New prostate cancer biomarkers validated by independent studies

Isabelle Guyon+, Thomas A. Stamey*, Herbert A. Fritsche++, and Stephen D. Barnhill+
+ Health Discovery Corporation, Savannah, Georgia
++ University of Texas, M.D. Anderson Cancer Center in Houston, Texas
* Department of Urology at Stanford University School of Medicine

Summary
With our analytical tools we have analyzed **microarray data** with over **20,000 genes**. We have discovered **molecular patterns characteristic** of **zonal** and **histological tissue classification** of the prostate. We have found promising **prostate cancer biomarkers** yielding **90% sensitivity and specificity** on completely independently collected data. We have also found promising **diagnostic markers** and **new drug targets** for **BPH**.

1 . Background

Prostate cancer is a deadly disease…
There are an estimated 300,000 new cases and 41,000 deaths from prostate cancer each year in the US alone.

… but most patients do not die of it!
Clinical studies indicate that 50% of men over age 50 and 70% of men over age 70 have prostate cancer. However, of all men diagnosed with prostate cancer during their lifetimes only 3% die of Prostate Cancer!

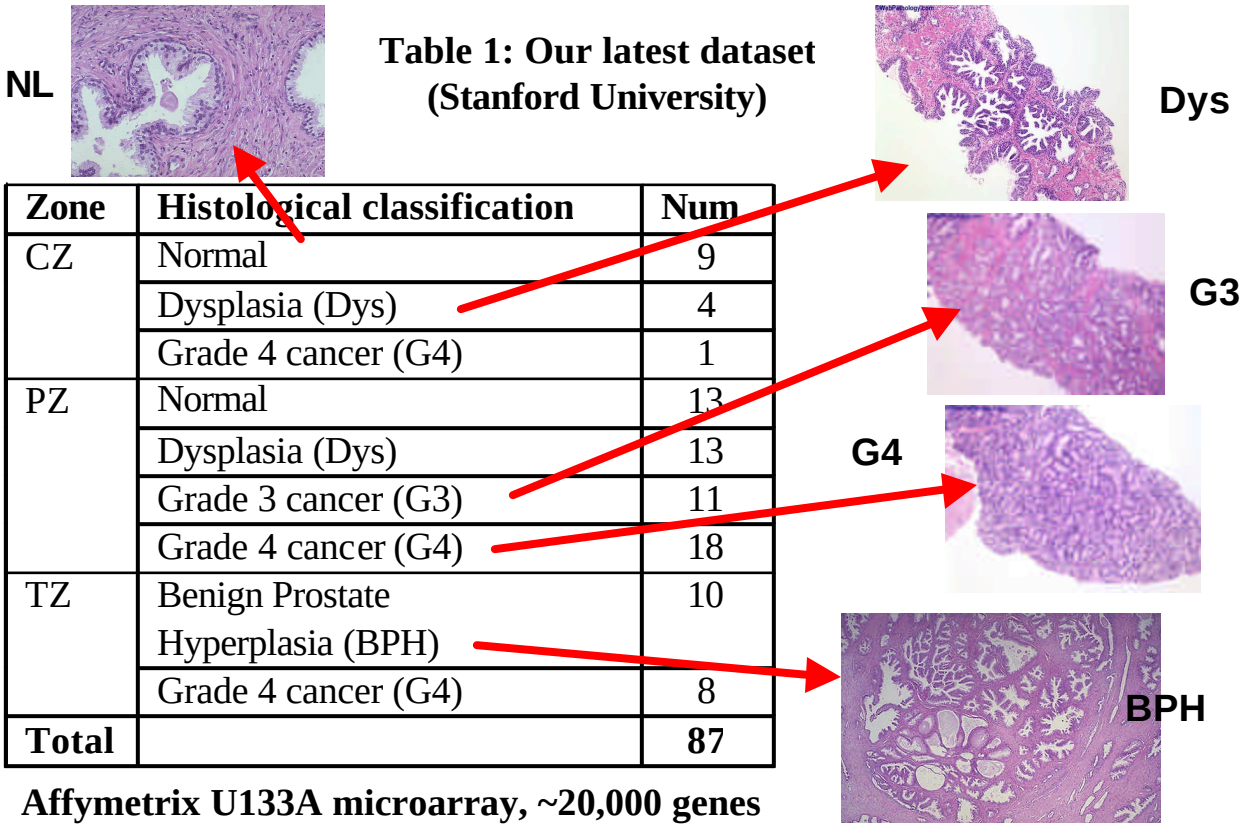
The era of PSA is over
The standard blood test measuring the level of prostate specific antigen (PSA) has recently been shown to reveal mostly benign prostate hyperplasia (BPH), a condition that is not life threatening, NOT Prostate Cancer. Current serum PSA based cancer detection strategies result in an unacceptable, high false positive rate, this leads to **many unnecessary biopsies** (only ¼ of biopsies are positive for cancer) and result in over-diagnosis of prostate cancers, which do not need to be treated. This leads to **many unnecessary prostatectomies.** In this contrext, there is a critical need for a PSA replacement that can accurately discriminate between latent, non-aggressive cancer and clinically significant Prostate Cancer. The results of our analysis are important steps towards that end.

Zones and grades
The prostate is traditionally though of being composed of three zones: the peripheral zone (PZ) has the largest number cancers (60-80%), and the worst prognosis. The central zone CZ has only

Table 1: Our latest dataset (Stanford University)

Zone	Histological classification	Num
CZ	Normal	9
	Dysplasia (Dys)	4
	Grade 4 cancer (G4)	1
PZ	Normal	13
	Dysplasia (Dys)	13
	Grade 3 cancer (G3)	11
	Grade 4 cancer (G4)	18
TZ	Benign Prostate Hyperplasia (BPH)	10
	Grade 4 cancer (G4)	8
Total		**87**

Affymetrix U133A microarray, ~20,000 genes



5-20% of the cancers. The transitional zone (TZ) is essentially composed of tissue growing in the aging man known as benign prostate hyperplasia or BPH. Only10-18% of the cancers occur in the TZ. Healthy and malignant prostate tissues can be recognized by histological staining and classified into healthy, BPH, or cancer. Several grades of cancers can also be distinguished, ranging from 1 to 5. Curability is inversely related to the fraction of cancer tissue of grade 4 or 5. Dysplasia is a tissue anomaly thought of being a precursor of cancer. Our study takes into account **zonal and histological tissue classification**.

2 . Material and Methods

Three elements are key to success in biomarker discover:

– **Samples:** the availability of good samples with accurate clinical annotations.
– **Measurements:** a sensitive measuring assay, good experimental design and sample processing.
– **Analysis:** the use of state-of-the art analytical biomarker discovery software.

With our patented analytical tools, which have proven to be superior in several studies, we have analyzed high quality microarray data from Stanford University and other sources.

An exceptional and challenging dataset

We performed our biomarker discovery with a dataset of 87 prostate tissues of different zones and histological grades. The data, provided by the team of Prof. Stamey at Stanford University (Table 1), is unique in the sense that it contains tissues that are **laser microdissected** and

carefully labeled. The tissues coming from frozen sections of the prostates of patients having undergone prostatectomy were analyzed with an Affymetrix U133A microarray, which reveals the gene expression of over **20,000 genes**.

Validation data

In addition we have had access to data from a previous Stanford study in which 67 samples were analyzed with the HuGeneFL Affymetrix array (Table 2), and publicly available data from Oncomine (Table 3.) The first Stamey study recorded only the expression of ~6500 genes, but had zonal and histological annotations. The Oncomine data recoded ~12500 genes and do not have zonal and histological annotations. We use these additional data for validation purpose.

Table 2: First Stanford study.

Zones	**Histology**	**Num**
CZ	Normal	3
PZ	Normal	5
	Stroma	1
	Dysplasia	3
	Grade 3	10
	Grade 4	27
TZ	BPH	18
Total		**67**

HuGeneFL microarray ~6500 genes

Table 3: Oncomine data.

Source	**Histology**	**Num**
Febbo [1]	Normal	50
	Tumor	52
LaTulippe [2]	Normal	3
	Tumor	23
Welsh [3]	Normal	9
	Tumor	27
Total		**164**

HuGeneFL microarray ~6500 genes

Analytical methods

The gene expression coefficients are ranked with the Area under the ROC curve (AUC) of **individual genes** to select genes most characteristic of disease tissues (cancer of BPH.) The AUC is the area under the curve plotting sensitivity *v.s.* specificity. The sensitivity is defined as the rate of successful disease tissue classification and the specificity is the rate of successful control tissue classification. Thus our ranking method allows us to assess the **classification power** of individual genes. The **statistical significance** of the genes selected with this criterion can be assessed with the Wilcoxon-Mann-Withney test and a "false discovery rate", which estimates the fraction of insignificant genes, can thus be computed. Visualization of the tissue samples as scatter plots in the space generated by two relevant genes allows us to verify **zonal and histological grouping.**

The validity of the genes is further challenged with the validation data. For this purpose, we must match the probes of the various arrays and normalize the data. We **matched the genes** in the different arrays by probe sequence, gene ID, and gene description (Stanford new & old: 2346 matched genes; Stanford new & Oncomine: 6839 matched genes.) We then normalized the data matrices **iteratively** in lines and columns. We then used **two validation methods**:

- Gene set "enrichment": We selected the top 50 ranking genes of the discovery dataset and examined how fast they are encountered in the ranking performed with the validation dataset.
- Classification prediction: We selected genes and trained a classifier using the discovery dataset; we computed the classification accuracy using the

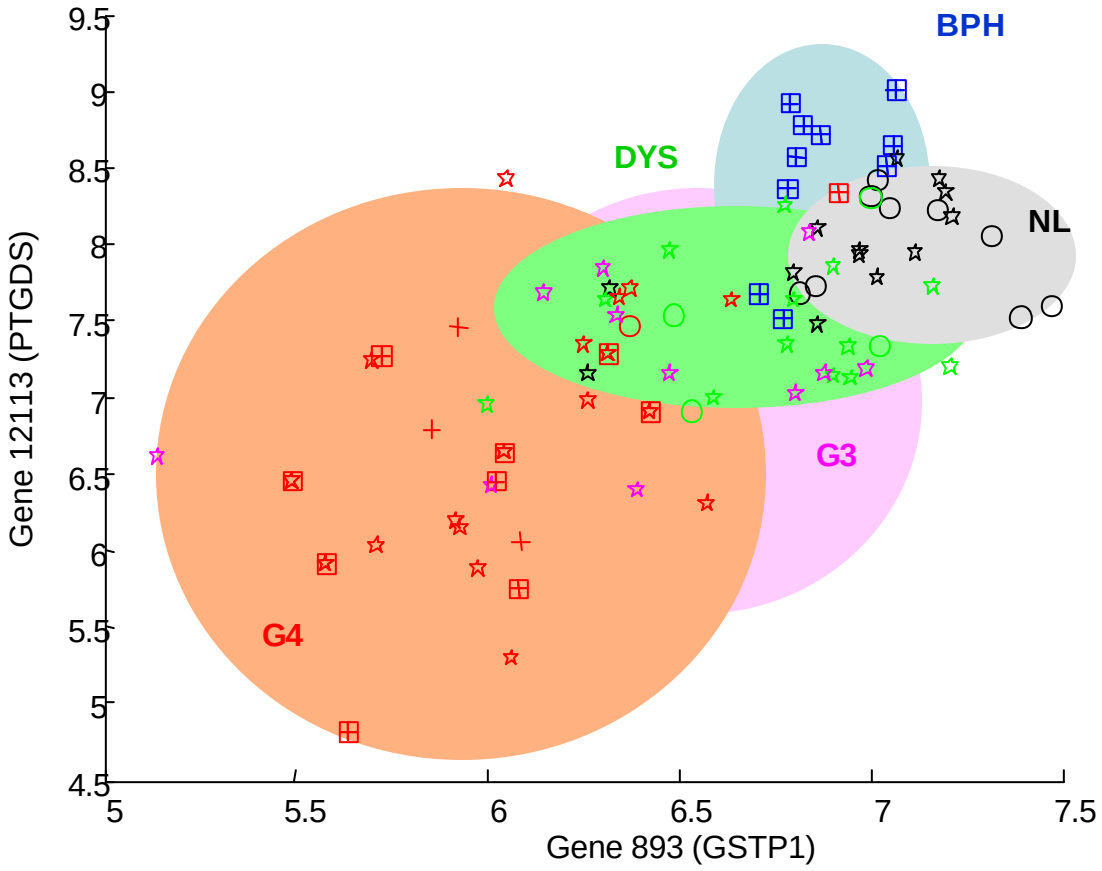


Figure 1: Genes characteristic of cancer.

validation dataset. We use regularized linear classifiers analogous to **Support Vector Machines** (SVM.)

As a last verification, we swapped the discovery and validation datasets and performed again the same analyses. Finally, we pooled the samples from several studies and perform a consensus gene ranking.

3. Results

Molecular signatures

In one experiment, we separated our U133A samples from the latest Stanford study into a discovery set and a test set. The test set consisted of 10 BPH samples and 10 cancer samples.[1] The normalized expression levels of the samples for the two genes separating best cancer from control are shown in Figure 1. The colors code for the histological classification (Red=G4, Pink=G3, Green=Dysplasia, Blue=BPH, Black=Normal.) The symbols indicate the zones (+=TZ, o=CZ, *=PZ.) Symbols enclosed in a square indicate the test samples, not used to select the genes. Remarkably, those two genes group well the samples by histological classes. Furthermore, they seem to indicate disease progression. Finally, the test samples are well classified and the BPH samples (which are all test samples) group in a cluster that is distinct from other control samples.

[1] These samples having all been processed at a later date than the other samples, we wanted to verify that no experimental artifact had been introduced.

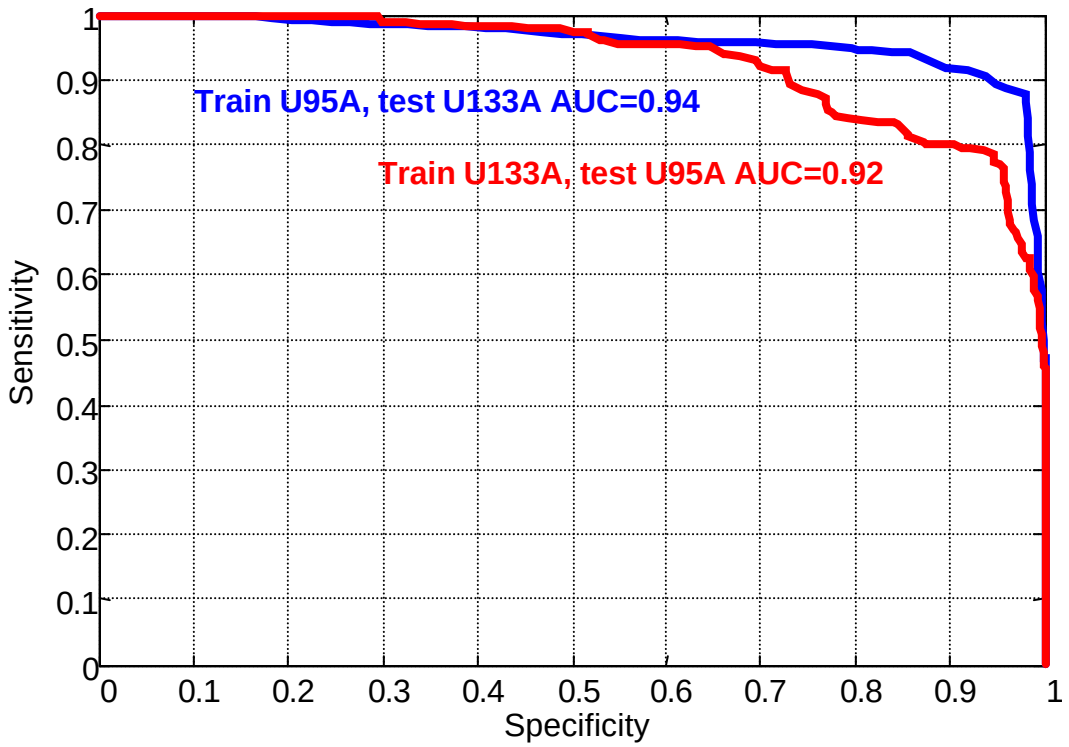


Figure 2: Classification accuracy validation. Ten genes are selected and used for training with one dataset. The other is used for testing.

We also found genes that are specific of BPH, *i.e.* which separate well BPH from all other histological classes (not shown.)

Biomarker validation

Using the statistical methods described, we estimated the false discovery rate of the genes most discriminative of cancer tissues to be less than 0.3%. For the genes most discriminative of BPH, it is no more than 0.2%.

Genes selected to separate tumor from controls in the discovery dataset (U133A Stanford data) are found to be also discriminative in the validation dataset (Oncomine data) and vice versa, with **AUCs above 0.9** (Figure 2.) We performed a similar validation for the BPH genes using the two Stanford studies; we also found AUCs larger than 0.9. In either case, only **ten genes** were necessary to achieve these results.

We also verified that the gene ranking obtained with one data set is enriched in genes selected from the other dataset. Approximately half of the top 50 genes of one study are found in the top 200 ranking genes of the other study.

Consensus ranking

Having validated the data, we pooled the samples of several studies to produce a consensus ranking. Hundred of genes are found significant and are candidate biomarkers. To illustrate the results, we show in Figure 3 the heat map of gene expression coefficients of the top 50 genes most discriminative of BPH in the two Stanford studies.

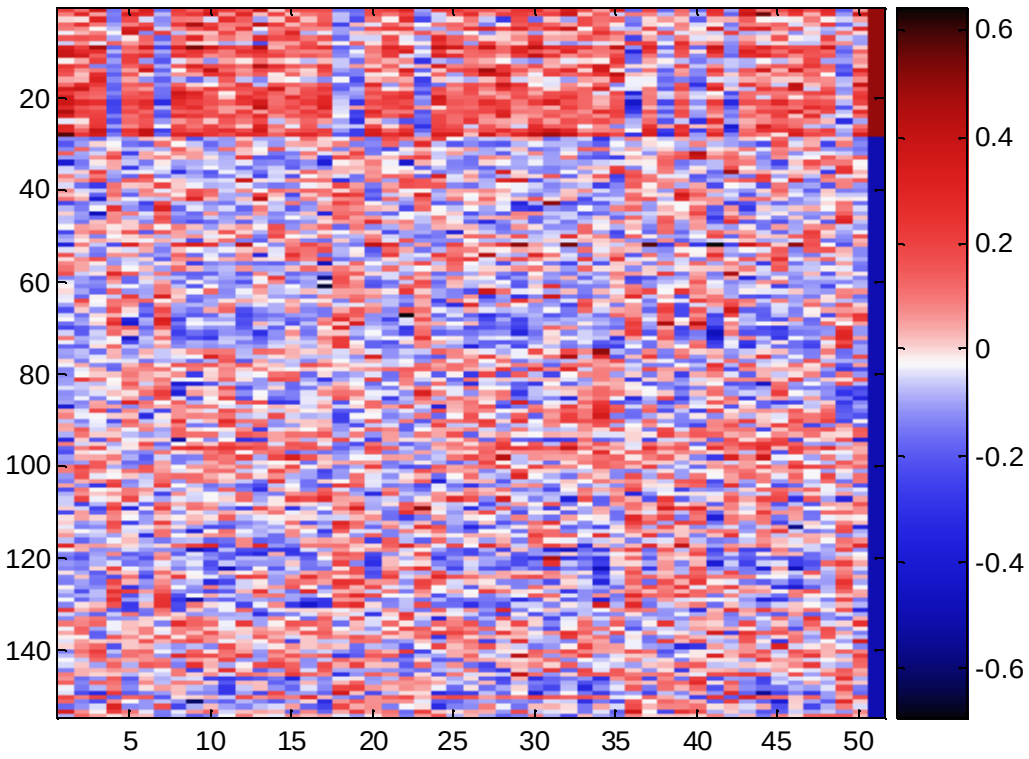


Figure 3: Heat map of the gene expression coefficients from the samples of the two Stanford studies. The lines of the matrix represent the samples. We show the 50 genes most discriminative of BPH (columns.) The last column indicates the tissue category (red at the top=BPH, blue at the bottom=others.)

Development of serum assays

Having, identified genes in prostate tissue, which show the presence of cancer or BPH, we shall now select those genes, which will **most likely be present in blood**. From our initial work, **hundreds of genes are suitable** for this analysis and we know that as few as **ten genes suffice** to get high classification accuracy. Analytical techniques such as Recursive Feature Elimination Support Vector Machine (RFE SVMs) may be instrumental in optimizing the final gene set. Partnering with Proteomics companies, we will develop serum assays for those gene products and perform clinical studies to validate the biomarkers as a prostate cancer and BPH tests.

[1a] Gene expression correlates of clinical prostate cancer behavior Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, D'Amico A, Richie J, Lander E, Loda M, Kantoff P, Golub T, Sellers W. Cancer Cell. 2002 Mar;1(2):203-9.

[1b] Use of expression analysis to predict outcome after radical prostatectomy , Phillip G. Febbo and William R. Sellers, The journal of urology, Vol. 170, S11–S20, December 2003.

[2] Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. LaTulippe E, Satagopan J, Smith A, Scher H, Scardino P, Reuter V, Gerald WL. Cancer Res. 2002 Aug 1;62(15):4499-506.

[3] Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer.
Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF Jr, Hampton GM.
Cancer Res. 2001 Aug 15;61(16):5974-8.