American Journal of Pathology, Vol. 163, No. 2, August 2003

Gene Discovery in Bladder Cancer Progression using cDNA Microarrays

Marta Sanchez-Carbayo,* Nicholas D. Socci,† Juan Jose Lozano,‡ Wentian Li,§ Elizabeth Charytonowicz,* Thomas J. Belbin,† Michael B. Prystowsky,† Angel R. Ortiz,‡ Geoffrey Childs,† and Carlos Cordon-Cardo*

From the Division of Molecular Pathology, Memorial Sloan-Kettering Cancer Center, New York; the Departments of Molecular Genetics and Pathology and the Seaver Center for Bioinformatics,† Albert Einstein College of Medicine, Bronx; the Department of Physiology and Biophysics,‡ Mt. Sinai School of Medicine, New York; and the Center for Genomics and Human Genetics,§ North Shore-Long Island Jewish Research Institute, Manhasset, New York*

To identify gene expression changes along progression of bladder cancer, we compared the expression profiles of early-stage and advanced bladder tumors using cDNA microarrays containing 17,842 known genes and expressed sequence tags. The application of bootstrapping techniques to hierarchical clustering segregated early-stage and invasive transitional carcinomas into two main clusters. Multidimensional analysis confirmed these clusters and more importantly, it separated carcinoma *in situ* from papillary superficial lesions and subgroups within early-stage and invasive tumors displaying different overall survival. Additionally, it recognized early-stage tumors showing gene profiles similar to invasive disease. Different techniques including standard *t*-test, single-gene logistic regression, and support vector machine algorithms were applied to identify relevant genes involved in bladder cancer progression. Cytokeratin 20, neuropilin-2, p21, and p33ING1 were selected among the top ranked molecular targets differentially expressed and validated by immunohistochemistry using tissue microarrays (n = 173). Their expression patterns were significantly associated with pathological stage, tumor grade, and altered retinoblastoma (RB) expression. Moreover, p33ING1 expression levels were significantly associated with overall survival. Analysis of the annotation of the most significant genes revealed the relevance of critical genes and pathways during bladder cancer progression, including the overexpression of oncogenic genes such as *DEK* in superficial tumors or immune response genes such as Cd86 antigen in invasive disease. Gene profiling successfully classified bladder tumors based on their progression and clinical outcome. The present study has identified molecular biomarkers of potential clinical significance and critical molecular targets associated with bladder cancer progression. *(Am J Pathol 2003, 163:505–516)*

Transitional cell carcinomas (TCCs) of the bladder define a group of histologically and genetically diverse cancers that account for ~4% of all adult malignancies with an annual incidence of ~53,200 cases in the United States.[1] Early-stage TCC has been classified into two groups with distinct clinical behavior and different molecular profiles. Superficial low-grade tumors (Ta) are always papillary and may recur but rarely progress, whereas high-grade tumors can be either papillary or flat lesions (Tis) and often progress to invasive disease.[2] Clinically, patients diagnosed with localized stage have a 5-year relative survival rate of 93%. However, patients presenting with regional and distant stage have 5-year relative survival rates of 49% and 6%, respectively.[1] Among the molecular events that characterize superficial papillary noninvasive bladder tumors are deletions affecting the long arm of chromosome 9, and activation of certain oncogenes, such as *H-RAS*, alterations identified only in a subset of invasive bladder neoplasms.[2] Deletions of 13q at the *RB* locus and 17p at the *TP53* locus, as well as 18q (*DCC* locus) and 5q (*APC* locus) losses have been reported in invasive bladder transitional carcinomas, but are absent in papillary noninvasive tumors.[2,3] A remaining challenge in bladder cancer is to define targets characteristic of aggressive early-stage tumors before they recur or progress into invasive disease. The present study was designed to identify critical molecular targets altered along the progression of bladder cancer using cDNA microarrays containing a high number of genes and expressed sequence tags (ESTs). Because molecular classification of tumor samples was not the main goal but rather gene discovery, bladder tumors representative of early and late stages were selected to cover the extremes in the natural history of bladder cancer progression. Several biocomputational methods were applied to identify targets displaying maximal expression differences among early- and late-stage tumors, and their clinical

Supported by the Seaver Foundation for Bioinformatics at the Albert Einstein College of Medicine of Yeshiva University (to N. D. S.).

Accepted for publication April 17, 2003.

Address reprint requests to Marta Sanchez-Carbayo, Ph.D., Division of Molecular Pathology, Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021. E-mail: sanchezm@mskcc.org.

impact was assessed using a larger cohort of patients with bladder cancer spotted onto tissue microarrays.

Materials and Methods

Cell Lines and Tumor Samples for cDNA Analysis

Four bladder cancer cell lines: T24, J82, RT4, and HT1197, were obtained from American Type Culture Collection (Rockville, MD) and maintained following standard procedures. All cells were grown and harvested at 75% to 90% confluence no longer than four to six passages. Total RNA of cell lines was extracted using RNeasy (Qiagen, Valencia, CA). Fifteen patients with bladder cancer were included for the expression-profiling study. Specimens were collected under an institutional review board-approved tissue procurement protocol. Bladder tumors embedded in OCT were macrodissected to ensure a minimum of 75% of tumor cells. Total RNA from bladder tumors was isolated in two steps using TRIzol (Life Technologies, Carlsbad, CA), followed by RNeasy purification.

cDNA Microarray Preparation, Hybridization, and Image Acquisition

A set of 17,842 sequence-verified human IMAGE cDNA clones, representing both known genes and ESTs, were polymerase chain reaction-amplified and spotted onto polylysine-coated microscope slides by the Albert Einstein College of Medicine microarray facility.[4] Five μg of total RNA from each bladder tissue and pool of cell lines was linearly amplified using a single round.[5] Amplified cRNA obtained from bladder tumors were labeled with Cy5 (red) (Amersham Biosciences AB, Uppsala, Sweden) and hybridized against amplified cRNA from the pool containing equal RNA quantities of the four cell lines labeled with Cy3 (green) (Amersham Biosciences AB). After hybridization, slides were washed, dried, and scanned by an Axon automated laser scanner. GenePix software was used for gridding and signal intensities calculation.[5,6]

Collection and Analysis of the Data of the cDNA Microarrays

Normalization

cDNA microarrays were normalized using an intensity-dependent algorithm.[5] Normalized fold changes in gene expression were then used to further analyze and cluster the bladder tumors.[6,7]

Clustering

Before clustering, the data were filtered to select genes having both significant average intensities and fold changes removing the background of unchanging genes. It was required that genes have a fold change of at least three (up or down) and an average intensity greater than 316 for at least two samples using the geometric average of the two channel intensities. This filter reduced the number of genes from 17,842 to 15,609. The set of 15 bladder tumor samples were then analyzed using the hierarchical clustering with the Ward linkage method, combined with nonparametric bootstrap resampling and consensus tree building to determine the support for sample groupings.[6,7]

Gene Ranking

Several scoring methods were applied to rank genes that could separate early-stage tumors from invasive organ-confined lesions and those developing metastatic bladder disease. Initially, the Mann-Whitney rank sum test was applied to identify genes differentially expressed between the two significant clusters.[8] Only genes showing a P value < 0.05 were considered for further analysis.

The method of single-variable logistic regression was applied labeling samples based on the histopathological records of the tumors analyzed in this study.[9,10] Single variable logistic regression was performed for all 15,650 genes/ESTs according to the maximum likelihood. The decay of the maximum likelihood as a function of the rank is approximated a power-law function or Zipf's law.[9,10]

Multidimensional analysis was then performed taking the 5616 genes providing data simultaneously in all of the tumors. The gene-expression matrix (N experiments $\times$ M genes) was first pretreated. All columns (M genes) were renormalized with the transformation $z = (x - \mu)/\delta$, that is, a new matrix (**X**-matrix) was created in which each column has a mean equal to 0 and variance equal to 1 (z-transformation). The **X**-matrix was then studied with Q-mode factor analysis (FA).[11,12] A supervised method was applied to detect the combination of genes in the **X**-matrix that can optimally be able to explain these groups. Specifically, the hyperplane that optimally separates the previously defined groups of samples was obtained by means of a support vector machine (SVM) algorithm.[13,14] The FA scores were then projected onto the characteristic vector of this hyperplane and then were z-transformed. The sorted z-scores were used to select the important descriptors able to separate the groups. A z-score cutoff of 2 was used to obtain a subset of cDNAs with the best discriminant properties.

We were interested in obtaining insights into the biochemical pathways involved in bladder tumor progression. We searched for the biological function index of the genes under study in gene ontology (GO).[14] These indexes were extracted in batch from SOURCE (*http://genome-www.standford.edu/source*). Biological processes according to GO were available only for 1044 genes of the 5616 under analysis. We identified by t-test the genes differentially expressed among each of the four groups generated by FA-SVM, as well as among superficial lesions (pooling groups 1 and 2) and invasive tumors (pooling groups 3 and 4) *versus* the rest of the experiments. The average t-test is calculated within the different groups as a measure of the enrichment of each biological process in the different groups. Only genes with t-test P

values < 0.01 were considered for further analysis. The statistical significance of the association of groups and biological processes (GO indexes) was evaluated by means of the hypergeometric distribution.[15,16] We computed the probability that at least *x* genes (with a *t*-test *P* value <0.01) were annotated within any given biological process in a random subset of *n* genes, where *N* denoted the total number of annotated genes (1,044) in the entire dataset, and *A* the number of these genes with a particular annotation. The following criteria were required for calculating the significance of the biological processes. Only biological processes (GO indexes) with more than five members in the set were selected for further analysis, only annotated genes with *t*-test *P* value <0.01 within each biological process were considered, and finally we required that at least two genes reached this significance within a biological process.[15,16] Hypergeometric *P* values <0.05 were considered significant, but we also focused on those marginally suboptimals, to compensate for the scarcity of GO annotations and the limitation that nonannotated genes could not be included in this analysis. Finally, significant cDNAs (according to a *t*-test), and belonging to significant or marginally significant GO pathways (*P* value <0.01) were grouped and represented using a color-graded spectrum.[15,16] Detailed information regarding the analyses are available on the web page for this article (*http://mskcc/GCL/BladderGenomics/cDNA*).

Clinical Validation of the Results

Tissue Samples in Tissue Microarrays

Three different bladder cancer microarrays were used in this study,[17] including a total of 173 bladder primary TCC tumors obtained under institutional review board-approved protocol. A total of 40 superficial and 64 invasive TCC tumors were analyzed in two microarrays. These tumors correspond to grade 1 ($n = 24$), grade 2 ($n = 8$), and grade 3 ($n = 82$) lesions. The third tissue microarray comprised a cohort of 69 bladder primary TCC cases with known p53/pRB status and annotated follow-up, including 2 superficial and 67 invasive lesions.

Immunohistochemistry

Protein patterns of expression were assessed at the microanatomical level using both cytospins from cell lines studied (data not shown) and tissue microarrays outlined above. Standard avidin-biotin immunoperoxidase procedures were applied for immunohistochemistry. We used the following panel of mouse monoclonal antibodies: neuropilin-2 (np-2) (clone 54; BD Transduction Labs, Lexington, KY); cytokeratin 20 (clone Ks20.8; DAKO, Glostrup, Denmark); cyclin E (clone cyE05; Neomarkers, Fremont, CA); p53 (clone 1801; Calbiochem, Cambridge, MA); total pRB (clone 3C8; QED Bioscience, San Diego, CA); under-phosphorylated pRB (clone G99-549; BD Transduction Labs); ninjurin (clone 50; BD Transduction Labs); p33ING1 (clone CAB1; BD Transduction Labs); and p21/WAF1 (clone Ab-1, Calbiochem). Control tissues for specificity assessment were used according to the manufacturers' recommendations. We used a 20% cutoff for p53 staining, 10% for p21, and 25% for cyclin E.[18] There is no consensus on the cutoffs of the immunohistochemical expression of the other markers, and thus they were analyzed as continuous variables, or taking several cutoffs when considered as categorical.

Statistical Analysis

All TCCs ($n = 173$) were used for the analysis of association between p53 and pRB with np-2, cytokeratin 20, cyclin E, and p21. These cases were also used for evaluating marker expression *versus* histopathological stage and tumor grade, using the nonparametric Mann-Whitney and Kruskall-Wallis tests.[8] The consensus value of the representative cores from each tumor sample arrayed was used for statistical analyses.

We analyzed the relationship of the cluster analysis of the bladder tumors to which expression profiling was performed with overall survival. Additionally, the association of the markers identified in the cDNA microarray analysis to outcome was also evaluated using a subset of 69 cTCC cases for which follow-up was available. Overall survival time was defined as the months elapsed between transurethral resection or cystectomy and death from disease (or the last follow-up date). Patients who were alive at the last follow-up or lost to follow-up were censored. For survival analysis, bootstrapping cluster and biomarkers were analyzed as categorical variables. The association of the marker expression levels with overall survival was analyzed using the Wald test, and the log-rank test was used to examine their relationship when different cutoffs were applied.[8] Survival curves were plotted using standard Kaplan-Meier methodology.[8] Associations between markers were analyzed using Kendall's tau *b*-test using the SPSS statistical package (version 8.0).

Results

Experimental Design

We performed the present analysis under the basis of two complementary sets of experiments. Initially, early-stage and advanced bladder tumors were analyzed using cDNA microarrays to identify differentially expressed genes among histopathologically distinct tumors representative of the natural history of bladder cancer progression (Table 1). The transcriptome of 15 bladder tumors was compared against a pool containing equal RNA quantities of four bladder cancer cell lines using cDNA microarrays containing 17,842 known genes and ESTs. This was followed by the study of the potential clinical significance of the selected targets identified by cDNA microarrays, which were validated at the microanatomical level using immunohistochemistry on tissue microarrays containing well-characterized bladder carcinomas. A cohort of early-stage and invasive bladder neoplasms was used to evaluate the association between biomarkers and histopathological stage and grade ($n = 173$). A

Table 1. Clinical, Histopathological, and Epidemiological Characteristics of the Patients to Whose Tumors cDNA Expression Profiling Analysis Was Performed,

Patient ID	BT	MD	Age	Sex	TNM	Carcinoma *in situ*	Prostate cancer	Smoking habit	Familiar cancer history	Follow-up (months)	Clinical outcome
174	1	1	67	M	TISG3N0	YES	YES	NO	NO	19	NED
160	1	1	83	M	TAG1N0	YES	YES	NO	NO	20	NED
157	1	1	61	M	TISG3N0	YES	NO	NO	NO	44	NED
134	1	1	80	M	T3BG3NO	YES	YES	YES	YES	13	NED
169	1	2	75	F	TAG3NO	NO	NO	NO	NK	47	NED
165	1	2	75	M	TAG1NO	NO	NO	NO	NO	18	NED
163	1	2	65	M	TAG1NO	NO	NO	YES	YES	22	NED
162	1	2	60	M	TISG3NO	YES	YES	YES	YES	20	NED
170	2	3	55	M	T4G3N0M1	YES	YES	YES	NK	4	DOD
168	2	3	61	M	T4BN1M2	YES	YES	YES	NK	1	DOD
141	2	4	49	F	TAG3M10	NO	NO	YES	YES	3	DOD
135	2	4	64	F	T4BG3N1	YES	NO	YES	YES	11	DOD
133	2	4	83	M	T3BG3NO	YES	YES	YES	YES	1	DOD
130	2	4	72	M	T3BG3N0	NO	YES	NO	NO	19	NED
124	2	4	59	F	T3AN1M0	NO	NO	YES	NO	17	REC

BT, Bootstrap clustering; MD, multidimensional grouping; sex, M (male), F: (female); TNM, tumor node metastases; clinical outcome, NED (no evidence of disease), DOD (death of disease), REC (recurrence).

subset of these bladder tumors ($n = 69$), with characterized p53 and pRB alterations and clinical follow-up, was used to delineate associations between potential novel biomarkers, cell-cycle regulators, and patient outcome.

Molecular Classification of Bladder Tumors Using cDNA Microarrays

The use of unsupervised hierarchical clustering combined with nonparametric bootstrap analysis classified primary bladder carcinomas based on their histopathological criteria. Overall, we observed that the early-stage tumors clustered together and were segregated from invasive TCC. Interestingly, within invasive bladder tumors we found that cases developing metastasis and displaying a shorter survival could be distinguished from those belonging to patients displaying a longer survival and organ-confined disease. The bootstrap resampling technique was able to establish a high confidence on these clusters (Figure 1A). Patients whose tumor samples were subjected to gene profiling had a median follow-up of 18 months (mean, 17.3 months; range, 1 to 47 months) (Table 1). The early-stage and the invasive clusters also grouped patients with different overall survival (Figure 1B). Because the identified clusters were associated with histopathology and clinical outcome, these results revealed the diagnostic and prognostic utility of unsupervised clustering even with a small number of patients. A multidimensional analysis of the 4729 genes providing expression data in all of the tumors, revealed four groups of gene profiles consistent with the previously identified clusters[11,12,19] (Figure 1C, Table 1). Moreover, we observed that gene expression profiling could separate carcinoma *in situ* from papillary superficial lesions. Two subgroups within early-stage and invasive tumors displaying different clinical outcome were also identified. Furthermore, the expression profiles of certain early-stage tumors displayed similar expression portraits than some organ-confined invasive lesions (Figure 1D). Thus, gene profile analyses provided predictive information regarding patients with early-stage bladder cancer and their likelihood to progress into invasive disease.

Identification of Genes Differentially Expressed between Early-Stage and Invasive Bladder Cancer

Several scoring methods were used for gene identification. Standard *t*-test and single-gene logistic regression were applied to rank the genes according to their ability to separate early-stage from invasive tumors (Table 2). First, the Mann-Whitney rank sum test was applied as a standard means for gene identification between these two groups.[8] The goal was to identify genes differentially expressed between the early-stage and invasive clusters. We observed that the first 120 genes correctly classified the samples contained in each cluster ($P = 0.033$). Two genes, p21 and cyclin E, were selected for further study because of their participation in the p53 and RB signaling pathways, both of which are frequently altered in bladder cancer progression.[3,12,13,20] We then performed a single-gene variable logistic regression analysis as a standard classification/discrimination model to rank genes by their classification performance.[9,10] In this case, the goal was to identify genes differentially expressed between early-stage disease and invasive tumors. Our results demonstrated that any of the 92 top-ranked genes could differentiate early-stage *versus* invasive lesions based on their gene expression levels (Figure 2). For those genes ranked from 93 to 500 using this analysis, a maximum of three misclassifications could be obtained. We chose to focus on the 92 genes that provided no misclassification among early-stage *versus* invasive tumors. Two genes from this initial group, cytokeratin 20 and neuropilin-2 (np-2), which are soluble proteins with potential role for tumor marker development, were selected and studied by immunohistochemistry on tissue microarrays. Finally, we continued to further elucidate the genes that best characterized each of

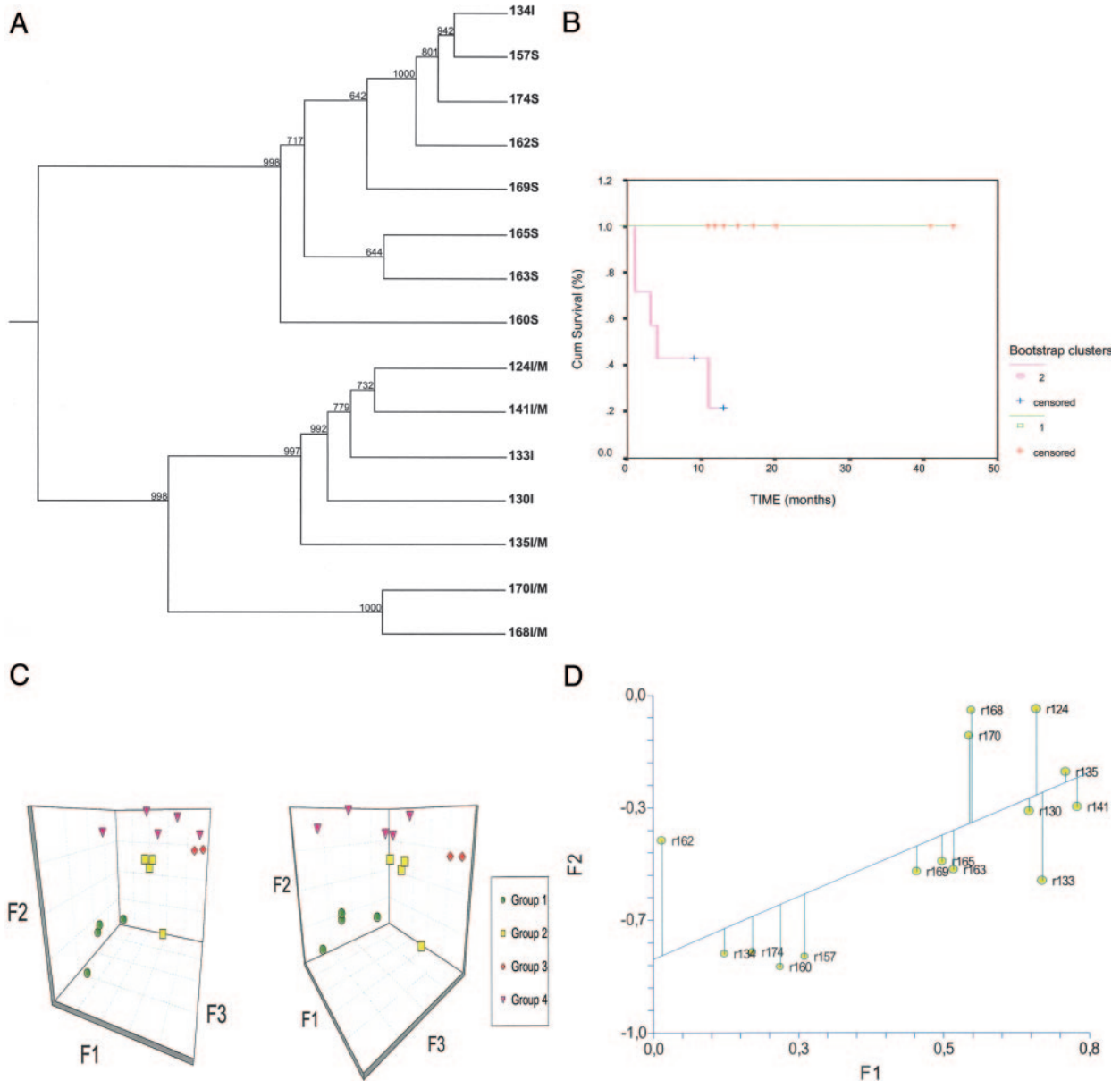

Figure 1. A: Hierarchical clustering using bootstrap resampling techniques classified these bladder tumors according to histopathological criteria. A tree is constructed by finding for each node the pairing that occurred most often in the 1000 separate trials displaying this count at each node of the tree. The number on each node represents how many times that samples to the right are grouped together out of a total of 1000 tries, a larger number indicates tight clustering. S, superficial bladder tumors; I, organ-confined invasive bladder tumors; I/M, invasive bladder tumors developing metastatic disease. **B:** Kaplan-Meier survival analysis of patients with superficial (cluster 1) and invasive (cluster 2) bladder tumors stratified by bootstrap clusters. These clusters containing the superficial and invasive tumors were found to be significantly associated with overall survival (log rank, $P = 0.0025$). **C:** Multidimensional analysis: four groups of expression profiles were identified by factor analysis and these were consistent with the superficial (groups 1 and 2) and invasive (groups 3 and 4) clusters. These two scatterplots illustrate different orientations of the three-dimensional plot. **D:** FA-SVM analysis: the **line** shows the association among the distribution of the expression profiles of these patients with the progression of the disease in these bladder cancer patients. Early-stage tumors are at the **bottom left** and highly invasive tumors at the **top right**. The expression profiles of certain superficial tumors (163, 165, and 169) were more similar to some organ-confined invasive lesions.

the four groups generated by the multidimensional approach.[11,12,19] Groups 1 and 2 included tumors with no evidence of disease after follow-up. Group 3 included the most advanced tumors, those developing metastatic disease and presenting shorter survival. Group 4 represented those primary tumors associated with metastatic and nonmetastatic lesions (Table 1). One gene from each of these groups 3 and 4, ninjurin and p33ING1, respectively, were selected for further evaluation of their potential prognostic value.

Association of the Selected Biomarkers with Clinicopathological and Molecular Variables

The potential role of the identified target genes in the staging of patients with bladder cancer was analyzed using tissue microarrays containing TCC of different stages and grades. Cytokeratin 20, np-2, p21, and p33ING1 were differentially expressed in early-stage and invasive tumors. In the subset of patients analyzed, we observed a significant correlation between the expres-

Table 2. Summary of the Top-Ranked Genes Obtained from the Scoring Methods Applied in This Study

Test	Accession	Gene name
Mann-Whitney test		
$P = 0.033$	R25377	DEK oncogene
	AA449831	Growth factor receptor-bound protein 2 (GRB2)
	AA862434	PROTEASOME CHAIN 7 PRECURSOR
	AA775415	SMT3 suppressor of mif two 3
	AA935560	Relaxin 2
	W49619	N-cadherin
	AA450265	PCNA
	T96829	Cyclin-E
	AA599093	Rab3-GAP regulatory domain
	AA465593	PSMA3
	R69307	P21
	R24543	NET1
	AA043806	Beta 3-endonexin
	AA485052	Proteasome (prosome, macropain) 26S subunit
Single-variable logistic regression		
92 Top-ranked	AA705060	IGF2R, insulin-like growth factor 2 receptor
	H58736	DMR protein
	AA482325	Cytokeratin K20
	H70815	Neuropilin-2
	AA157797	EGF-like module EMR2
	N66933	Tumor suppressor pp32r1
	AA024832	Glypican 6, leukocyte differentiation antigen (CD84)
	W56308	Glutathione peroxidase (GPX2)
	AA496780	RAB7, member RAS oncogene family
	W31919	TTN gene for titin
	AA460365	Neurexin III-alpha
	AA436871	Syntaxin 3A (STX3A)
	AA488526	Nucleolar phosphoprotein p130
	N94428	E1A binding protein p300 (EP300)
Multidimensional analysis		
Group 1	AA416585	ACE-related carboxypeptidase ACE2
	AA056381	TYROSYL-TRNA SYNTHETASE
	R24258	Protein kinase C
	N67578	Aquaporin-5
Group 2	H98694	PI-3-kinase-related kinase SMG-1
	AA488073	Mucin (PEM)
	R98047	7q31 (TES)
Group 3	AA406020	ISG15
	H28734	Glutamate receptor, ionotropic, AMPA 2 (GRIA2)
	AA026167	22q11
	AA455302	ING1 tumor suppressor
Group 4	H73591	Ctochrome b5 outer mitochondrial membrane precursor
	R00884	Dihydrofolate reductase (DHFR)
	T86959	Fibrillin
	AA625806	Ninjurin

sion of these proteins with tumor stage and grade (Table 3A). Levels of p33ING1 expression were easily detectable in normal urothelium and in the majority of early-stage TCC, but were much lower in invasive tumors (Figure 3, A and B). Because both *TP53* and *RB* signaling pathways are frequently altered during bladder cancer progression, we also evaluated the association of these biomarkers with p53 and pRB status. We found that the expression of cyclin E was significantly associated with p53 expression; cytokeratin 20, np-2, p21, and p33ING1 were all associated with altered pRB expression (Table 3B). A significant correlation between p33ING1 expression with cyclin E and p21 was also noted.

The prognostic utility of cytokeratin 20, np-2, p21, and cyclin E was evaluated using the 69 TCCs for which clinical follow-up was available. We observed that only the expression of p33ING1 was significantly associated with overall survival ($P = 0.02$). Patients displaying a higher expression of p33ING1 showed a shorter survival than those with low expression of this protein (Figure 3C). Overall, the results demonstrated that the expression of these selected molecular markers was associated with tumor stage, grade, p53/pRB expression, and overall survival.

Molecular Pathways Involved in Bladder Cancer

A multivariate analysis,[16] was applied to detect the combination of genes that can optimally explain the four main groups identified by multidimensional analysis as well as to discriminate early-stage (pooling together groups 1 and 2) *versus* invasive (pooling together groups 3 and 4) TCCs. We focused on genes that were differentially expressed in each of these groups *versus* the remaining samples. We then grouped these genes according to the

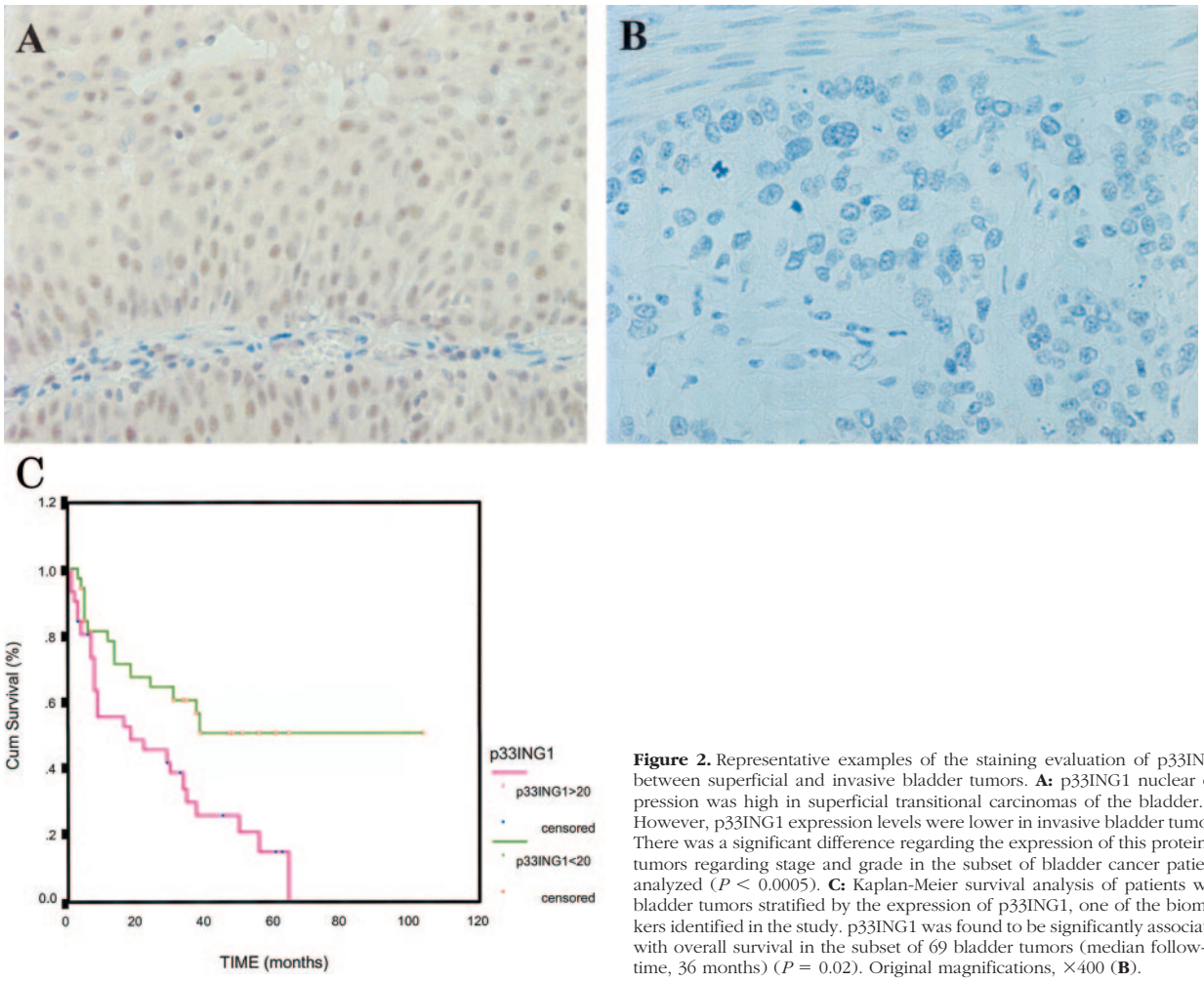

Figure 2. Representative examples of the staining evaluation of p33ING1 between superficial and invasive bladder tumors. **A:** p33ING1 nuclear expression was high in superficial transitional carcinomas of the bladder. **B:** However, p33ING1 expression levels were lower in invasive bladder tumors. There was a significant difference regarding the expression of this protein in tumors regarding stage and grade in the subset of bladder cancer patients analyzed ($P < 0.0005$). **C:** Kaplan-Meier survival analysis of patients with bladder tumors stratified by the expression of p33ING1, one of the biomarkers identified in the study. p33ING1 was found to be significantly associated with overall survival in the subset of 69 bladder tumors (median follow-up time, 36 months) ($P = 0.02$). Original magnifications, ×400 (**B**).

molecular pathways in which they are involved according to the GO database. Table 4 summarizes the overexpressed and underexpressed molecular functional pathways with at least two genes differentially expressed in each of these groups. The analysis of early-stage lesions *versus* invasive tumors together revealed the relevance of overexpression of oncogenic genes, such as *DEK* in superficial tumors, *versus* the overexpression of immune response genes, such as cd86 or lipid metabolism-related genes, in invasive tumors. The subset of overexpressed genes with the most optimal discriminatory properties is shown in Figure 4.[13,14] The complete ranked gene list will be available in the web site created for this manuscript (*http://mskcc/GCL/BladderGenomics/cDNA*).

Discussion

The molecular discrimination between early-stage and invasive bladder cancer has already been described in previous analysis of pools of bladder tumors of different stage and grades using oligonucleotide microarrays.[21,22] Bladder cancer cell lines derived from histopathologically distinct tumors have been also classified based on their expression profiles.[23] It is then not surprising that in our study hierarchical clustering analyses segregated early-stage and advanced invasive tumors, considering that patients were selected to represent early and late events of bladder cancer progression. Thus, individual tumors have been classified according to histopathological criteria using a different microarray platform. Bootstrap resampling was used to assess the robustness of the clustering results. The relevant finding in terms of the molecular classification of these tumors was that multidimensional analyses not only supported such clustering but also separated carcinoma *in situ versus* superficial papillary lesions, and segregated subgroups within noninvasive and invasive disease with different clinical outcome. Moreover, it identified those early-stage tumors with expression profiles more similar to invasive lesions. The use of a selected relative low number of bladder tumors representing the early and late stages of the disease and a high number of genes and ESTs allowed to identify critical molecular targets and molecular pathways involved in bladder cancer progression. Different independent biocomputational techniques were applied for gene identification, rendering complementary information. The molecules selected for validation in tissue microarrays were associated with tumor stage and

Table 3. Association between the Expression of the Identified Markers with: (A) Tumor Stage and Grade; and (B) with the Expression of p53 and pRB

A) Histopathology

Biomarker	Stage		Grade	
	Number of cases	*P* value	Number of cases	*P* value
p21	150	<0.0005	147	<0.0005
Cyclin E	155	NS	153	NS
Cytokeratin 20	155	0.005	153	<0.005
Neuropilin-2	157	<0.0005	154	<0.0005
Ninjurin	147	NS	144	NS
p33ING1	146	<0.0005	145	<0.0005

B) Associations with p53 and pRB expression

Biomarker	p53			Underphosphorylated RB		
	Number of cases	Kendall's tau-b	*P* value	Number of cases	Kendall's tau-b	*P* value
p21		NS		146	0.316	<0.0005*
Cyclin E	55	0.203	0.001		NS	
Cytokeratin 20		NS		151	0.233	0.0005
Neuropilin-2		NS		152	0.217	<0.0005
Ninjurin		NS			NS	
p33ING1		NS		143	0.370	<0.0005*

*Significant also for total pRB.

grade, as well as altered p53 and pRB expression. Furthermore, p33ING1 showed a significant association with patient survival. Overall, both clusters and individual targets showed clinical value for subtype classification and prognosis of patients with bladder cancer.

Apart from the classification of these tumor samples, the analytical approach undertaken focused on identifying genes and pathways related to bladder cancer progression. After the segregation of tumor subtypes by gene clustering, our data analysis focused on gene identification methods and validation using tissue microarrays on a larger separate cohort of patients with bladder cancer. Different algorithms were applied for gene identification providing distinct ranked gene lists.[9] The Mann-Whitney test is a standard means for gene identification between two groups. In this case, p21 and cyclin E were selected for further validation. p21 was found to be associated with tumor stage and grade, in accordance with previous series.[18,24] However, we did not find any association between cyclin E expression and these clinicopathological variables.[25,26] Single-variable logistic regression is a standard classification/discrimination model to rank gene by their classification performance.[9,10] This method takes into consideration the dispersion of the expression data of each gene within each group under comparison, in our case, early-stage and invasive tumors. The Mann-Whitney test compares the median expression of a certain gene between two groups, np2 and cytokeratin-20, known soluble proteins, were selected for further validation, and their expression was significantly associated with tumor stage and grade. np-2 is a transmembrane receptor for semaphorins (mediators of neuronal guidance), and for several angiogenic factors, including vascular endothelial growth factor (VEGF) 145 and VEGF 165.[27–29] It has been reported that osteosarcomas overexpressing np-2 had increased vascularity and poorer prognosis, suggesting that np-2 acts as a VEGF-amplifier in these tumors.[29] An association between np-2 expression and tumor progression has also been reported for certain neoplasms, including prostate and lung cancer.[30,31] The association of cytokeratin 20 with stage and grade had previously been reported for bladder cancer both in tissue and urine specimens.[32–34] Comparing tumor samples by cDNA microarrays representing both early and late stage in bladder cancer progression has identified single genes that separate early-stage and invasive tumors. Two different scoring methods, *t*-test and single logistic regression analysis, have provided more than 200 genes that segregate early-stage and invasive groups of bladder tumors under cDNA analysis. That single genes could discriminate closely related cancer subtypes as reported in breast tumors,[35] is unusual. Our study has compared tumor samples representative of the opposite early and late stages of the disease, a situation in which a single gene is more likely to detect these differences. We have evaluated only six of these genes based on the availability of antibodies for immunohistochemical studies. Further studies are warranted to evaluate the potential prognostic relevance of these and other genes revealed to be involved in bladder cancer progression in a larger series of followed-up bladder cancer patients.

Not only we were interested in identification of genes with histopathological diagnostic properties. We also focused on genes with potential prognostic value. In this regard, the characterization of genes included in the subgroups rendered by the FA-SVM analysis was of interest. Ninjurin and p33ING1 were selected among the target genes differentially expressed in the poor prognostic groups identified by multivariate analysis. Ninjurin, a nerve injury-induced protein involved in neuronal growth, is known to be altered in hepatocellular carcinoma,[35]

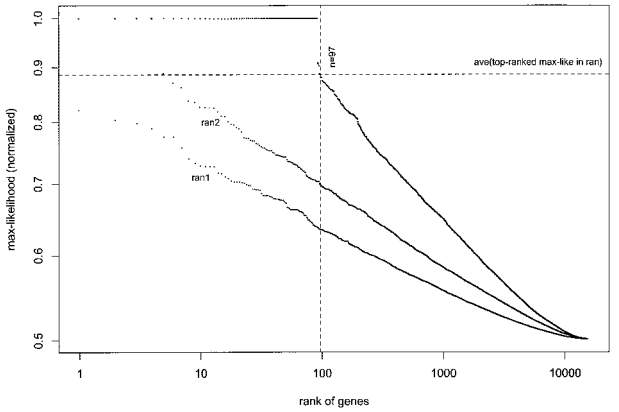

Figure 3. Single-gene logistic regression analysis. The maximum likelihood of single-gene logistic regression that classifies superficial (S) *versus* invasive and metastatic (I+M) samples is ranked and plotted against the rank (in logarithm scale). Each **point** represents a gene of a total of 15,650 genes. Two plots are also drawn for the similar analysis with sample label permuted randomly (ran1 and ran2). When such permutation analysis is repeated many times, the average value of the maximum likelihood of the top-rank gene can be calculated by a mathematical formula,[10] and is represented by the **horizontal line**. Ninety-seven genes were shown to provide better likelihood values than the best value from the randomized data set. RAN, random permutation number.

acute lymphoblastic leukemia,[36] and was reported to be down-regulated by p53.[37] Nevertheless, ninjurin was not found to be significantly associated with tumor stage and grade in the cohort of bladder cancer patients analyzed. Although p33ING1 was not clearly differentially expressed among the superficial and invasive tumors analyzed by the cDNA microarrays, the expression of p33ING1 was significantly associated with tumor stage and grade in the larger cohort of patients spotted onto the tissue microarrays. p33ING1 has been reported to cooperate with p53 in blocking cell proliferation and enhancing apoptosis, and it has been ascribed as a candidate tumor suppressor gene.[38] Remarkably, it correlated with overall patient survival in our subset of 69 followed-up patients that consisted of a majority of invasive TCCs. Shorter survival was observed in invasive patients with squamous differentiation, a feature known to be associated with aggressive clinical behavior.[2,3] Its expression has also been reported to be involved in the progression of lymphoid tumors.[39] We found that down-regulation of p33ING1 was associated with pRB, p21, and cyclin E expression, but not with p53.[40] Overall, the mechanistic networks by which p33ING1 is involved in bladder cancer progression need further validation with specific studies targeting these issues, out of the scope of the experimental design of this study.

Multidimensional analysis provides a descriptive visualization of associations among tumor samples and gene expression patterns. Factor analysis, in particular, allows a soft classification of samples and genes. That is, its main advantage is that a particular gene can be relevant in more than one group. This method is unsupervised. When additional information is available, as in this case, supervised analysis is preferred. We incorporate this using a SVM layer on top of the FA results. SVM supervised algorithms have mostly been applied for subclassification of tumors and predictive purposes.[22] We have also applied a mixed Q-mode FA-SVM model[11–13] together with an analysis of functional annotations on the selected genes,[14] to obtain insight on the molecular processes

Table 4. Relation of Biological Functional Annotations Most Overexpressed or Underexpressed with Their Hypergeometric *P* Values (Using a Cutoff of 0.10)

	Group	Hypergeometric *P* value (<0.10)	Number of genes (*t*-test p < 0.01)	Number of genes within each BP among the 1,044 annotated genes
Biological process (BP) overexpression				
Excretion	1	0.0088340	3	9
RNA_splicing	1	0.0602896	2	8
Epidermal_differentiation	3	0.0601808	2	7
Lipid_metabolism	3	0.0054039	4	13
Immune_response	4	0.0021934	5	18
Neurogenesis	4	0.0296986	4	22
Oncogenesis	5	0.0989333	4	55
Development	6	0.0510570	6	36
Immune_response	6	0.0429155	4	18
Lipid_metabolism	6	0.0710055	3	13
Protein_complex_assembly	6	0.0948138	2	7
Biological process (BP) underexpression				
Protein_modification	1	0.0000444	5	18
Cell_adhesion	4	0.0867590	2	29
Regulation_of_cell_cycle	4	0.0486711	2	21
Regulation_of_transcription_from_Pol_II	4	0.0328320	2	17
Development	5	0.0510570	6	36
Immune_response	5	0.0429155	4	18
Lipid_metabolism	5	0.0710055	3	13
Protein_complex_assembly	5	0.0948138	2	7
Oncogenesis	6	0.0989333	4	55

Shown is the GO biological process, the group where that biological process is found significantly expressed as compared to the other groups, and the associated *P* value. Groups 1 to 4 are defined by the FA-SVM, group 5 includes superficial lesions (pooling groups 1 and 2), and group 6 includes invasive tumors (pooling groups 3 and 4). The number of genes within each biological process with a *t*-test *P* value lower than 0.01, the number of genes among the 1044 annotated ones with annotated function within a group are also included.

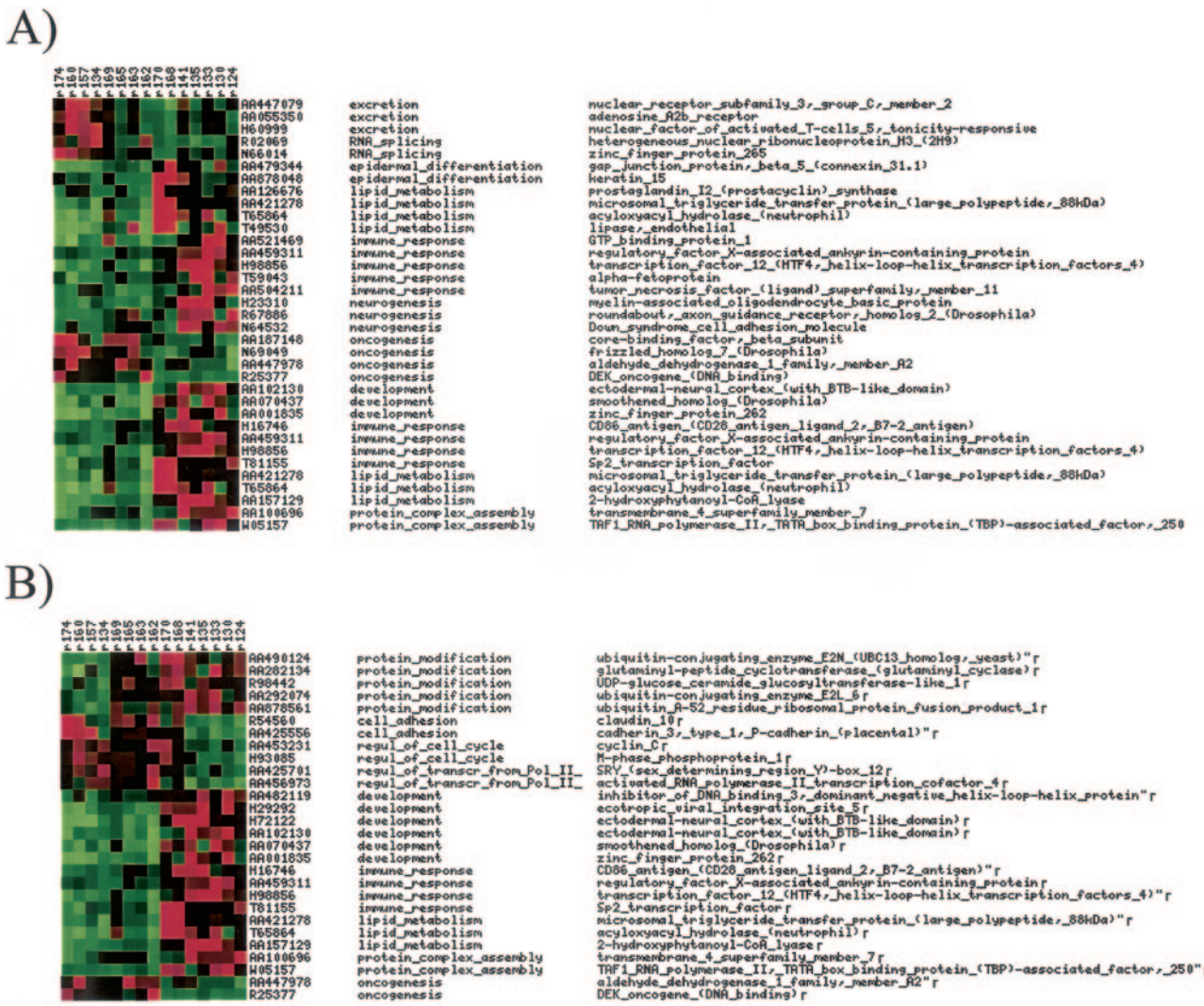

Figure 4. Set of selected cDNAs annotated with GO biological function indexes showing a P value <0.01 in the t-test of group discrimination. We separately include these significant genes belonging to the biological functional annotations that provided a significant ($P < 0.05$) or marginally significant hypergeometric P value ($P < 0.10$). **A:** Overexpressed genes belonging to each of the significant annotations. **B:** Underexpressed genes belonging to each of the significant annotations. It is worth noting that many of the overexpressed genes in superficial tumors were underexpressed in the invasive tumors and vice versa.

taking place along bladder cancer progression. This approach was possible because the tumor samples under study could be representative of both early and late events associated with progression of these uroepithelial tumors. Initial events significantly altered in early stages include the overexpression of cell membrane receptors involved in intracellular signaling processes, excretion, and RNA-splicing genes, as well as the underexpression of ubiquitinazing enzymes. Regarding invasive subgroups, there was a remarkable overexpression of lipid metabolism and epidermal differentiation targets such as keratin 15. We also observed overexpression of immune response and neurogenesis-related genes also involved in remodeling of the extracellular matrix.[22] A remarkable underexpression of cell adhesion, and regulation of transcriptional and cell cycle processes was observed in the most advanced invasive tumors, as previously reported.[21–23] The analysis of early-stage *versus* invasive tumors identified novel oncogenes altered in bladder cancer such as *DEK*, and revealed that oncogenetic events are among the most overexpressed networks in superficial bladder cancer, including the *WNT* pathway represented by frizzled proteins. Interestingly, genes such as *DEK* were found significantly expressed among bladder cancer subtypes in the other gene identification algorithms used in the present study. Immune response and development-related genes as well as those involved in protein complex assembly, were shown to be among the most overexpressed genes in invasive tumors.[21,22] We also observed that lipid metabolism-related genes, some of them involved in angiogenesis pathways, are overexpressed in invasive tumors, when compared to early-stage lesions. The observation that functional annotations that were significantly overexpressed in superficial tumors (group 5) were also found to be significantly underexpressed in invasive lesions (group 6), provided confidence to our analysis. Mitotic spindle checkpoint-related genes such as *Cdc16*, and apoptosis network related genes such as *BCL2*, were among the most significantly expressed genes in bladder cancer progression (Figure

4). However, functional annotation for these and other important pathways, possibly altered in bladder cancer progression, did not reach significance as a group pathway under the restrictions that we took in our approach. This may also be because of the low density of functional annotation provided by GO, or the variability of the expression of genes annotated in these pathways according to GO among the tumors under study.

In summary, large-scale survey transcript profiling of individual bladder tumors using cDNA array analysis has contributed to a biologically oriented classification of bladder cancer. Carcinomas *in situ* and papillary superficial lesions were shown to display differential expression profiles. Organ-confined invasive tumors and those developing regional or distant metastasis also displayed differential gene expression profiles. The combination of cDNA and tissue microarrays has facilitated identification and validation of novel molecular targets of potential clinical significance. Clusters and individual targets have been shown to provide a novel means for molecular diagnosis and outcome prediction of patients with bladder cancer. Overall, molecular gene profiling has revealed molecular targets and pathways associated with bladder cancer progression.

References

1. Cancer Facts and Figures 2000. American Cancer Society, 2000
2. Reuter VE, Melamed MR: The lower urinary tract. Diagnostic Surgical Pathology. Edited by SS Sternberg. New York, Raven Press, 1989, pp 1355–1392
3. Dalbagni G, Presti J, Reuter V, Fair WR, Cordon-Cardo C: Genetic alterations in bladder cancer. Lancet 1993, 324:469–471
4. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G: Making and reading microarrays. Nat Genet 1999, 21:15–19
5. Wang E, Miller LD, Ohnmacht GA, Liu ET, Marincola FM: High-fidelity mRNA amplification for gene profiling. Nature Biotechnol 2000, 18: 457–459
6. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998, 95:14863–14868
7. Felsenstein J: Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985, 39:783–791
8. Dawson-Saunders B, Trapp RG: Estimating and comparing means. Basic and Clinical Biostatistics, ed 2. Norwalk, Appleton & Lange, 1994, pp 110–112
9. Li W, Yang Y: How many genes are needed for a discriminant microarray data analysis? Methods of Microarray Data Analysis. Edited by SM Lin, KF Johnson. Kluwer Academic, 2002, pp 137–150
10. Li W, Yang Y: Zipf's law in importance of genes for cancer classification using microarray data. J Theor Biol 2002, 219:539–551
11. Reyment R, Joreskog KG: R-mode methods. Applied Factor Analysis in the Natural Sciences. Cambridge, Cambridge University Press, 1996, pp 89–128
12. De las Rivas J, Lozano JJ, Ortiz AR: Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny and deduced evolutionary patterns. Genome Res 2002, 12:567–583
13. Cristiani N, Shawe-Taylor J: Kernel-induced featured spaces. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, Cambridge University Press, 2000, pp 26–46
14. Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L: Large-scale protein annotation through gene ontology. Genome Res 2002, 12:785–794
15. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. Nat Genet 1999, 22:281–285
16. Boldrick JC, Alizadeh AA, Diehn M, Dudoit S, Liu CL, Belcher CE, Botstein D, Staudt LM, Brown PO, Relman DA: Stereotyped and specific gene expression programs in human innate immune responses to bacteria. Proc Natl Acad Sci USA 2002, 99:972–977
17. Axel Hoos, Urist MJ, Stojadinovic A, Mastorides S, Dudas ME, Leung DH, Kuo D, Brennan MF, Lewis JJ, Cordon-Cardo C: Validation of tissue microarrays for immunohistochemical profiling of cancer specimens using the example of human fibroblastic tumors. Am J Pathol 2001, 158:1245–1251
18. Lu M, Wikman F, Orntoft TF, Charytonowicz E, Rabbani F, Zhang Z, Dalbagni G, Pohar KS, Yu G, Cordon-Cardo C: Impact of alterations affecting the p53 pathway in bladder cancer on clinical outcome, assessed by conventional and array-based methods. Clin Cancer Res 2002, 8:171–179
19. Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987, 4:406–425
20. Cordon-Cardo C, Zhang ZF, Dalbagni G, Drobnjak M, Charytonowicz E, Hu SX, Xu HJ, Reuter VE, Benedict WF: Cooperative effects of p53 and pRB alterations in primary superficial bladder tumors. Cancer Res 1997, 57:1217–1221
21. Thykjaer T, Workman C, Kruhoffer M, Demtroder K, Wolf H, Andersen LD, Frederiksen CM, Knudsen S, Orntoft TF: Identification of gene expression patterns in superficial and invasive human bladder cancer. Cancer Res 2001, 61:2492–2499
22. Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dudoit S, Wolf H, Orntoft TF: Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet 2003, 33:90–96
23. Sanchez-Carbayo M, Socci ND, Charytonowicz E, Lu M, Prystowsky M, Childs G, Cordon-Cardo C: Molecular profiling of bladder cancer using cDNA microarrays: defining histogenesis and biological phenotypes. Cancer Res 2002, 62:6973–6980
24. Stein JP, Ginsberg DA, Grossfeld GD, Chatterjee SJ, Esrig D, Dickinson MG, Groshen S, Taylor CR, Jones PA, Skinner DG, Cote RJ: Effect of p21WAF1/CIP1 expression on tumor progression in bladder cancer. J Natl Cancer Inst 1998, 90:1072–1079
25. Makiyama K, Masuda M, Takano Y, Iki M, Asakura T, Suwa Y, Noguchi S, Hosaka M: Cyclin E overexpression in transitional cell carcinoma of the bladder. Cancer Lett 2000, 151:193–198
26. Richter J, Wagner U, Kononen J, Fijan A, Bruderer J, Schmid U, Ackermann D, Maurer R, Alund G, Knonagel H, Rist M, Wilber K, Anabitarte M, Hering F, Hardmeier T, Schonenberger A, Flury R, Jager P, Fehr JL, Schraml P, Moch H, Mihatsch MJ, Gasser T, Kallioniemi OP, Sauter G: High-throughput tissue microarray analysis of cyclin E gene amplification and overexpression in urinary bladder cancer. Am J Pathol 2000, 157:787–794
27. Gluzman-Poltorak Z, Cohen T, Herzog Y, Neufeld G: Neuropilin-2 is a receptor for the vascular endothelial growth factor (VEGF) forms VEGF-145 and VEGF-165. J Biol Chem 2000, 275:18040–18045
28. Banerjee SK, Zoubine MN, Tran TM, Weston AP, Campbell DR: Overexpression of vascular endothelial growth factor164 and its coreceptor neuropilin-1 in estrogen-induced rat pituitary tumors and GH3 rat pituitary tumor cells. Int J Oncol 2000, 16:253–260
29. Handa A, Tokunaga T, Tsuchida T, Lee YH, Kijima H, Yamazaki H, Ueyama Y, Fukuda H, Nakamura M: Neuropilin-2 expression affects the increased vascularization and is a prognostic factor in osteosarcoma. Int J Oncol 2000, 17:291–295
30. Latil A, Bieche I, Pesche S, Valeri A, Fournier G, Cussenot O, Lidereau R: VEGF overexpression in clinically localized prostate tumors and neuropilin-1 overexpression in metastatic forms. Int J Cancer 2000, 89:167–171
31. Tomizawa Y, Sekido Y, Kondo M, Gao B, Yokota J, Roche J, Drabkin H, Lerman MI, Gazdar AF, Minna JD: Inhibition of lung cancer cell growth and induction of apoptosis after reexpression of 3p21.3 candidate tumor suppressor gene SEMA3B. Proc Natl Acad Sci USA 2001, 98:13954–13959
32. Harnden P, Allam A, Joyce AD, Patel A, Selby P, Southgate J: Cytokeratin 20 expression by non-invasive transitional cell carcinomas: potential for distinguishing recurrent from non-recurrent disease. Histopathology 1995, 27:169–174
33. Buchumensky V, Klein A, Zemer R, Kessler OJ, Zimlichman S, Nissenkorn I: Cytokeratin 20: a new marker for early detection of bladder cell carcinoma? J Urol 1998, 160:1971–1974

34. Harnden P, Mahmood N, Southgate J: Expression of cytokeratin 20 redefines urothelial papillomas of the bladder. Lancet 1999, 353:974–977
35. Kim JW, Moon AR, Kim JH, Yoon SY, Oh GT, Choe YK, Choe IS: Up-regulation of ninjurin expression in human hepatocellular carcinoma associated with cirrhosis and chronic viral hepatitis. Mol Cell 2001, 11:151–157
36. Chen JS, Coustan-Smith E, Suzuki T, Neale GA, Mihara K, Pui CH, Campana D: Identification of novel markers for monitoring minimal residual disease in acute lymphoblastic leukemia. Blood 2001, 97:2115–2120
37. Kannan K, Amariglio N, Rechavi G, Jakob-Hirsch J, Kela I, Kaminski N, Getz G, Domany E, Givol D: DNA microarrays identification of primary and secondary target genes regulated by p53. Oncogene 2001, 20:2225–2234
38. Garkavtsev I, Kazarov A, Gudkov A, Riabowol K: Suppression of the novel growth inhibitor p33ING1 promotes neoplastic transformation. Nat Genet 1996, 14:415–420
39. Ohmori M, Nagai M, Tasaka T, Koeffler HP, Toyama T, Riabowol K, Takahara J: Decreased expression of p33ING1 mRNA in lymphoid malignancies. Am J Hematol 1999, 62:118–119
40. Cheung Jr KJ, Bush JA, Jia W, Li G: Expression of the novel tumour suppressor p33(ING1) is independent of p53. Br J Cancer 2000, 83:1468–1472