SUPPLEMENT

Contents

1. Extended description of the Data Sets.

Datasets 1 and 2 (**Table 1**) are based on post-prostatectomy frozen tissue samples obtained by informed consent using IRB-approved and HIPPA-compliant protocols. All tissues, except where noted, were collected at surgery and escorted to pathology for expedited review, dissection, and snap freezing in liquid nitrogen. RNA for expression analysis was prepared directly from frozen tissue following dissection of OCT (optimum cutting temperature compound) blocks with the aid of a cryostat. For expression analysis 50 micrograms (10 micrograms for biopsy tissue) of total RNA samples were processed for hybridization to Affymetrix GeneChips (www.affymetrix.com).

Dataset 1 contains expression data from multiple sources. First, there were 109 post prostatectomy frozen tissue samples from 87 patients of the UCI/NIH SPECS (Strategic Partners for the Evaluation of Cancer Signatures) project (**http://www.pathology.uci.edu/faculty/mercola/UCISPECSHome.html**).

. Two different types of tissue samples were analyzed from 22 of these 87 patients; one sample type was enriched for tumor. The other sample type contained stroma from cases of prostate cancer, but with this stroma generally located more than 15 mm from the tumor, and usually in the contralateral lobe. Second, there were 27 prostate biopsy specimens obtained as fresh snap frozen biopsy cores from 18 normal prostates. These samples were obtained from the control untreated subjects of a clinical trial to evaluate the role of Difluoromethylornithine (DFMO) to decrease the prostate size of normal men. Ten of these were collected before the treatment period, and eight were collected after the treatment period had ended (3). Third, 13 prostates from non-prostate cancer donors were obtained from the rapid autopsy program of the Sun Health Research Institute, a consortium member of the UCI SPECS program with an average patient age of 82 years.

Dataset 2 contains expression data from 136 samples from 82 patients of the UCI SPECS prostate program. Expression data from 65 samples that consisted predominately of tumor was used as a test dataset. 71 of the tumor-bearing samples were manually microdissected to obtain tumor-adjacent stroma which was used for validation of the Diagnostic Classifier. For manual microdissection, the tumor-bearing tissue was embedded in an OCT (optimum cutting temperature compound, Fisher Scientific Inc.) block then mounted in

a cryostat. Frozen sections were stained using hematoxylin and eosin (H and E) to visualize the location of the tumor. Then the OCT-embedded block was etched with a single straight cut with a scalpel to divide the embedded tissue into a tumor zone and tumor-adjacent stroma. Subsequent cryosections produced two halves at the site of the etched cut and were separately used for H and E staining and examined to confirm their composition. Multiple frozen sections of the tumor-adjacent stroma were then pooled and used for RNA preparation and microarray hybridization. A final frozen section was used for H and E staining and examined to confirm that it was free of tumor cells. The pooled tumor-adjacent stroma was then used for RNA preparation and expression analysis.

Expression data for Datasets 1 and 2 are publicly available in the GEO database (http://www.ncbi.nlm.nih.gov/geo) with accession number GSE17951 (Dataset 1) and GSE8218 (Dataset 2). For Datasets 1 and 2, the distributions for the four principal cell types (tumor epithelial cells, stroma cells, epithelial cells of BPH, and epithelial cells of dilated cystic glands) were estimated as follows. A frozen section was taken immediately above the sections pooled for RNA preparation and again immediately below the pooled sections. Each of these extra sections was reviewed by three pathologists (Dataset 1) or four (Dataset 2), whose estimates were averaged as described (4). The distributions of tumor percentage for Dataset 1 and 2 are shown in **Figure 1(a)** and **1(b)**.

Dataset 3 consists of a series of 79 samples (6) of the UCI SPECS prostate program. Relative transcript levels were measured with Affymetrix U133A chips. The cell composition was not documented at the time when the data were collected. Cell composition was estimated by use of multigene signatures that are invariant with tumor surgical pathology parameters of Gleason and stage by the CellPred program(7) which confirmed that all 79 samples bear tumor, with tumor content ranging from 24% to 87% (**Figure 1(c)**).

Dataset 4 (8) is composed of 57 samples from 44 patients of the UCI SPECS prostate program, including 13 samples of stroma near tumor and 44 tumor-bearing samples. Tumor percentage (ranging from 0% to 80%, **Figure 1(d))** was approximated by using the CellPred program.(7)

The U133Plus2 platform used for Dataset 1 has about 30,000 probe sets whereas the U133A used for Datasets 2, 3 and 4, contains 22,000 of these probe sets. Normalization was carried out across multiple

datasets using the ~22,000 probe sets in common to all Datasets. First, Dataset 1 was quantile-normalized using function 'normalizeQuantiles' of LIMMA routine (9). Datasets 2 - 4 were then quantile-normalized by referencing normalized Dataset 1 with a modified function 'REFnormalizeQuantiles' which was coded by ZJ and is available at the SPECS website (http://www.pathology.uci.edu/faculty/mercola/UCISPECSHome.html).

2. Extended description of the multiple linear regression (MLR) model and epithelial genes filtration.

A multiple linear regression (MLR) model was used to describe the observed Affymetrix intensity of a gene as the summation of the contributions from different types of cells given the pathological cell constitution data:

$$g = \beta_0 + \sum_{j=1}^{C} \beta_j p_j + e, \qquad (1)$$

where g is the expression value for a gene, $p's$ are the percentage data determined by the pathologists, and $\beta's$ are the expression coefficients associated with different cell types. In equation (1), C is the number of tissue types under consideration. In the current study, three major tissue types were included, *i.e.*, tumor, stroma and BPH. β_j is the estimate of the relative expression level in cell type j (*i.e.* the expression coefficient) compared to the overall mean expression level β_0. The regression model was applied to the patient cases in Dataset 1 to obtain the model parameters ($\beta's$) and their corresponding p-values, which were then used to aid subsequent gene screening. The application to prostate cancer expression data and validation by immunohistochemistry and by correlation of derive β_j values with LCM-derived samples assayed by qPCR has been described (4).

We used the cell-type specific expression coefficients ($\beta's$) to identify genes that are largely expressed in stroma using three criteria: 1. Genes that are expressed at the RNA level in tumor epithelial cells at greater than 10% of the expression level in stroma cells, *i.e.*, $\beta_S > 10 \times \beta_T$, where β_T and β_S are defined by Equation (1) above. 2. $\beta_S > 0$. 3. $p(\beta_S) < 0.1$. Criteria 2 and 3 select genes that are significantly expressed in stroma cells. In the MLR model, criterion 3 has two implications: either the gene is expressed in stroma cells but not in tumor cells ($\beta_S > 0$ and $\beta_T < 0$) and is retained or the gene is expressed in both stroma cells and tumor cells ($\beta_S > 0$ and $\beta_T > 0$) but is only retained if ($\beta_S > 10 \times \beta_T$).

3. Extended description of the comparison with random-gene classifiers.

To further validate our 131-probe set Diagnostic Classifier, we carried out 100 randomized experiments using the 22,277 probe sets of the U133A platform which was used for the original 13 tumor-bearing training cases. In each experiment we randomly selected 2,210 probe sets from the 12,901 probe sets remaining after subtracting 9376 aging related probe sets from the entire 22,277 probe sets. The remaining probe sets were screened with the same MLR criteria used for the development of the 131-probe set classifier, *i.e.*, 1. $\beta_S > 0$, 2. $p(\beta_S) < 0.1$ and 3. $\beta_S > 10 \times \beta_T$. The genes that survived MLR filter were used to develop a classifier with PAM exactly as for the 131-probe set classifier. PAM selected an average of 6.2 (standard deviation = 2.3) probe sets (<< 131) and the average performance of these random-gene classifiers based on the tests of other datasets is summarized in **Table S1.**

The random classifiers were biased towards calling almost all samples as being normal, leading to a statistically significant under-calling of sets of tumor samples, (*e.g.* **Table S1**, line 2) and a statistically significant "success" in calling normal samples (*e.g.* **Table S1**, lines 6-8 and 13). The overall accuracy was around 35%, no different from random accuracy, indicating that the results obtained with the 131-probe set classifier cannot be attributed to chance.

4. Extended description of the comparison with Dakhova et al. (1)and Richardson et al. (2)

Two recent studies describe expression analysis results for subclasses of the stroma of prostate cancer (1, 2). In one study (2), 44 (39 unique) genes were identified as differentially expressed between intratumor stroma and normal stroma using Affymetrix U133 Plus 2.0 GeneChips on five paired LCM intratumor stroma and matched normal stroma. The microarray data from this study is not publicly available and a detailed comparison is not possible. Several of the 44 genes were recognized as differentially expressed in our analysis; however, none survived the age and tumor epithelial cell expression filters applied here. In another study (1), Agilent 44K gene expression arrays were used on 17 paired laser-captured microdissected (LCM) reactive stroma samples and matched normal stroma samples. 1,141 genes were identified as differentially expressed between "reactive" and normal stroma. Reactive stroma has been studied in detail (1, 2) and is a form of stroma very near to tumors which differs from normal stroma in histological appearance, cell composition and gene and protein expression. Prostate cancer cases with defined reactive stroma exhibit a significantly decreased postprostatectomy disease-free survival (10). We downloaded the raw microarray Dataset from public Gene Expression Omnibus database (accession SE11682) in order to assess the agreement with the 339 probe set basis set with the data of Dakhova *et al.* (1). The 339 probe sets were mapped to 557 genes on the Agilent array, out of which 38 genes were among the 2967 Agilent genes that exhibited significant differential expression ($p < 0.05$). Thirty one of 38 showed concordance in differential expression between the two studies (**Table S2**). Additional similarities were likely to have been masked by platform-specific effects (Affymetrix versus Agilent). This overlap of 31 concordant genes between the two lists of 339 and 557 genes exceeds that expected by chance alone ($\boldsymbol{p} = 0.0001$, see **Table S2**). As expected, these genes alone successfully categorize the cases of Dakhova *et al.* (1) into reactive and normal stroma cases (**Figure S4**).

The differences in the genes identified in the two studies may be at least partly explained by the fact that our study was designed to identify as many changes as possible that were common to all stroma in the presence of tumor. In contrast Dakhova *et al.* (1) used only "reactive" stroma as defined by the Masson's trichrome staining pattern. Tumors exhibiting the reactive stroma pattern have been associated with poor

postprostatectomy disease-free survival (11). The overlap in the lists may include expression changes that occur in reactive stroma, thereby strengthening the PAM classifier for samples near the poor prognosis tumors. Thus, the overlapping genes of our Diagnostic Classifier also likely have prognostic significance (*cf*. **Figure S4**).

5. Extended description of RNA preparation and the XP-PCR protocol.

RNA preparation. RNA preparation. All microarray hybridizations performed here were carried out using Affymetrix U133 plus 2.0 array expression arrays (Santa Clara, CA) including 54,675 probe sets to analyze >47,000 transcripts, including 38,500 well-characterized human genes. Isolated total RNA samples were processed as recommended by the manufacturer. In brief, total RNA was prepared from frozen tissue sections by direct dissolution in the reagents of the RNeasy (Qiagen, Chatsworth, CA) followed by passage of the RNeasy spin column. Eluted total RNAs were quantified with a portion of the recovered total RNA and adjusted to a final concentration of 1.25 mg/ml. All total RNA samples were assessed for quality by the application of a small amount of each sample (typically 25 to 250 ng/well) with a Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA). Single-stranded and then double stranded (ds) complement DNA was synthesized from the poly(A)+ mRNA present in the isolated total RNA (typically 10 mg total RNA starting material each sample reaction) using SuperScript Double-Stranded cDNA Synthesis Kits (Invitrogen, Carlsbad, CA) and poly (T)-nucleotide primers that contained a sequence recognized by T7RNA polymerase. A portion of the resulting ds-cDNA was used as a template to generate biotin-tagged complement cRNA from an in vitro transcription reaction, using Affymetrix GeneChip IVT Labeling Kit. Fifteen micrograms of the resulting biotin-tagged cRNA was fragmented to an average strand length of 100 bases (range, 35 to 200 bases) following prescribed protocols (Affymetrix GeneChip Expression Analysis Technical Manual). The product was hybridized at 45^{o} C with rotation for 16 hour (Affymetrix Hybridization Oven 640) to the expression arrays (UCI array core). The arrays were washed and then stained (streptavidin-phycoerythrin) using a Affymetrix Fluidics Station 450. The resulting array image and data were analyzed as described in the manuscript (**Materials and Methods; Results**).

XP-PCR. XP-PCR is a quantitative, multiplex RT-PCR methodology developed by AltheaDx and commercialized as the GeXP platform by Beckman Coulter (12). The amplification and detection platform allows for the expression analysis of multiple genes in a single reaction and has been applied in numerous ways including cancer diagnostics (13) and detailed compound treatment profiling (14). A defined

combination of gene specific and universal primers used in the reaction results in a series of fluorescently labeled PCR products whose size and quantity are measured using a capillary electrophoresis instrument (Beckman Coulter GeXP). For use in the quantitative analysis of FFPE, the XP-PCR technology utilizes XP-PCR optimized for Short Fragment Analysis (or XP-PCR SFA) to detect smaller RNA fragments. Primer design for AltheaDx's short-fragment multiplex RT-PCR assays involves gene specific primers that produce small amplicons (40-110nt), incorporating universal priming sequences and adding additional gene sequences to allow for size separation on a capillary electrophoresis instrument. Amplicons are kept short to maximize amplification sensitivity in short fragment RNA from typical FFPE samples. As in AltheaDx's standard quantitative, multiplex PCR assays, universal priming sequences on the 5' ends of all chimeric primers allow the PCR reaction to be primarily driven by universal primers, which are in excess, thus locking in the relative ratios of the gene targets as they are amplified. Gene specific spacer sequences extend the resulting PCR reactions into a readable range (>100nt) and create sufficient separation between amplicons (>2nt between each gene) for detection and resolution on the capillary electrophoresis instrument.

6. Table S1. Comparison of 131-probe set Diagnositic Classifier to classifiers generated from 'random' genes.

		Dataset	Case Num.	Accuracy %		Sensitivity %		Specificity %	
				i	ii	i	ii	i	ii
1	**Training set**	1	26 (13 + 13)	96.4	67.1	92.3	32.5	100	97.1
	Test set								
	Tumor								
2	Tumor-bearing	1	55 (68 - 13)	96.4	8.7	96.4	8.7	NA	NA
3	Tumor-bearing	2	65	100	12.9	100	12.9	NA	NA
4	Tumor-bearing	3	79	100	13.4	100	13.4	NA	NA
5	Tumor-bearing	4	44	100	15.9	100	15.9	NA	NA
	Normal								
6	Biopsies (5)	1	7	100	98.8	NA	NA	100	98.8
7	Biopsies (2)	1	5	60.0	100	NA	NA	60.0	100
8	Rapid autopsies	1	13	92.3	67.5	NA	NA	92.3	67.5
	Manuel Midrodissected/LCM								
9	Tumor-adjacent Stroma	2	71	97.1	13.6	97.1	13.6	NA	NA
10	Tumor-adjacent Stroma	4	13	100	15.9	100	15.9	NA	NA
11	Tumor-adjacent Stroma	1	12	75.0	5.8	75.0	5.8	NA	NA
12	Tumor-bearing	5	12	100	19.2	100	19.2	NA	NA
13	Pooled normal stroma	5	4	100	79.4	NA	NA	100	79.4

7. Table S2. Concordance of 38 overlapping genes/probe sets of the 339 probe sets (basis) of the Diagnostic Classifier with the sign of differential change of Dakhova et al. (1).

Affymetrix Probe Set ID	Agilent Probe ID	Gene Symbol	This Study	Dakova et al.
205554_s_at	25330	DNASE1L3	up	up
207332_s_at	6474	TFRC	up	up
207332_s_at	33074	TFRC	up	down
209765_at	27257	ADAM19	down	up
206331_at	41822	CALCRL	down	up
201655_s_at	19543	HSPG2	down	up
207437_at	22289	NOVA1	down	up
205954_at	40101	RXRG	down	up
210432_s_at	19493	SCN3A	down	up
219902_at	10102	BHMT2	down	down
212097_at	5848	CAV1	down	down
212097_at	8348	CAV1	down	down
212097_at	40981	CAV1	down	down
208792_s_at	32464	CLU	down	down
213428_s_at	21788	COL6A1	down	down
209015_s_at	1064	DNAJB6	down	down
218435_at	12280	DNAJC15	down	down
204410_at	32431	EIF1AY	down	down
207876_s_at	6002	FLNC	down	down
205674_x_at	29788	FXYD2	down	down
211275_s_at	43496	GYG1	down	down
205561_at	22259	KCTD17	down	down
216096_s_at	24526	NRXN1	down	down
209915_s_at	24526	NRXN1	down	down
204940_at	32154	PLN	down	down
204939_s_at	32154	PLN	down	down
203456_at	43556	PRAF2	down	down
208131_s_at	2097	PTGIS	down	down
212610_at	38709	PTPN11	down	down
208789_at	13320	PTRF	down	down
212887_at	44512	SEC23A	down	down
201312_s_at	38622	SH3BGRL	down	down
213203_at	12610	SNAPC5	down	down
213203_at	26477	SNAPC5	down	down
218087_s_at	9944	SORBS1	down	down
202440_s_at	5316	ST5	down	down
212457_at	2313	TFE3	down	down
213480_at	41809	VAMP4	down	down

8. Table S3. Function enrichment analysis.

Gene enrichment analysis was done using DAVID (http://david.abcc.ncifcrf.gov/) (15) for the most significant gene ontology terms, pathways, and functional categories associated with the 131 classifier genes. Genes presented in Affymetrix U133A GeneChip were used as the background for the enrichment analysis. EASE score, a modified Fisher Exact P-Value, was used to rank the statistical significance of enrichment in an annotation category. Fisher Exact ***p***-Value = 0 represents perfect enrichment. Usually ***p***-Value is equal or smaller than 0.05 to be considered as strongly enrichment. The ten most significant enriched categories are shown. Each gene may have multiple entries.

Category	AFFY_ID	Gene Name
Anatomical structure development		
1	206874_s_at	collagen, type xvii, alpha 1
2	205303_at	potassium inwardly-rectifying channel, subfamily j, member 8
3	209915_s_at	neurexin 1
4	205973_at	fasciculation and elongation protein zeta 1 (zygin i)
5	210198_s_at	proteolipid protein 1 (pelizaeus-merzbacher disease, spastic paraplegia 2, uncomplicated)
6	205611_at	tumor necrosis factor (ligand) superfamily, member 12
7	206289_at	homeobox a4
8	218818_at	four and a half lim domains 3
9	210280_at	myelin protein zero (charcot-marie-tooth neuropathy 1b)
10	214023_x_at	tubulin, beta 2b
11	210632_s_at	sarcoglycan, alpha (50kda dystrophin-associated glycoprotein)
12	216894_x_at	cyclin-dependent kinase inhibitor 1c (p57, kip2)
13	212457_at	transcription factor binding to ighm enhancer 3
14	213808_at	adam metallopeptidase domain 23
15	201431_s_at	dihydropyrimidinase-like 3
16	214122_at	pdz and lim domain 7 (enigma)
17	215306_at	luteinizing hormone/choriogonadotropin receptor
18	202565_s_at	supervillin
19	212120_at	ras homolog gene family, member q
20	211964_at	collagen, type iv, alpha 2
21	205132_at	actin, alpha, cardiac muscle
22	210869_s_at 209086_x_at 211340_s_at 209087_x_at	melanoma cell adhesion molecule
23	209169_at	glycoprotein m6b
24	204736_s_at	chondroitin sulfate proteoglycan 4 (melanoma-associated)
25	204777_s_at	mal, t-cell differentiation protein
26	209686_at	s100 calcium binding protein, beta (neural)
27	214212_x_at	pleckstrin homology domain containing, family c (with ferm domain) member 1

28	216500_at	lectin, galactoside-binding, soluble, 1 (galectin 1)
29	210319_x_at	msh homeobox homolog 2 (drosophila)
30	212097_at	caveolin 1, caveolae protein, 22kda
31	206382_s_at	brain-derived neurotrophic factor
32	204159_at	cyclin-dependent kinase inhibitor 2c (p18, inhibits cdk4)
33	204939_s_at	phospholamban
	204940_at	
34	209843_s_at	sry (sex determining region y)-box 10
35	202806_at	drebrin 1
36	204584_at	l1 cell adhesion molecule

System development

1	206874_s_at	collagen, type xvii, alpha 1
2	205303_at	potassium inwardly-rectifying channel, subfamily j, member 8
3	209915_s_at	neurexin 1
4	205973_at	fasciculation and elongation protein zeta 1 (zygin i)
5	210198_s_at	proteolipid protein 1 (pelizaeus-merzbacher disease, spastic paraplegia 2, uncomplicated)
6	205611_at	tumor necrosis factor (ligand) superfamily, member 12
7	218818_at	four and a half lim domains 3
8	210280_at	myelin protein zero (charcot-marie-tooth neuropathy 1b)
9	214023_x_at	tubulin, beta 2b
10	210632_s_at	sarcoglycan, alpha (50kda dystrophin-associated glycoprotein)
11	216894_x_at	cyclin-dependent kinase inhibitor 1c (p57, kip2)
12	212457_at	transcription factor binding to ighm enhancer 3
13	213808_at	adam metallopeptidase domain 23
14	201431_s_at	dihydropyrimidinase-like 3
15	214122_at	pdz and lim domain 7 (enigma)
16	215306_at	luteinizing hormone/choriogonadotropin receptor
17	202565_s_at	supervillin
18	211964_at	collagen, type iv, alpha 2
19	205132_at	actin, alpha, cardiac muscle
20	209169_at	glycoprotein m6b
21	204736_s_at	chondroitin sulfate proteoglycan 4 (melanoma-associated)
22	204777_s_at	mal, t-cell differentiation protein
23	209686_at	s100 calcium binding protein, beta (neural)
24	216500_at	lectin, galactoside-binding, soluble, 1 (galectin 1)
25	210319_x_at	msh homeobox homolog 2 (drosophila)
26	206382_s_at	brain-derived neurotrophic factor
27	204159_at	cyclin-dependent kinase inhibitor 2c (p18, inhibits cdk4)
28	204939_s_at	phospholamban
	204940_at	
29	202806_at	drebrin 1
30	204584_at	l1 cell adhesion molecule

Developmental process

1	206874_s_at	collagen, type xvii, alpha 1
2	209015_s_at	dnaj (hsp40) homolog, subfamily b, member 6
3	205303_at	potassium inwardly-rectifying channel, subfamily j, member 8
4	209915_s_at	neurexin 1
5	205973_at	fasciculation and elongation protein zeta 1 (zygin i)
6	205611_at	tumor necrosis factor (ligand) superfamily, member 12
7	210198_s_at	proteolipid protein 1 (pelizaeus-merzbacher disease, spastic paraplegia

		2, uncomplicated)
8	206289_at	homeobox a4
9	218818_at	four and a half lim domains 3
10	212274_at	lipin 1
11	210280_at	myelin protein zero (charcot-marie-tooth neuropathy 1b)
12	202931_x_at	bridging integrator 1
	210201_x_at	
	214439_x_at	
13	214023_x_at	tubulin, beta 2b
14	201841_s_at	heat shock 27kda protein 1
15	210632_s_at	sarcoglycan, alpha (50kda dystrophin-associated glycoprotein)
16	214306_at	optic atrophy 1 (autosomal dominant)
17	216894_x_at	cyclin-dependent kinase inhibitor 1c (p57, kip2)
18	212457_at	transcription factor binding to ighm enhancer 3
19	213808_at	adam metallopeptidase domain 23
20	214122_at	pdz and lim domain 7 (enigma)
21	201431_s_at	dihydropyrimidinase-like 3
22	215306_at	luteinizing hormone/choriogonadotropin receptor
23	202565_s_at	supervillin
24	212120_at	ras homolog gene family, member q
25	211964_at	collagen, type iv, alpha 2
26	205132_at	actin, alpha, cardiac muscle
27	210869_s_at	melanoma cell adhesion molecule
	209086_x_at	
	211340_s_at	
	209087_x_at	
28	204628_s_at	integrin, beta 3 (platelet glycoprotein iiia, antigen cd61)
	204627_s_at	
29	204736_s_at	chondroitin sulfate proteoglycan 4 (melanoma-associated)
30	209169_at	glycoprotein m6b
31	204777_s_at	mal, t-cell differentiation protein
32	209686_at	s100 calcium binding protein, beta (neural)
33	214212_x_at	pleckstrin homology domain containing, family c (with ferm domain) member 1
34	216500_at	lectin, galactoside-binding, soluble, 1 (galectin 1)
35	210319_x_at	msh homeobox homolog 2 (drosophila)
36	212097_at	caveolin 1, caveolae protein, 22kda
37	206382_s_at	brain-derived neurotrophic factor
38	209651_at	transforming growth factor beta 1 induced transcript 1
39	204159_at	cyclin-dependent kinase inhibitor 2c (p18, inhibits cdk4)
40	204939_s_at	phospholamban
	204940_at	
41	209843_s_at	sry (sex determining region y)-box 10
42	202806_at	drebrin 1
43	206874_s_at	ste20-like kinase (yeast)
44	204584_at	l1 cell adhesion molecule

Disease mutation

1	221667_s_at	heat shock 22kda protein 8
2	206874_s_at	collagen, type xvii, alpha 1
3	210198_s_at	proteolipid protein 1 (pelizaeus-merzbacher disease, spastic paraplegia 2, uncomplicated)
4	206024_at	4-hydroxyphenylpyruvate dioxygenase
5	207554_x_at	thromboxane a2 receptor

	336_at	
6	202931_x_at	bridging integrator 1
	210201_x_at	
	214439_x_at	
7	210280_at	myelin protein zero (charcot-marie-tooth neuropathy 1b)
8	201841_s_at	heat shock 27kda protein 1
9	210632_s_at	sarcoglycan, alpha (50kda dystrophin-associated glycoprotein)
10	214306_at	optic atrophy 1 (autosomal dominant)
11	216894_x_at	cyclin-dependent kinase inhibitor 1c (p57, kip2)
12	205433_at	butyrylcholinesterase
13	215306_at	luteinizing hormone/choriogonadotropin receptor
14	218660_at	dysferlin, limb girdle muscular dystrophy 2b (autosomal recessive)
15	205132_at	actin, alpha, cardiac muscle
16	201843_s_at	egf-containing fibulin-like extracellular matrix protein 1
17	204628_s_at	integrin, beta 3 (platelet glycoprotein iiia, antigen cd61)
	204627_s_at	
18	205231_s_at	epilepsy, progressive myoclonus type 2a, lafora disease (laforin)
19	204365_s_at	receptor accessory protein 1
20	210319_x_at	msh homeobox homolog 2 (drosophila)
21	212097_at	caveolin 1, caveolae protein, 22kda
22	206382_s_at	brain-derived neurotrophic factor
23	204159_at	cyclin-dependent kinase inhibitor 2c (p18, inhibits cdk4)
24	204939_s_at	phospholamban
	204940_at	
25	209843_s_at	sry (sex determining region y)-box 10
26	204584_at	l1 cell adhesion molecule
Multicellular organismal development		
1	206874_s_at	collagen, type xvii, alpha 1
2	205303_at	potassium inwardly-rectifying channel, subfamily j, member 8
3	209915_s_at	neurexin 1
4	205973_at	fasciculation and elongation protein zeta 1 (zygin i)
5	210198_s_at	proteolipid protein 1 (pelizaeus-merzbacher disease, spastic paraplegia 2, uncomplicated)
6	205611_at	tumor necrosis factor (ligand) superfamily, member 12
7	206289_at	homeobox a4
8	218818_at	four and a half lim domains 3
9	202931_x_at	bridging integrator 1
	210201_x_at	
	214439_x_at	
10	210280_at	myelin protein zero (charcot-marie-tooth neuropathy 1b)
11	214023_x_at	tubulin, beta 2b
12	210632_s_at	sarcoglycan, alpha (50kda dystrophin-associated glycoprotein)
13	216894_x_at	cyclin-dependent kinase inhibitor 1c (p57, kip2)
14	212457_at	transcription factor binding to ighm enhancer 3
15	213808_at	adam metallopeptidase domain 23
16	201431_s_at	dihydropyrimidinase-like 3
17	214122_at	pdz and lim domain 7 (enigma)
18	215306_at	luteinizing hormone/choriogonadotropin receptor
19	202565_s_at	supervillin
20	211964_at	collagen, type iv, alpha 2
21	205132_at	actin, alpha, cardiac muscle
22	204628_s_at	integrin, beta 3 (platelet glycoprotein iiia, antigen cd61)
	204627_s_at	

23	209169_at	glycoprotein m6b
24	204736_s_at	chondroitin sulfate proteoglycan 4 (melanoma-associated)
25	204777_s_at	mal, t-cell differentiation protein
26	209686_at	s100 calcium binding protein, beta (neural)
27	216500_at	lectin, galactoside-binding, soluble, 1 (galectin 1)
28	210319_x_at	msh homeobox homolog 2 (drosophila)
29	212097_at	caveolin 1, caveolae protein, 22kda
30	206382_s_at	brain-derived neurotrophic factor
31	204159_at	cyclin-dependent kinase inhibitor 2c (p18, inhibits cdk4)
32	204939_s_at	phospholamban
	204940_at	
33	202806_at	drebrin 1
34	204584_at	l1 cell adhesion molecule

Cytoskeleton

1	203389_at	kinesin family member 3c
2	203151_at	microtubule-associated protein 1a
3	202565_s_at	supervillin
4	212565_at	serine/threonine kinase 38 like
5	205132_at	actin, alpha, cardiac muscle
6	205973_at	fasciculation and elongation protein zeta 1 (zygin i)
7	221246_x_at	tensin 1
8	218818_at	four and a half lim domains 3
9	209191_at	tubulin, beta 6
10	207876_s_at	filamin c, gamma (actin binding protein 280)
11	202931_x_at	bridging integrator 1
	210201_x_at	
	214439_x_at	
12	214212_x_at	pleckstrin homology domain containing, family c (with ferm domain) member 1
13	214023_x_at	tubulin, beta 2b
14	213847_at	peripherin
15	201841_s_at	heat shock 27kda protein 1
16	210632_s_at	sarcoglycan, alpha (50kda dystrophin-associated glycoprotein)
17	209651_at	transforming growth factor beta 1 induced transcript 1
18	202806_at	drebrin 1
19	214122_at	pdz and lim domain 7 (enigma)

Cytoskeleton organization and biogenesis

1	203389_at	kinesin family member 3c
2	209015_s_at	dnaj (hsp40) homolog, subfamily b, member 6
3	212793_at	dishevelled associated activator of morphogenesis 2
4	202565_s_at	supervillin
5	205132_at	actin, alpha, cardiac muscle
6	218818_at	four and a half lim domains 3
7	209191_at	tubulin, beta 6
8	214023_x_at	tubulin, beta 2b
9	214212_x_at	pleckstrin homology domain containing, family c (with ferm domain) member 1
10	213847_at	peripherin
11	214306_at	optic atrophy 1 (autosomal dominant)
12	202806_at	drebrin 1
13	214122_at	pdz and lim domain 7 (enigma)

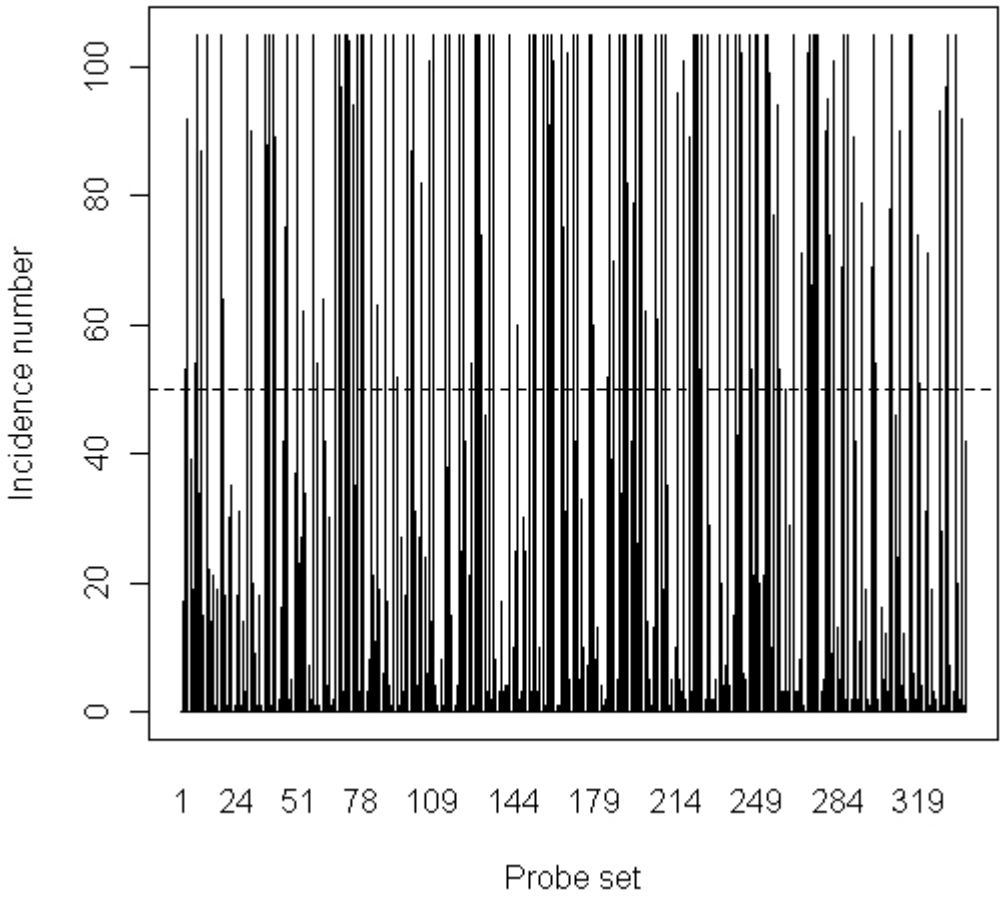
Cell-substrate junction assembly		
1	201389_at	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
2	221246_x_at	tensin 1
3	204628_s_at 204627_s_at	integrin, beta 3 (platelet glycoprotein iiia, antigen cd61)
Actin cytoskeleton		
1	207876_s_at	filamin c, gamma (actin binding protein 280)
2	202931_x_at 210201_x_at 214439_x_at	bridging integrator 1
3	214212_x_at	pleckstrin homology domain containing, family c (with ferm domain) member 1
4	212565_at	serine/threonine kinase 38 like
5	202565_s_at	supervillin
6	205132_at	actin, alpha, cardiac muscle
7	202806_at	drebrin 1
8	218818_at	four and a half lim domains 3
9	214122_at	pdz and lim domain 7 (enigma)
Cytoplasm organization and biogenesis		
1	210280_at	myelin protein zero (charcot-marie-tooth neuropathy 1b)
2	201389_at	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
3	221246_x_at	tensin 1
4	204628_s_at 204627_s_at	integrin, beta 3 (platelet glycoprotein iiia, antigen cd61)
Cell-substrate junction assembly		
1	201389_at	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
2	221246_x_at	tensin 1
3	204628_s_at 204627_s_at	integrin, beta 3 (platelet glycoprotein iiia, antigen cd61)
Actin cytoskeleton		
1	207876_s_at	filamin c, gamma (actin binding protein 280)
2	202931_x_at 210201_x_at 214439_x_at	bridging integrator 1
3	214212_x_at	pleckstrin homology domain containing, family c (with ferm domain) member 1
4	212565_at	serine/threonine kinase 38 like
5	202565_s_at	supervillin
6	205132_at	actin, alpha, cardiac muscle
7	202806_at	drebrin 1
8	218818_at	four and a half lim domains 3
9	214122_at	pdz and lim domain 7 (enigma)
Cytoplasm organization and biogenesis		
1	210280_at	myelin protein zero (charcot-marie-tooth neuropathy 1b)

2	201389_at	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
3	221246_x_at	tensin 1
4	204628_s_at 204627_s_at	integrin, beta 3 (platelet glycoprotein iiia, antigen cd61)

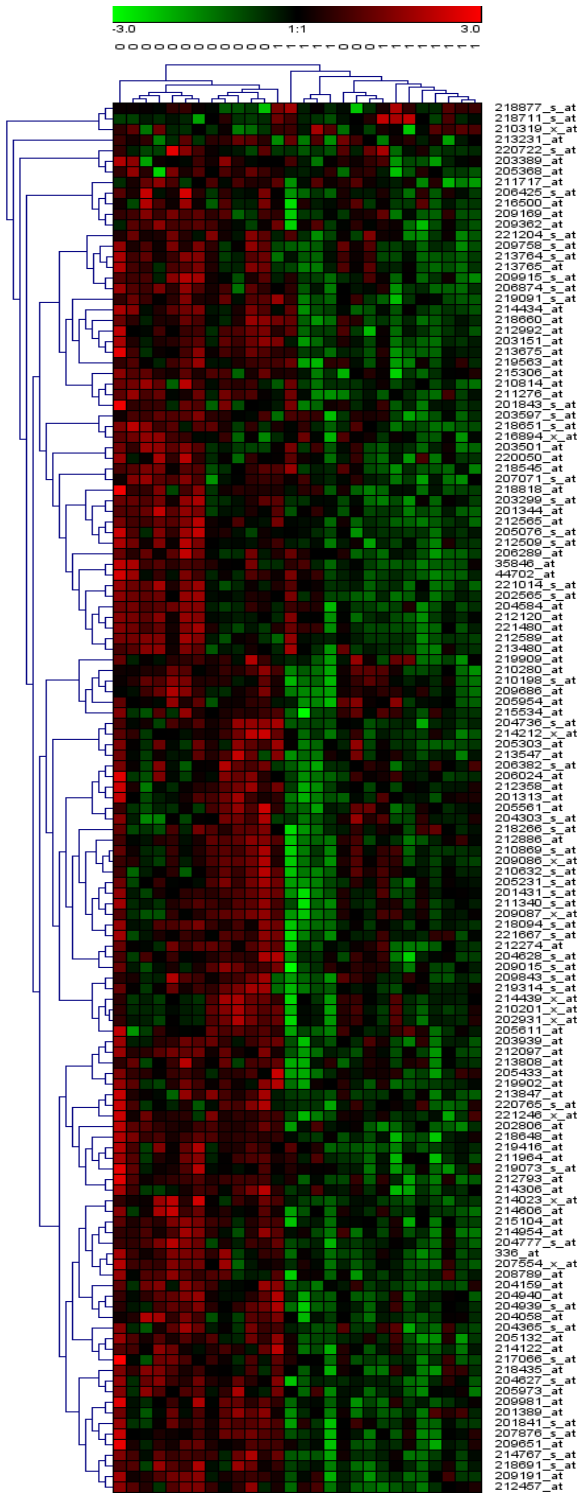
9. Table S4. PCR validation of preferential expression in stroma by representative genes of the Diagnostic Classifier.

Attribute \ Gene	**EFEMP1**	**CAV1**	**CCDC69**	**BHMT2**
Probability	3.92313E-05	4.32E-15	0.003158	9.16E-14
Mean (Stroma)	1.097978466	2.308086	0.14483	1.514679
Mean (Tumor)	0.591863008	0.894584	0.086355	0.61488
Ratio (Stroma/Tumor)	1.8	2.6	1.6	2.5

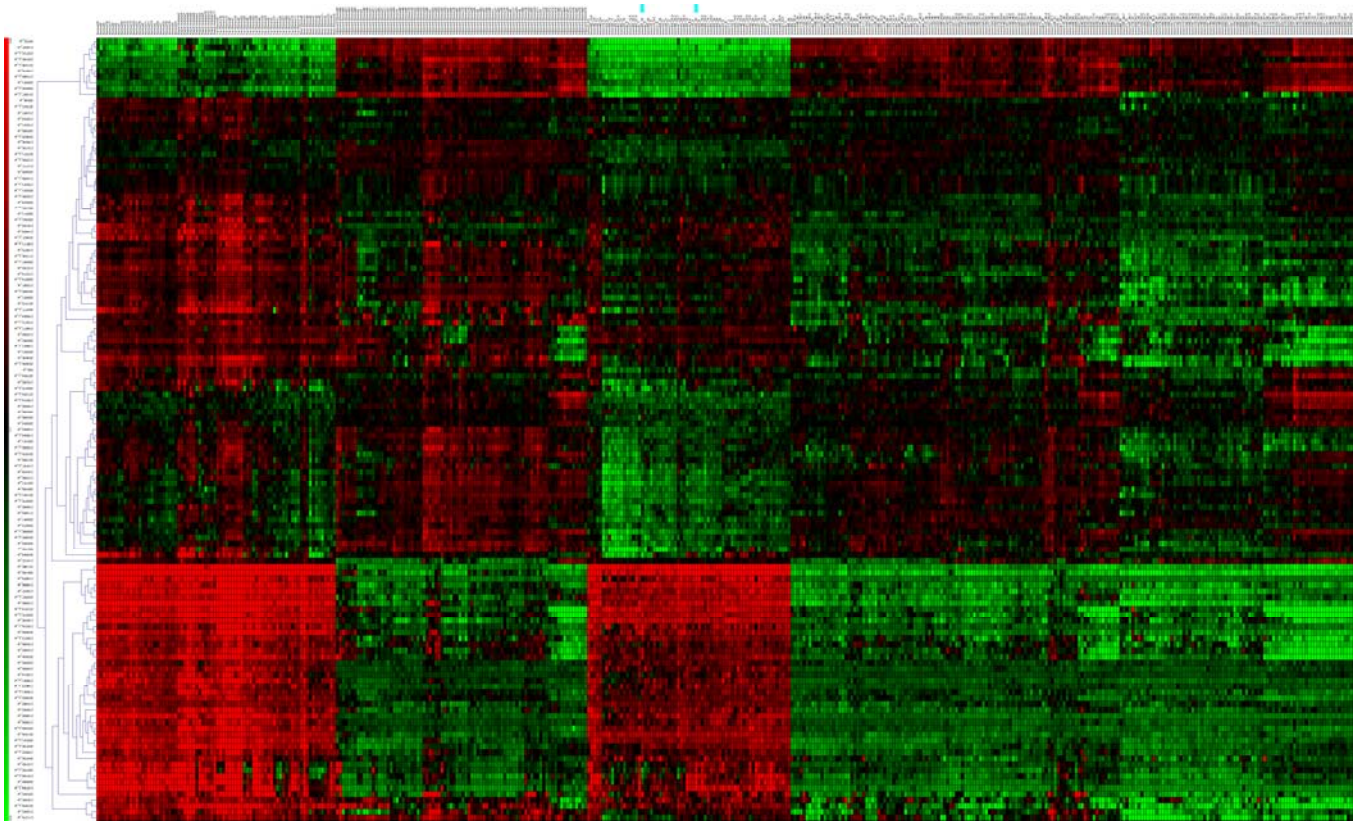
10. Figure S1. The incidence numbers of 339 probe sets obtained by 105-fold permutation procedure for gene selection. The dashed horizontal line marks the incidence number = 50. All probe sets with an incidence of >50 were selected for training using PAM using all 15 normal biopsy and the 13 original minimum tumor-bearing stroma cases.



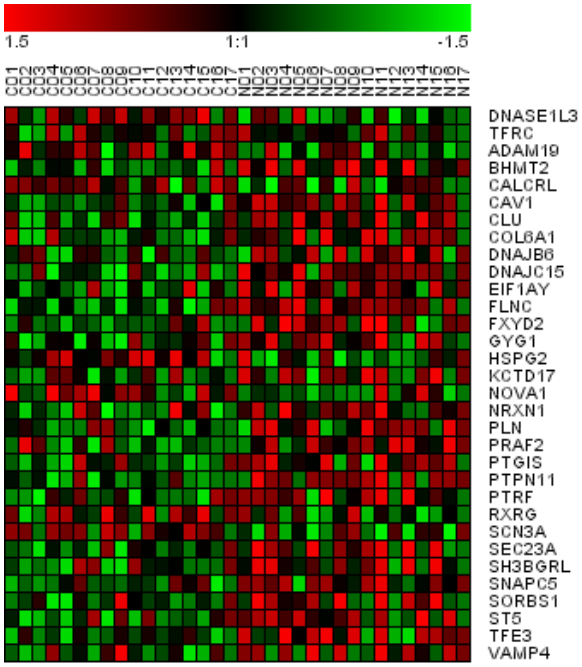
11. Figure S2. Heatmap using the 131 classifier to categorize all training cases.



12. Figure S3. Heatmap of all 364 test samples used in this study as categorized by the 131 probe set Diagnostic Classifier. The numbered bars above the various groups of cases indicate cases corresponding to the numbered case sets of **Table 2**.



13. Figure S4. Comparison with Dakhova et al. (1). Cluster diagram of the cases of Dakhova et al. (1) using 38 overlapping genes.



Comparison with Dakhova *et al.* (1). In Dakhova *et al.* (1), 1,141 unique genes were identified as differential expressed between reactive stroma and normal stroma using Agilent 44K gene expression array on seventeen paired laser-captured microdissected reactive stroma and matched normal stroma. While the original study did not provide the full differential expressed gene list, we downloaded the raw microarray data set from public Gene Expression Omnibus database (accession SE11682) for the purpose of assessing the agreements of our stroma classifier with Dakhova *et al. t* tests were used to for differential analysis and 2967 genes (6.6% of 45015 genes on the array) were identified as differential expressed with *p* cut off of 0.05. We used this loose criterion to generate a relatively larger gene list to be compared with our classifier basis set of 339 probe sets (the basis set selected by the 105-fold permuted selection procedure (**Figure S1**)) were mapped to 557 genes on the Agilent array, out of which 38 genes were in the 2967 gene list. 31 out of 38 showed concordance between two studies (**Table S2**). This overlap of 31 concordant genes between the two lists of 339 and 557 genes exceeds that expected by chance alone (***p*** = 0.0001, see **Table S2**). As expected,

these genes alone successfully categorize the cases of Dakhova *et al.* (1) into reactive and normal stroma cases (**Figure S4**).

The heatmap was made for the 38 genes using the Agilent microarray data set (SE11682). The top two genes were identified as up-regulated in our study and the bottom 30 genes were down-regulated (the comparison with Dakhova *et al.* (1) is shown in **Table S2)**. There were several probe sets corresponding to the same gene name, to avoid redundant mapping, the expression values were averaged. Sample names start with letter C represent tumor adjacent stroma and sample names start with letter N represent remote stroma.

The significance of the 38-31-gene concordance was assessed by comparison to a random model with using simulation, where "38" denoted the number of common genes in terms of gene identity between two studies while "31" indicated how many genes out of these "38" identity-concordant genes also concur on alteration tendency. We randomly, independently and respectively selected 557 probe sets and 2,502 probe sets derived above from the total of 37,765 probe sets of Agilent basis. If the number of common probe sets between these two sets of randomly selected genes is equal to or greater than 38 and no less than 31 of these identity-concordant probe sets have similar alteration direction, we increased the simulation count, C, by 1. We repeated this process by 10000 times. The *p* values associated with this test is defined as C/10000. In this simulation study, the *p* value was 0.003, indicating that the observed 38 overlapping probe sets can not be explained by chance and therefore our study and that of Dakhova *et al.* (1) are independent studies that are mutually supportive (an an algorithm in R is available at http://www.pathology.uci.edu/faculty/mercola/UCISPECSHome.html).

14. References for the Supplement

1. Dakhova O, Ozen M, Creighton CJ, et al. Global gene expression analysis of reactive stroma in prostate cancer. Clin Cancer Res 2009;15:3979-89.
2. Richardson AM, Woodson K, Wang Y, et al. Global expression analysis of prostate cancer-associated stroma and epithelia. Diagn Mol Pathol 2007;16:189-97.
3. Simoneau AR, Gerner EW, Nagle R, et al. The effect of difluoromethylornithine on decreasing prostate size and polyamines in men: results of a year-long phase IIb randomized placebo-controlled chemoprevention trial. Cancer Epidemiol Biomarkers Prev 2008;17:292-9.
4. Stuart RO, Wachsman William, Berry Charles C., Arden Karen, Goodison Steven, Klacansky Igor, McClelland Michael, Wang-Rodriquez Jessica, Wasserman Linda, Sawyers, Ann, Yipeng, Wang, Kalcheva, Iveata, Tarin David, Mercola Dan. In silico dissection of cell-type associated patterns of gene expression in prostate cancer. Proceeding of the National Academy of Sciences USA 2004;101:615-20.
5. Véronique Baron1 GDG, Anja Krones-Herzig2, Thierry Virolle3, Antonella Calogero2, Rafael Urcis1 and Dan Mercola1, 4. In 112th Internatl. Conf. On Gene Therapy, Hotel Del Coronado, December 12-14, 2002. Cancer Gene Therapy, 2002;9(12), Suppl. 1: S57. INHIBITION OF EGR-1 EXPRESSION RESTRAINS TRANSFORMATION OF PROSTATE CANCER CELLS AND DELAYS CANCER PROGRESSION. 2002.
6. Stephenson AJ, Smith A, Kattan MW, et al. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. Cancer 2005;104:290-8.
7. Wang Y, Xiao-Qin Xia, Zhenyu Jia, Anne Sawyers, Huazhen Yao, Jessica Wang-Rodriquez, Michael McClelland, Dan Mercola. In silico estimates of tissue components in surgical samples based on expression profiling data using. . Cancer Research (in press) [algorythm available at http://webarraydborg/webarray/indexhtml] 2010; .
8. Liu P, Ramachandran S, Ali Seyed M, et al. Sex-determining region Y box 4 is a transforming oncogene in human prostate cancer cells. Cancer Res 2006;66:4011-9.
9. Dalgaard P. Statistics and Computing: Introductory Statistics with R. pp. 260, Springer-Verlag Inc., NY. 2002.
10. Ayala G, Tuxhorn JA, Wheeler TM, et al. Reactive stroma as a predictor of biochemical-free recurrence in prostate cancer. Clin Cancer Res 2003a;9:4792-801.
11. Yanagisawa N, Li R, Rowley D, et al. Stromogenic prostatic carcinoma pattern (carcinomas with reactive stromal grade 3) in needle biopsies predicts biochemical recurrence-free survival in patients after radical prostatectomy. Hum Pathol 2007;38:1611-20.
12. Loehrlein C, Pollart, D., Shaler, T., Stephens, K., Tan, Y., Wong, L., and Monforte, J. , inventor Methods for Analysis of Gene Expression. U.S. Patent No. 6,618,679. 2003.
. 2003.
13. Chen QR, Vansant G, Oades K, et al. Diagnosis of the small round blue cell tumors using multiplex polymerase chain reaction. J Mol Diagn 2007;9:80-8.
14. Monks A HC, Pezzoli P, Kondapaka S, Vansant G, Petersen KD, Sehested M, Monforte J, Shoemaker RH. . Gene Expression-Signature of Nelinostat in Cell Lines is Specific for Histone Deacetylase Inhibitor Treatment, With a Corresponding Signature in Xenografts. . Anticancer Drugs 2009;20:682-92.
15. Dennis G, Jr., Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003;4:P3.