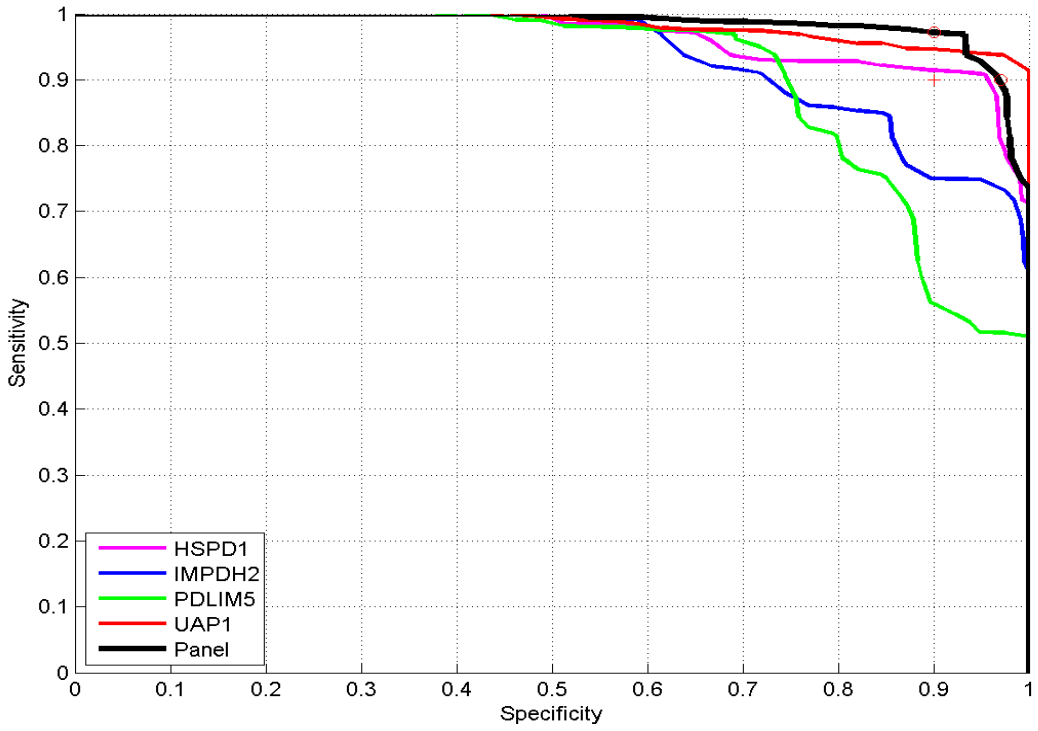
$P(\text{cancer}) = 1 / (1+\exp(-aS))$.

Using logistic regression [17], the authors obtained the following estimates for the parameters a and b: a=2.53 and b=5.94. These can be readjusted as more data become available.

DISCUSSION

The authors analyzed a microarray dataset of over 20,000 genes and identified a molecular signature of 4 genes, which are highly overexpressed in grade 3 and grade 4 prostate cancer cells when compared with noncancer prostate cells and BPH tissue. The discriminative power of this gene signature was

Figure 4. **RT-PCR platform**. ROC curves on test data for the individual genes selected and signature (Panel). AUC=0.97; (sensitivity, specificity): (0.9, 0.97), (0.97, 0.9). doi: 10.3834/uij.1944-5784.2009.08.06f4



validated on an independent microarray database (Oncomine). The gene signature retained its high predictive accuracy when assessed by real-time RT-PCR assays.

Prostate cancer is a complex disease involving many gene pathways, which include apoptosis, cellular proliferation, inflammation, and angiogenesis. Hence, the authors expected that a set of complementary genes might include genes from different pathways, related to the male reproduction system and to cancer mechanisms. The 4 genes identified in the present study are representative of cancer-associated pathways (Table 5). UAP1 is a sperm-associated antigen involved in aminosugar metabolism and is associated with androgen response [18], male infertility [19], and cancer [20]. PDLIM5 is part of several signaling pathways including protein kinase C (PKC) and was found overexpressed in prostate cancer [19,21,22]. IMPDH2 is part of the guanine nucleotide biosynthesis pathway and is related to apoptosis [23]. HSPD1 is heat shock protein (chaperonin) involved in protein folding and apoptosis [24]. Although much work remains in order to fully characterize the role of these 4 genes in prostate cancer, these literature reports give credence to their potential role in the development of prostate cancer. The present data demonstrate their diagnostic application in this gene signature test for the detection of prostate cancer cells in biopsy and prostatectomy tissues. Additional studies are underway to assess the gene expression in prostatic intraepithelial neoplasia (PIN) and high grade PIN, as well as the stromal component surrounding the cancer foci. Other studies are underway to assess the potential diagnostic utility of this 4-gene signature in prostate cells that are released into urine.

Table 5. Information on the 4 Genes. doi: 10.3834/uij.1944-5784.2009.08.06t5

Name	UAP1	PDLIM5	IMPDH2	HSPD1
Unigene	Hs.21293 Hs.492859	Hs.7780 Hs.480311	Hs.75432 Hs.476231	Hs.79037 Hs.632539
Chromosome	1q23.3	4,527 cR	3p21.2	2q33.1
Genbank	S73498	AL049969.1	J04208	BC002676.1
EC #	2.7.7.23	NA	1.1.1.205	NA
Function	Aminosugar metabolism. Sperm associated antigen.	Part of several signaling pathways including PKC	Guanine nucleotide synthesis. Related to apoptosis.	Heat shock protein (chaperonin). Protein folding, apoptosis.

Abbreviation: PKC, protein kinase C

ACKNOWLEDGEMENTS

This work would not have been possible without the contribution of many people who collected, prepared, and processed the prostate cancer samples. For the discovery data, we are very grateful to Thomas A. Stamey, John McNeal, and Rosalie Nolley for their thorough selection and preparation of samples, and to Janet Warrington and her team at Affymetrix for processing the samples. Paul Choppa's team at Clarient is gratefully acknowledged for preparing the RT-PCR data.

Conflict of Interest

Isabelle Guyon: Paid consultant to sponsor; Board membership with sponsor; Equity ownership/stock holder for mentioned product; Patent inventor for mentioned product.
Herbert A. Fritsche: Board membership with sponsor.
Paul Choppa: Employed by Clarient.
Li-Ying Yang: No conflict.
Stephen D. Barnhill: Equity ownership/stock holder for mentioned product.

REFERENCES

[1] Jemal A, Siegel R, Ward E, et al. Cancer statistics 2008. *CA Cancer J Clin*. 2008;58(2):71-96.

[2] Brooks JD. Microarray analysis in prostate cancer research. *Curr Opin Urol.* 2002;12(5):395-399.

[3] Calvo A, Gonzalez-Moreno O, Yoon CY, et al. Prostate cancer and the genomic revolution: Advances using microarray analyses. *Mutat Res*. 2005;576(1-2):66-79.

[4] Elek J, Park KH, Narayanan R. Microarray-based expression profiling in prostate tumors. *In Vivo*. 2000;14(1):173-182.

[5] Saramaki OR, Porkka KP, Vessella RL, Visakorpi T. Genetic aberrations in prostate cancer by microarray analysis. *Int J Cancer*. 2006;119(6):1322-1329.

[6] Xu J, Stolk JA, Zhang X, et al. Identification of differentially expressed genes in human prostate cancer using subtraction and microarray. *Cancer Res*. 2000;60(6):1677-1682.

[7] Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY: Association for Computer Machinery; 1992; 144-152.

[8] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1):389-422.

[9] Stamey TA, Warrington JA, Caldwell MC, et al. Molecular genetic profiling of Gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia. *J Urol*. 2001;166(6):2171-2177.

[10] Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*. 2004;6(1):1-6.

[11] Singh D, Febbo P, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002;1(2):203-209.

[12] Febbo P, Sellers W. Use of expression analysis to predict outcome after radical prostatectomy. *J Urol*. 2003;170(6 Pt 2): S11–S20.

[13] LaTulippe E, Satagopan J, Smith A, et al. Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res*. 2002;62(15):4499-4506.

[14] Welsh JB, Sapinoso LM, Su AI, et al. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res*. 2001;61(16):5974-5978.

[15] Guyon I, Gunn S, Nikravesh M, Zadeh LA, eds. *Feature Extraction Foundations and Applications*. Berlin, Germany: Springer-Verlag; 2006.

[16] Agresti A, Coull B. Approximate is better than 'exact' for interval estimated binomial proportions. *Am Statistician*. 1998;52:119-126.

[17] Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York, NY: Chichester Wiley; 2000.

[18] DePrimo SE, Diehn M, Nelson JB, et al. Transcriptional programs activated by exposure of human prostate cancer cells to androgen. *Genome Biol*. 2002; 3(7): RESEARCH0032.

[19] Luo JH, Yu YP, Cieply K, et al. Gene expression analysis of prostate cancers. *Mol Carcinog*. 2002;33(1):25-35.

[20] Diekman AB, Olson G, Goldberg E. Expression of the human antigen SPAG2 in the testis and localization to the outer dense fibers in spermatozoa. *Mol Reprod Dev.* 1998;50(3):284-293.

[21] Febbo PG, Sellers WR. Use of expression analysis to predict outcome after radical prostatectomy. *J Urol*. 2003;170(6 Pt 2):S11-S20.

[22] Henshall SM, Afar DE, Hiller J, et al. Survival analysis of genome-wide gene expression profiles of prostate cancers identifies new prognostic targets of disease relapse. *Cancer Res*. 2003;63(14):4196-4203.

[23] Grusch M, Rosenberger G, Fuhrmann G, et al. Benzamide riboside induces apoptosis independent of Cdc25A expression in human ovarian carcinoma N.1 cells. *Cell Death Differ*. 1999;6(8):736-744.

[24] Sarto C, Binz P-A, Mocarellli P. Heat shock proteins in human cancer. *Electrophoresis*. 2000;21:1218-1226.