BMC Bioinformatics

Research article

Expression profiling of human renal carcinomas with functional taxonomic analysis

Michael A Gieseg*[1], Theresa Cody[1], Michael Z Man[1], Steven J Madore[1], Mark A Rubin[2] and Eric P Kaldjian[1,2]

Address: [1]Pfizer Global Research & Development, 2800 Plymouth Rd, Ann Arbor, Michigan 48105, USA and [2]Department of Pathology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA

E-mail: Michael A Gieseg* - Michael.Gieseg@pfizer.com; Theresa Cody - Theresa.Cody@pfizer.com; Michael Z Man - Michael.Man@pfizer.com; Steven J Madore - Steven.Madore@pfizer.com; Mark A Rubin - marubin@umich.edu; Eric P Kaldjian - Eric.Kaldjian@pfizer.com

*Corresponding author

Abstract

Background: Molecular characterization has contributed to the understanding of the inception, progression, treatment and prognosis of cancer. Nucleic acid array-based technologies extend molecular characterization of tumors to thousands of gene products. To effectively discriminate between tumor sub-types, reliable laboratory techniques and analytic methods are required.

Results: We derived mRNA expression profiles from 21 human tissue samples (eight normal kidneys and 13 kidney tumors) and two pooled samples using the Affymetrix GeneChip platform. A panel of ten clustering algorithms combined with four data pre-processing methods identified a consensus cluster dendrogram in 18 of 40 analyses and of these 16 used a logarithmic transformation. Within the consensus dendrogram the expression profiles of the samples grouped according to tissue type; clear cell and chromophobe carcinomas displayed distinctly different gene expression patterns. By using a rigorous statistical selection based method we identified 355 genes that showed significant ($p < 0.001$) gene expression changes in clear cell renal carcinomas compared to normal kidney. These genes were classified with a tool to conceptualize expression patterns called "Functional Taxonomy". Each tumor type had a distinct "signature," with a high number of genes in the categories of Metabolism, Signal Transduction, and Cellular and Matrix Organization and Adhesion.

Conclusions: Affymetrix GeneChip profiling differentiated clear cell and chromophobe carcinomas from one another and from normal kidney cortex. Clustering methods that used logarithmic transformation of data sets produced dendrograms consistent with the sample biology. Functional taxonomy provided a practical approach to the interpretation of gene expression data.

Background

Pathologic classification of tumors has been traditionally based on microscopic appearance. Although morphology often correlates with natural history of disease, tumors of a given pattern may have a broad prognostic range and different responses to treatment. Molecular methods, such as the evaluation of hormone receptors in breast carcinoma, have been effectively employed to further charac-

terize tumors [1]. Nucleic acid array-based technologies extend molecular characterization by providing a biochemical snapshot, or profile, of cellular activity that encompasses thousands of gene products [2]. Potential applications beyond diagnosis and prognosis are diverse, and include treatment response stratification of patients in clinical trials, assessment of relevance to human safety of drug-associated tumors in animal carcinogenicity studies, and the development of more pertinent animal xenograft models of cancer therapy. Successful application of array-based tools depends on establishing robust laboratory and computational methods that effectively and reliably discriminate between tumor types. Recent reports have demonstrated the power of such tools to distinguish between clinically meaningful subsets of cancer [3,4].

Renal cell carcinoma (RCCa) represents approximately 3% of all human malignancies with an incidence of 7 per 100,000 individuals. Of these individuals about 40% present with metastatic disease and a further third will develop distant metastases during the postoperative course. The most effective therapy for RCCa localized to the kidney is surgery and a metastatic tumor is practically incurable. There is a low response to biological modifiers and the treatment is generally only palliative [5].

We evaluated the RNA expression profiles of renal cell carcinomas (RCCa) using the Affymetrix GeneChip platform, comparing mRNA expression profiles from a total of 21 human tissue samples and two pooled samples. The 21 samples consisted of eight normal kidneys, nine clear cell carcinomas (CC), two chromophobe carcinomas (Chr), one urothelial carcinoma and one metanephric adenoma. Expression profiles from a pilot data set of ten samples were analyzed using multiple clustering algorithms. Genes were then selected from the pilot data set using a fold-change criteria or from all of the normal and CC samples using a p-value. Genes that were identified from both the pilot and the complete data set were categorized according to a hierarchical classification scheme based on functional attributes of encoded proteins.

Results
Clustering of the pilot data set
The gene expression data from a pilot data set that consisted of ten samples (patients 1 – 4, consisting of two CC, two Chr, four normals, two pooled samples) were analyzed using hierarchical clustering to identify structure within the data set. Pooled samples were included to determine whether a combined sample yielded an expression profile representative of the individual samples. To determine the relatedness of the samples, clustering analysis was performed. Ten different clustering algorithms using four methods of pre-processing the data sets were applied to identify the most consistent sample-clustering

pattern. The rationale behind this approach was to avoid the bias inherent in any single clustering method and to determine the most appropriate clustering method for this data type. Genes that were considered "present" above background by the Affymetrix software in at least one of the samples were included in the analyses. This reduced the data set to 4571 genes per sample. The 40 sample clusters were evaluated to see if their dendrograms fitted the expected sample biology (Table 2). The most consistent cluster dendrogram, present in 18 of 40 analyses, did indeed match the sample biology (Figure 2A) and for 16 of these analyses a logarithmic transformation was used. In the consensus dendrogram, the two CC, two Chr and four normal samples each clustered in separate groups. Interestingly, the dendrogram suggested that the expression pattern of the normal samples was more similar to the CC than to the Chr samples. As expected, the pooled normal sample clustered with the normal kidney samples. The pooled tumor sample clustered more closely to the CC than to the Chr, possibly due to the greater similarity between the two CC, and consequent weighting of the pooled sample toward the more uniform CC expression profile.
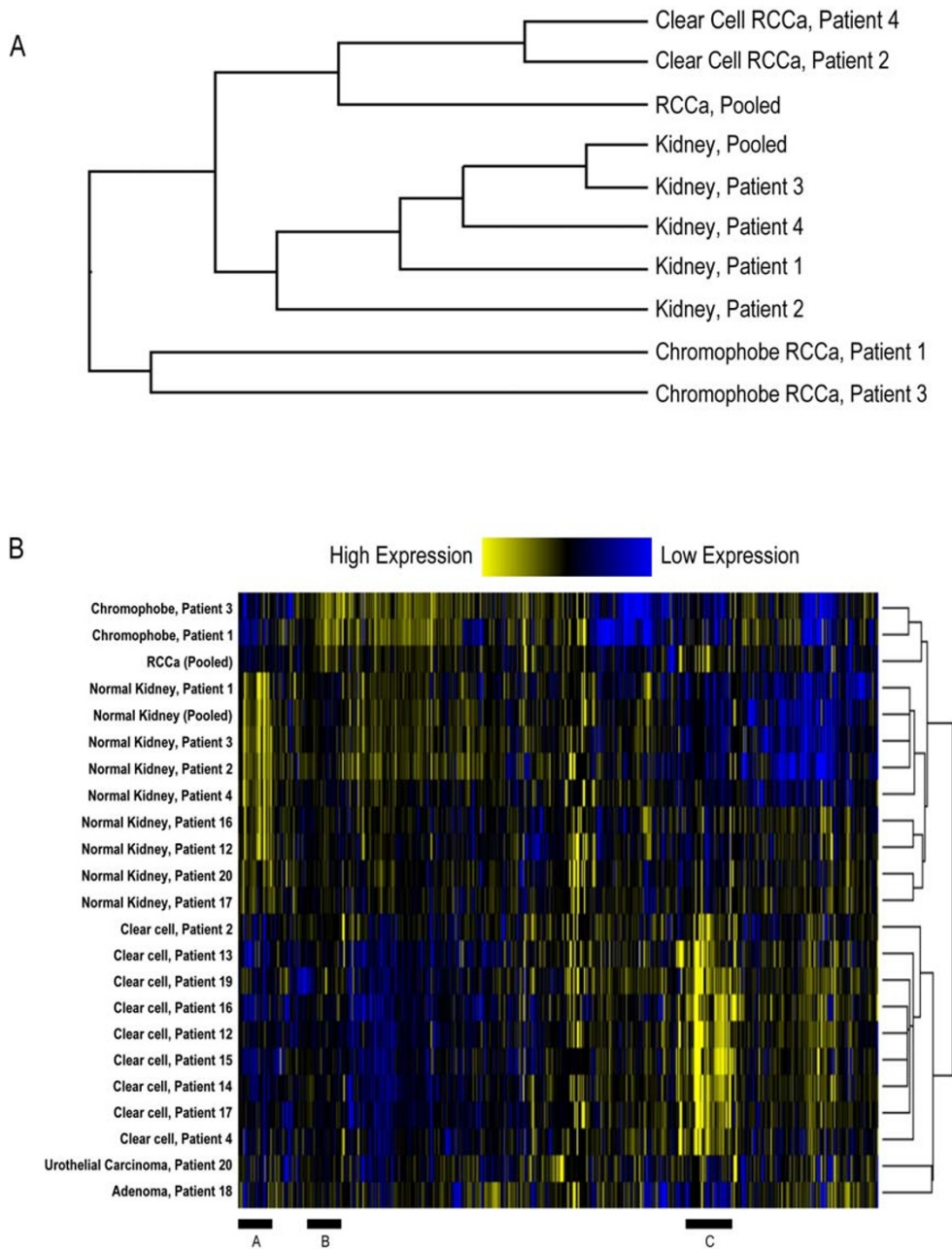
Clustering of the complete data set
To expand our data set we obtained a further 13 human tissue samples (patients 12–20), including four normal kidneys and nine kidney tumors. These were profiled in the same manner as the first 10 samples. From this combined data set (23 samples) we selected genes classified as Present (see Methods) at least once (5372 genes) and then clustered the log-transformed data with average linkage analysis. This method had previously produced a dendrogram that matched the expected sample biology in the pilot data set. The sample dendrogram (Figure 2B) showed that the relatedness of the samples was similar to that observed with the pilot data set. As seen previously in the pilot data set the normal samples clustered together off a single node. However the Chr also clustered off this node and now appeared more similar to normal samples than CC. The two CC included in the pilot data set (patients 2 and 4) now clustered within a larger CC group that include the additional CC samples. From this node there also appeared to be two outlier samples (patients 18 and 20). Pathology reports on these samples revealed that these were not RCCa samples but instead patient 20 was a papillary urothelial carcinoma and the sample from patient 18 was not a carcinoma but a benign metanephric adenoma.

There were distinct patterns visible in the gene cluster that were conserved in the CC samples (Figure 2B, zone C), or the Chr samples (Figure 2B, zone B), or the normal kidney samples (Figure 2B, zone A). These patterns indicated that each subtype of tumor expressed a common set of genes
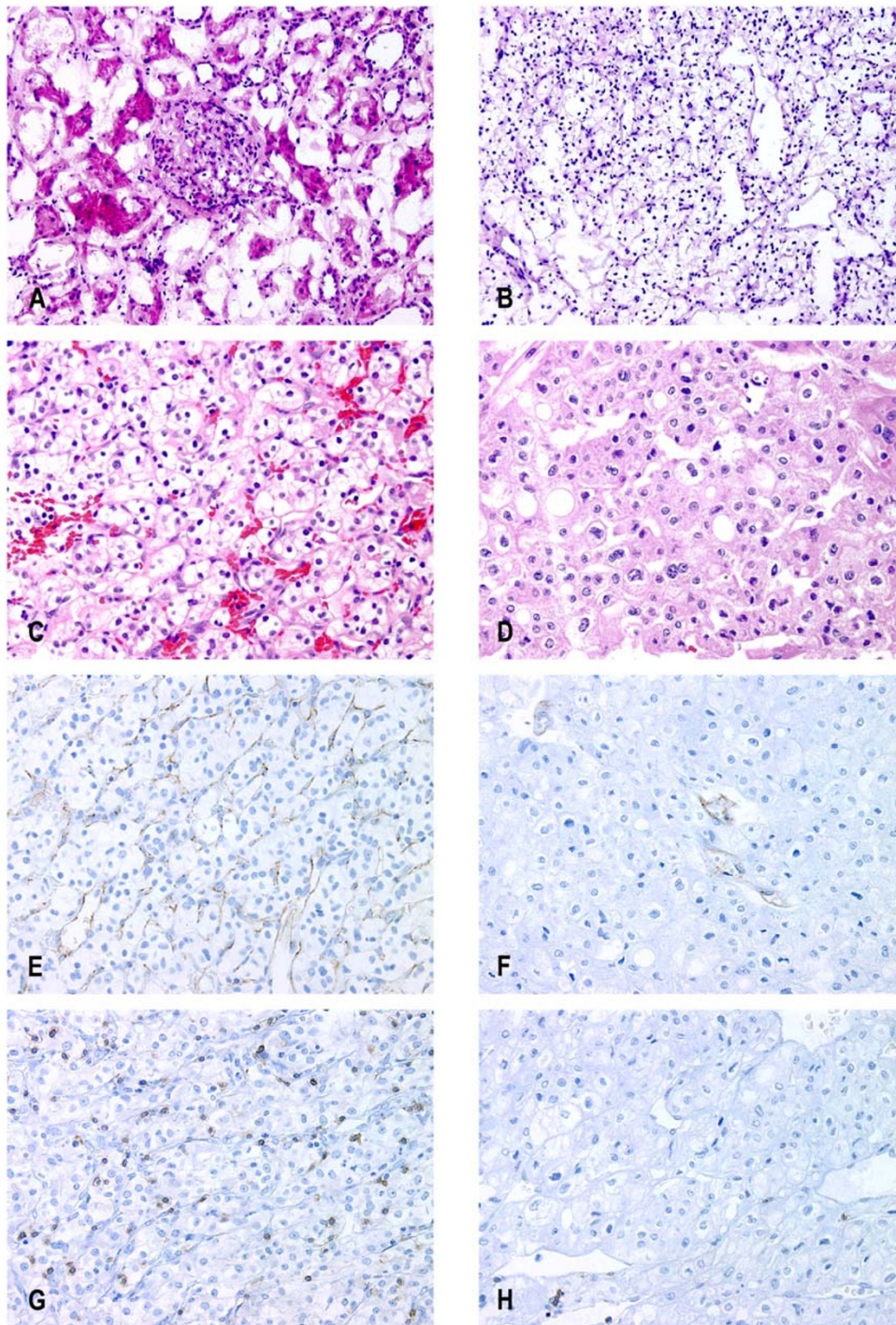
Figure 1
Histology of tissue samples. Frozen sections of normal kidney with glomerulus in center (**A**) and clear cell RCCa (**B**). Representative paraffin sections of clear cell (**C, E, G**) and chromophobe RCCa (**D, F, H**). Immunohistochemical stains for CD 31 (**E** and **F**) and CD 3 (**G** and **H**). Note that the clear cell tumor has an the extensive network of CD 31-positive vessels (**E**) and numerous scattered CD 3-positive lymphoctyes (**G**). In contrast, in the chromophobe there is a relative paucity of vessels (**F**) and few T cells (**H**). Original magnification:**A**, **B** 100× **C** – **H** 200×. Staining: **A** – **D**, hematoxylin and eosin; **E** – **H**, diaminobenzidine immunoperoxidase with hematoxylin.

Figure 2
Sample cluster of the pilot data set (A) and sample and gene cluster of the complete data set (B). The sample cluster dendrogram form the pilot data set **(A)** uses the average linkage clustering method and represents the most frequently occurring dendrogram out of the 40 derived for the data set. Horizontal distance indicates relatedness. The sample and gene cluster **(B)** from all 23 samples shows a colored representation of the gene expression data where columns are individual genes and rows are separate samples. The color in each cell of the table represents the median adjusted expression value of each gene.

Table 1: Details of immunohistochemistry

Antibody	Type	Dilution	Source	Digestion	Detection link
CD 31	Mouse monoclonal	1:20	DAKO	Protease 2, 32 min	Mouse (Vector)
CD 3	Rabbit polyclonal	1:1000	DAKO	Protease 2, 16 min	Rabbit (Vector)

Table 2: Summary of 40 sample clusters

	Data pre-processing method			
Clustering method	Unchanged	Log_{10} Transformation	Standardization to N(0,1)	Log_{10} transformation and Standardization to N(0,1)
Single linkage cluster analysis	N	Y	N	N
Average linkage cluster analysis	N	Y	N	Y
Complete linkage cluster analysis	N	Y	N	Y
Ward's minimum variance cluster analysis	N	Y	N	Y
Centroid hierarchical cluster analysis	N	Y	N	N
Density linkage cluster analysis	N	N	N	N
Flexible-beta cluster analysis	N	Y	N	Y
McQuitty's similarity analysis	N	Y	N	Y
Oblique principle component clustering, correlation P	N	Y	N	Y
Oblique principle component clustering, correlation S	Y	Y	Y	Y
Number of methods with cluster dendrograms that matched the sample biology	1	9	1	7

"Y" indicates the dendrogram matched the sample biology, "N" indicates it did not

that could be selected and further characterized. The urothelial carcinoma and metanephric adenoma appeared to share few of these genes commonly regulated between the other tissue types.

Functional taxonomy: Genes differentially regulated in CC and Chr

To select genes that were changed between CC and Chr in our pilot data set we used an arbitrary cutoff of 2 fold-change units in combination with the Affymetrix difference call (see Methods). Genes were selected according to criteria described in Table 3, which lists the number of selected genes in each of eight categories. A total of 456 genes were selected by these criteria.

In order to understand the molecular differences between Chr and CC RCCa on a broader scale, we developed a gene categorization system ("Functional Taxonomy," see Methods) in which genes were hierarchically grouped according to the cellular function of their protein products. The number of genes within each primary category was tallied and plotted to generate a first-level "signature" (Figure 3A). The functional taxonomy signature provided a graphical method to rapidly visualize broad gene expression characteristics of the tumors. The cellular function categories that contained the largest number of the 456 genes were Signal Transduction (97 genes), Cellular and Matrix Organization and Adhesion (56 genes), Metabolism (54 genes), and Unclassified (65 genes). The CC and Chr samples demonstrated differences in the numbers or patterns of increased and decreased genes in these categories and their subcategories. Almost twice as many genes in Signal Transduction were increased in CC compared to Chr; a similar number were decreased in both types. Within Cellular and Matrix Organization and Adhesion, the expression of nearly four times as many genes was increased in CC compared to Chr, which decreased expression of about three times as many genes as did CC. Subcategori-

Table 3: Number of selected genes in each category from pilot data set

Selection criterion[a]	Number of genes
Increase in all tumors	10
Decrease in all tumors	59
Increase in Chr	76
Decrease in Chr	99
Increase in CC	136
Decrease in CC	40
Diametric discriminator 1 (increase in Chr, decrease in CC)	13
Diametric discriminator 2 (increase in CC, decrease in Chr)	23
Total	**456**

[a]Gene selection based on two fold-change threshold in tumor relative to normal.

zation (Figure 3C) revealed a predominance of genes coding for extracellular matrix proteins (16 genes) and cellular adhesion molecules (10 genes) that were increased in expression in CC.

Functional taxonomy: Genes differentially regulated in CC
To determine the difference between normal kidney samples and CC samples we used a rigorous statistical approach based upon a calculated p-value (see methods) and the combined data set consisting of eight normal kidney samples and nine CC samples. We identified 142 genes that were significantly upregulated in CC compared to normal kidneys and 213 that were significantly down regulated. To determine the biological significance/function of these genes we used Functional Taxonomy to classify them into 16 cellular groupings (Figure 4A). The first level signature showed similar trends to the CC and Chr comparison (Figure 3A). The four categories that contained the most genes were Signal Transduction (78 genes), Cellular and Matrix Organization and Adhesion (47 genes), Metabolism (40 genes), and Unclassified (51 genes). These were the same categories that showed the most gene changes in the pilot data set. Subcategorization of both categories showed trends within CC samples that were similar to those observed in the pilot data set. Within the Signal Transduction category there were many genes associated with ligands, receptors and cytosolic factors (Figure 4B and Table 4). Gene expression changes within the Cellular Matrix Organization and Adhesion category were focused on extracellular matrix genes and cellular adhesion molecules (Figure 4C and Table 5).

Immunohistochemistry
The selected gene lists were reviewed for genes that corresponded to proteins that could be evaluated with immunohistochemistry. The data sets revealed that mRNA transcripts for CD 31 (PECAM) and the T-cell receptor

beta chain were increased in CC but not Chr. Since antibodies to CD 31 and CD 3 (which forms a complex with the T cell receptor) are reactive in fixed tissues, we used them to stain the tumors. CD 31 was present in CC in a prominent dense branching network of fine vessels surrounding the tumor cells (Figure 1E). In contrast, Chr had few CD 31-positive vessels present (Figure 1F). CD 3 stained numerous T lymphocytes scattered throughout the CC tumors (Figure 1G). These were not initially apparent in the hematoxylin and eosin sections, probably due to the similarity in size and appearance of the tumor cell nuclei and the lymphocytes. In contrast, there were almost no T cells in the Chr tumors (Figure 1H). These results indicated concordance between the mRNA expression profiles and the pathobiology of the RCCa tumor sub-types.

Discussion
Expression profiling of kidney tumors using the Affymetrix GeneChip distinctly separated four different tumors from each other, as well as from normal kidney cortex. This finding is consistent with the morphologic, karyotypic and clinical outcome differences between these tumor types [6,7]. There are many sample-clustering methods that may be applied to expression microarray data, none of which can be conclusively called "correct", since each algorithm makes different assumptions regarding the nature of the data. We used ten clustering methods combined with four ways of pre-processing the data sets to eliminate, or at least reduce bias in a pilot data set. The smaller pilot data set was used to simplify the interpretation of the results. A common cluster dendrogram was produced by 18 of 40 methods; 16 of these were from the 20 that employed logarithmic transformation of the data sets. The pattern was consistent with the biology of the sample with normal kidney, CC and Chr samples each grouping together (Figure 2A). That a logarithmic transformation gave the most meaningful cluster dendrograms
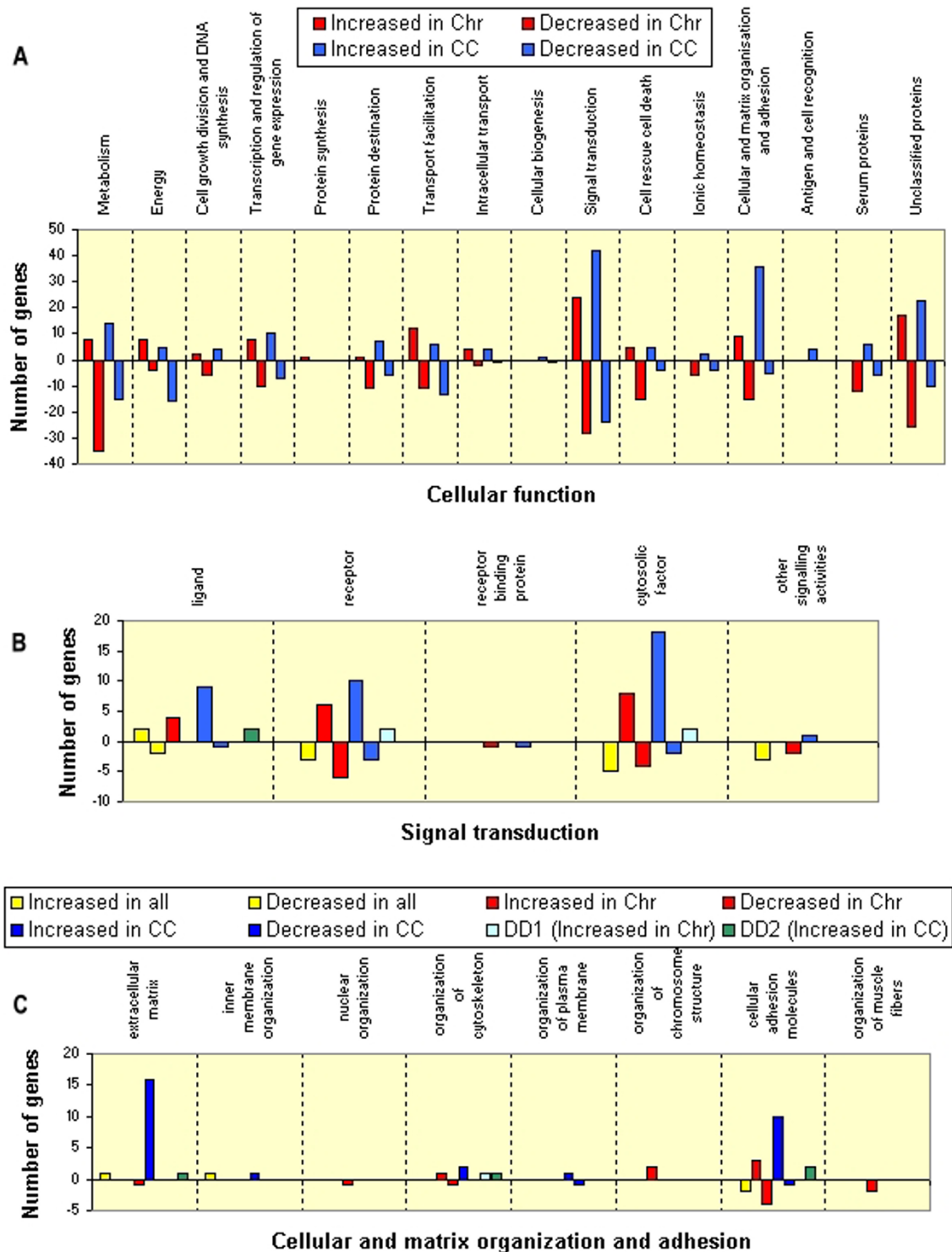
Figure 3
Functional taxonomy signatures of clear cell and chromophobe RCCa. The 456 genes that were selected according to the criteria in Table 3 were annotated and placed into 16 cellular function categories **(A)**. The number of genes in each cellular function category is displayed. A positive value indicates the genes were increased in expression while a negative value indicates the genes were decreased in expression. The second-level signatures for the subcategories of Signal Transduction, and of Cellular and Matrix Organization and Adhesion are shown in panels **(B)** and **(C)** respectively.
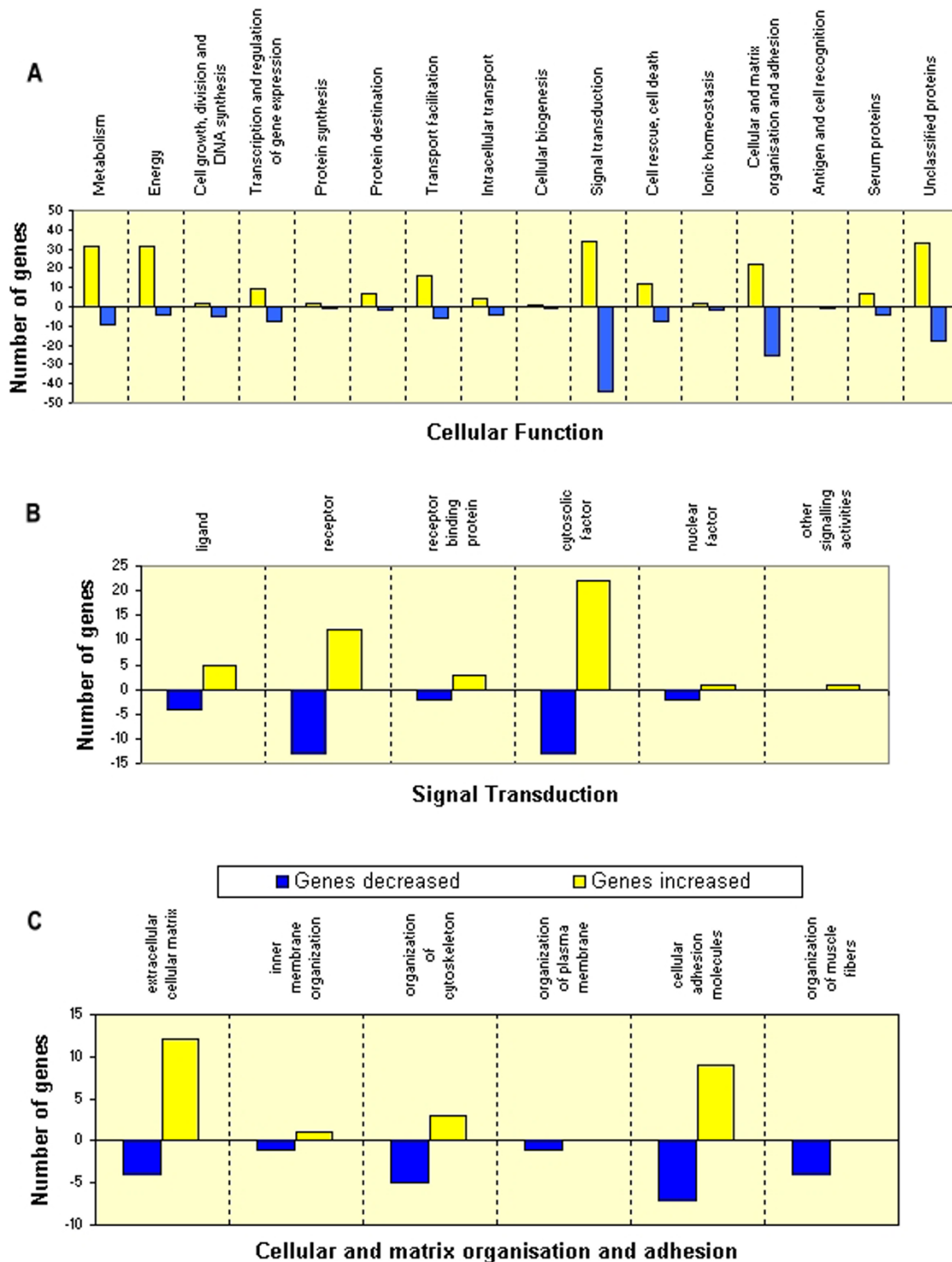
Figure 4
Functional taxonomy signatures of clear cell RCCa. The 355 genes that were selected from the complete data set were annotated and placed into 16 cellular function categories **(A)**. The number of genes in each cellular function category is displayed. A positive value indicates the genes were increased in expression while a negative value indicates the genes were decreased in expression. The second-level signatures for the subcategories of Signal Transduction, and of Cellular and Matrix Organization and Adhesion are shown in panels **(B)** and **(C)** respectively.

Table 4: Signal Transduction

	Ligand	Receptor	Receptor binding protein	Cytosolic factor	Nuclear factor	Other signaling activities
Increased in CC	*TNFSF7* *VEGF* *SCYA5* *TGFB1* *SELPLG*	*CXCR4* *FCER1G* *IFNGR1* *CSF1R* *LY117* *IFNAR2* *TNFRSF1A* *C3AR1* *CD3D* *CD14* *TNFRSF6* *LGALS9*	*IGFBP3* *SHC1* *NCK1*	*HPCAL1* *INHBB* *PTPN12* *CBLB* *GNAI2* *HCLS1* *GUCY1B3* *ARHGEF6* *ADCY7* *MTAP44** *GIT2* *ARHGDIB* *RRAD* *HCK* *GBP2* *PTPNS1* *NMI* *PRKCL1* *ANXA1* *PPP1CC* *RASA3** *RIT*	*IFI16*	*LCP2*
Decreased in CC	*EGF* *FGF1* *FGF9* *SEMA3B*	*NR3C2* *FGFR3* *GABRA3* *PTGER3* *NPY1R* *PCDH1* *GPC3* *PDGFRA* *EPOR* *CASR* *FOLR1* *ERBB2* *NR2F6*	*GRB7* *TRAF4*	*MIS1* *RAP1GA1* *GNAI1* *PVALB* *PRKCZ* *PPP2CB* *APLP2* *ANXA11* *RALBP1* *CRABP2* *GPS1* *PDE1A* *BCR*	*FOXI1* *HOXB3*	

* Not an approved HUGO name

is consistent with the distribution of the untransformed expression data being skewed to the left because the majority of genes have low expression levels. Standardization of the data assigns equal weight to each gene and, hence, increases the contribution of unreliable low expression genes. The use of logarithmic transformation, on the other hand, improves the spread of the data so the distribution is close to normal. It also re-adjusts the weight for each gene. For example, genes with high expression levels, which might be unreliable or biased due to saturation, will have lower weights in distance calculation. Therefore, the logarithmic transformation improves the calculation of distance for the subsequential clustering algorithms

and leads to uncovering the biological meaningful pattern within the data.

A comparison of the dendrograms from the pilot data set and complete data set reveals some surprising changes. In general the major structure of the dendrogram remained the same, CC, Chr and normal kidney all grouped separately. However, in the pilot data set the CC were more similar to normal kidney than Chr, while in the complete data set Chr were more similar to normal kidney. It is unclear why this larger data set changed the dendrogram and suggests that the subtle structure in the dendrogram was not as robust as it appeared. With fewer Chr compared to

CC it is not possible to draw any strong conclusions about relatedness of the Chr samples.

In order to visualize the functional patterns associated with a particular set of selected genes we used a simple, semi-hierarchical system to categorize genes according to the function of the proteins they encode, that we call Functional Taxonomy. There are challenges associated with the partly subjective nature of categorization of gene function, such as where to place a single gene product that is involved in several cellular tasks. Ideally, the categorization should consider multiple attributes of a protein. To this end, we propose three complementary classification schemes: (1) biochemical function, which categorizes according to molecular activity; (2) cellular function, which categorizes according to biological role at a cellular level; (3) tissue function, which categorizes according to anatomic or organ system location. In this paper we have visualized profiling results using the second of these schemes (cellular function) at three levels: primary categories, secondary categories, and individual genes (see Figure 3 and 4, Table 4 and 5). We have found Functional Taxonomy to be a useful visualization tool for understanding the differences in gene expression patterns between CC and Chr tumors. This system is similar in concept to what is currently being developed by the Gene Ontology Consortium [10].

The cellular function signatures of nine CC and two Chr revealed that the greatest number of gene expression changes for both tumor types occurred in the categories of Signal Transduction, Cellular and Matrix Organization and Adhesion, and Metabolism. This is consistent with current theories of neoplasia, which hypothesize that tumor cells modify their signaling pathways, establish new contacts with an altered extracellular matrix, and refashion their metabolic machinery.

There exists considerable literature on the expressed genes and gene products associated with RCCa. Using the selected gene sets from Table 3 and the p-values and fold-change values calculated from the eight normal kidneys and nine CC, we looked for concordance between our results and published reports. The genes *CA 9* (*carbonic anhydrase IX*), *CCND1* (*cyclin D1*), *CDH2* (*N-Cadherin*), *EGFR* (*epidermal growth factor receptor*) and *TGFA* (*tranforming growth factor alpha*) all showed increases in CC expression that matched the literature and had p-values ≤ 0.0061 [11–15]. The observed decrease in *CDH1* (*E-cadherin*) in CC (p-value = 0.0045) also matched previously published reports, as did the decrease in *VIM* (*vimentin*) expression in Chr RCCa [13,16]. *VIM* was also found to be increased in CC with a p-value = 0.0045, which was consistent with the literature. We detected a small increase in expression of *ICAM1* (*intercellular adhesion molecule 1*) in

CC (Fold-change = 1.9, p-value = 0.0081), which was also consistent with the literature [17].

The expression results for the genes *JUN* (*c-jun*) and *VHL* (*von Hippel-Lindau*) did not match the literature [18,19]. Nor did the result for *KRT7* (*cytokeratin 7*), which has been shown to be overexpressed in Chr [20]. Instead we found *KRT7* to be strongly repressed in CC (fold-change = -5.1, p-value = 0.0009). Yet, expression profiling using nucleic acid microarrays does not necessarily correlate with other forms of analysis for all genes [21,22]. This may be especially true when altered expression of a gene is reported to be present in a subset of a population of tumors, since a small sample number (as are the Chr samples in this study) may not include the alteration.

In the case of CD 31 and the T cell receptor beta chain, expression profiling results were concordant with immunohistochemical analysis of the tumors. The prevalence of the scattered T cells within the CC tumors was somewhat surprising, but entirely consistent with the biology of response to a treatment for RCCa, interleukin 2 (IL-2), since IL-2 activates lymphocytes against the tumor [23].

During preparation of this manuscript, an expression profiling study of seven renal neoplasms (four CC, 2 oncocytomas, and one Chr) was reported [24]. This study employed a different platform (Incyte glass slide cDNA microarray) and hybridization method (competitive tumor/normal binding), and used related but not identical gene selection criteria (two fold-change in expression versus normal kidney in at least two of the seven tumors). The study identified 189 genes that were differentially expressed in at least two tumors, and this gene set was also able to distinguish between CC and Chr tumor types. We suspect that a greater number of Chr-associated genes would have been selected in their study had there been at least two Chr samples, since a gene altered in expression only in the single Chr, but not in any of the oncocytomas or the CC, would not have been identified by the selection criteria.

Conclusion
The results of the present study demonstrate the power of Affymetrix GeneChip expression profiling to differentiate between morphologically distinct tissues that are descended from a common organ. In addition, they demonstrate the value of functional cataloging selected genes and visualizing the result in a graphical format.

Methods
Sample isolation, histology and immunohistochemistry
Renal cell carcinoma (RCCa) samples were collected from patients undergoing radical nephrectomy at the University of Michigan Medical Center. All samples and associated

Table 5: Cellular and matrix organization and adhesion

	Extra cellular matrix	Inner membrane organization	Organization of the cytoskeleton	Organization of the plasma membrane	Cellular adhesion molecules	Organization of muscle fibers
Increased in CC	P4HA1 COL4A2 COL1A1 PLOD2 CAV1 COL1A1 COL1A2 P4HB PLOD COL4A1 SERPINB1 FN1	CAV2	LCP1 TMSB10 TUBB		SPARC ITGAM PECAM1 LAMA4 ITGA5 GJA1 CLECSF2 ITGB2 TGFBI	
Decreased in CC	APM1* MFAP4 HYAL1 LAD1	PTPN3	PFN2 ABLIM KRT7 ARHGEF5 LLGL2	PALM	TM4SF3 CLDN10 PLXNB1 MUC1 TACSTD1 LU CDH1	MYL6 TNNC1 MYH10 ADF*

* Not an approved HUGO name

clinical data were obtained with Institutional Review Board approval. A total of 21 tissue samples (eight normal kidneys and 13 tumors) were obtained from 13 patients. Patients ranged in age from 40 to 83 years with four male patients and nine female patients. Nine patients were diagnosed with clear cell carcinomas (CC), two with chromophobe cell carcinomas (Chr), one with a papillary urothelial carcinoma, and one with a metanephric adenoma. For eight of the patients the resection specimens included unaffected kidney (termed "normal kidney") as well as tumor tissue. Tissue samples were fixed in formalin, or were immediately frozen at -80°C in optimal cutting temperature embedding medium (OCT, Sakura). Cryostat sections were prepared to confirm suitability for profiling (Figure 1A and 1B); only viable-appearing tissues were processed. Care was taken to ensure that normal tissue was not contaminated with tumor tissue and vice versa. Normal kidney samples were uniformly taken from renal cortex. Paraffin-embedded tissues were stained with hematoxylin and eosin for diagnostic evaluation. Representative morphologies are shown in Figure 1C and 1D. Immunohistochemical stains were performed on fixed tissues. Sections were deparaffinized, rehydrated and treated with 3% hydrogen peroxide. Stains were performed using an automated avidin-biotin complex method according to the manufacturer's protocol (Nexes IHC Staining System, Ventana Medical Systems), with details as indicated in Table 1.

RNA isolation
Total RNA was prepared from each sample (eight normal kidneys, 13 tumors). In addition, two pooled samples were made by mixing equal quantities of RNA from the kidney samples of patients 1 – 4 together, and by mixing equal quantities of RNA from the four RCCa samples also from patients 1 – 4. The frozen tissues were warmed briefly, allowing the OCT compound to soften slightly so that it could be rapidly dissected away without the tissues thawing. The tissues were then pulverized using a frozen steel block and hammer. RNA was extracted using Trizol reagent (Life Technologies) according to the manufacturer's protocol.

RNA labeling and GeneChip hybridization
Biotinylated target RNA was prepared from 15 µg of total RNA using the Affymetrix protocol. Labeled cRNA was hybridized on the HuGeneFL Affymetrix GeneChip® containing probes for approximately 5600 mRNAs corresponding to genes of known sequence. Each hybridization included control RNA transcripts. The hybridization reactions were processed and scanned according to standard Affymetrix protocols.

Data Analysis
The Affymetrix Microarray Suite and Data Mining Tool were used to calculate average difference (gene expression) values, fold-change values and difference calls from

the GeneChip fluorescent intensity data. All GeneChips were normalized to avoid differences in fluorescent intensity as described in the *Affymetrix GeneChip Expression Analysis Manual (version 3.1)*. The average intensity of each gene chip was scaled to 600, where average intensity was calculated by averaging all the gene expression values of every probe set on the array excluding the highest and lowest 2% of the values. If the scaling factor did not fall between 0.3 and 2, the GeneChip data was not used. Fold-change values were calculated as the ratio of gene expression values between each tumor-normal sample pair from the same patient, with the normal sample values used as the baseline. The difference call was calculated using four variables: the number of probes pairs that have changed in a certain direction, the ratio of increased probe pairs over decreased probe pairs, the log average ratio change, and the difference in the number of probe pairs changed (either positive or negative). The default values (explained in the *Affymetrix Manual*) were used in the difference call decision matrix. The resulting data sets were manipulated and filtered in Microsoft Access according to various criteria (see results) for selection of genes that showed altered expression in the tumors.

For the pilot data set, sample clusters were prepared using the software program SAS (SAS Institute, Cary, NC) on the gene expression values. The following clustering algorithms were used: seven hierarchical clustering methods (single linkage, average linkage, complete linkage, Ward's minimum variance, density linkage, centroid, and flexible-beta), a k-mean method, and an oblique principle component method employing either (1) Pearson correlation or (2) Spearman correlation based on rank. Negative values were set equal to one. Data sets were processed as follows for the clustering analysis: unchanged (raw data), $\log_{10}$ transformation, standardization to N(0,1), and a combination of both $\log_{10}$ transformation and standardization to N(0,1). This yielded a total of 40 sample cluster dendrograms. For the complete data set (21 samples plus two pooled samples) gene and sample clusters were prepared using Cluster and visualized with TreeView [25], using $\log_{10}$-transformed and median-centered average difference values.

To select genes for further analysis from the pilot data set, a two fold-change threshold was used in combination with the Affymetrix difference call of either increase (I), marginal increase (MI), decrease (D) and marginal decrease (MD). For genes selected as either increased or decreased in a particular RCCa subtype both samples had to show concordance in their fold-change and difference call, and both samples from the other subtype had to have a difference call of no change (NC) or a call that was opposite to that of the first subtype.

To select genes that were significantly changed in CC (nine samples) verses normal kidney (eight samples) a non-parametric Wilcoxon test was used to calculate a p-value for each gene. Genes were then selected that conformed to all of the following criteria: p-value ≤ 0.001; a mean average difference value ≥ 200; fold-change ≥ 1.1 for genes increased in CC relative to normal kidney or fold-change ≤ -1.1 for genes decreased in CC relative to normal kidney.

All selected genes were then annotated using online resources, such as Medline, OMIM and GeneCards. The information was used to build a database that included HUGO (Human Genome Organization) designated names, function and expression information, and any tumor associations.

Given the absence of available tools for visualization of tumor gene expression patterns on a broad scale, we devised a classification system called "Functional Taxonomy" that categorizes genes according to various functional attributes of the proteins they encode. One such functional attribute is a protein's biological role at a cellular level. Using a mammalian modification of the MIPS *Saccharomyces cerevisiae* functional categories [26], (Munich Information Center for Protein Sequences, [www.mips.biochem.mpg.de]) the genes were placed into a hierarchical classification scheme containing 16 primary categories of cellular function. The number of genes that fell within each primary category was counted and plotted graphically to generate a first-level "signature" (see Figure 3A and 4A). Each of the primary categories was further divided into subcategories for a more detailed second-level visualization of the data (see Figures 3B, 3C, 4B and 4C).

Authors' Contributions

Author 1, MAG, carried out sample preparation, RNA isolation, data analysis, functional taxonomy analysis and drafted the manuscript. Author 2, TC, carried out the immunohistochemistry. Author 3, MZM, performed the clustering analysis. Author 4, SJM, directed the team who carried out the Affymetrix GeneChip hybridizations and initial data processing. Author 4 MAR provided the samples and reviewed the manuscript. Author 6 EPK conceived the study and participated it its design and assisted in writing the manuscript.

Acknowledgments

References

1. Fitzgibbons PL, Page DL, Weaver D, Thor AD, Allred DC, Clark GM, Ruby SG, O'Malley F, Simpson JF, Connolly JL, *et al*: **Prognostic fac-**

tors in breast cancer. **College of American Pathologists Consensus Statement 1999.** *Arch Pathol Lab Med* 2000, **124:**966-978
2. Lander ES: **Array of hope.** *Nat Genet* 1999, **21:**3-4
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, *et al:* **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511
4. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OPY, *et al:* **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344:**539-548
5. Kovacs G: **Molecular genetics of human renal cell tumors.** *Nephrol Dial Transplant* 1996, **11(Suppl 6):**62-65
6. Zambrano NR, Lubensky IA, Merino MJ, Linehan WM, Walther MM: **Histopathology and molecular genetics of renal tumors: toward unification of a classification system.** *J Urol* 1999, **162:**1246-1258
7. Wilhelm M, Krause U, Kovacs G: **Diagnosis and prognosis of renal-cell tumors: a molecular approach.** *World J Urol* 1995, **13:**143-148
8. Cohen C, McCue PA, Derose PB: **Histogenesis of renal cell carcinoma and renal oncocytoma. An immunohistochemical study.** *Cancer* 1988, **62:**1946-1951
9. Storkel S, Steart PV, Drenckhahn D, Thoenes W: **The human chromophobe cell renal carcinoma: its probable relation to intercalated cells of the collecting duct.** *Virchows Arch B Cell Pathol Incl Mol Pathol* 1989, **56:**237-245
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al:* **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29
11. Liao SY, Aurelio ON, Jan K, Zavada J, Stanbridge EJ: **Identification of the MN/CA9 protein as a reliable diagnostic biomarker of clear cell carcinoma of the kidney.** *Cancer Res* 1997, **57:**2827-2831
12. Lin BT, Brynes RK, Gelb AB, McCourty A, Amin MB, Medeiros LJ: **Cyclin D1 expression in renal carcinomas and oncocytomas: an immunohistochemical study.** *Mod Pathol* 1998, **11:**1075-1081
13. Taki A, Nakatani Y, Misugi K, Yao M, Nagashima Y: **Chromophobe renal cell carcinoma: an immunohistochemical study of 21 Japanese cases.** *Mod Pathol* 1999, **12:**310-317
14. Weidner U, Peter S, Strohmeyer T, Hussnatter R, Ackermann R, Sies H: **Inverse relationship of epidermal growth factor receptor and HER2/neu gene expression in human renal cell carcinoma.** *Cancer Res* 1990, **50:**4504-4509
15. Knebelmann B, Ananth S, Cohen HT, Sukhatme VP: **Transforming growth factor alpha is a target for the von Hippel-Lindau tumor suppressor.** *Cancer Res* 1998, **58:**226-231
16. Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP, Sauter G: **High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma.** *Am J Pathol* 1999, **154:**981-986
17. Markovic-Lipkovski J, Brasanac D, Todorovic V, Muller CA, Muller GA: **Immunomorphological characteristics of renal cell carcinoma.** *Histol Histopathol* 1995, **10:**651-659
18. Koo AS, Chiu R, Soong J, Dekernion JB, Belldegrun A: **The expression of C-jun and junB mRNA in renal cell cancer and in vitro regulation by transforming growth factor beta 1 and tumor necrosis factor alpha I.** *J Urol* 1992, **148:**1314-1318
19. Brauch H, Weirich G, Brieger J, Glavac D, Rodl H, Eichinger M, Feurer M, Weidt E, Puranakanitstha C, Neuhaus C: **VHL alterations in human clear cell renal cell carcinoma: association with advanced tumor stage and a novel hot spot mutation.** *Cancer Res* 2000, **60:**1942-1948
20. Leroy X, Moukassa D, Copin MC, Saint F, Mazeman E, Gosselin B: **Utility of cytokeratin 7 for distinguishing chromophobe renal cell carcinoma from renal oncocytoma.** *Eur Urol* 2000, **37:**484-487
21. Ishii M, Si Hashimoto, Tsutsumi S, Wada Y, Matsushima K, Kodama T, H Aburatani: **Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis.** *Genomics* 2000, **68:**136-143
22. Taniguchi M, Miura K, Iwao H, Yamanaka S: **Quantitative Assessment of DNA Microarrays-Comparison with Northern Blot Analyses.** *Genomics* 2001, **71:**34-39
23. Bukowski RM: **Cytokine combinations: therapeutic use in patients with advanced renal cell carcinoma.** *Semin Oncol* 2000, **27:**204-212
24. Young AN, Amin MB, Moreno CS, Lim SD, Cohen C, Petros JA, Marshall FF, Neish AS: **Expression profiling of renal epithelial neoplasms: a method for tumor classification and discovery of diagnostic molecular markers.** *Am J Pathol* 2001, **158:**1639-1651
25. 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
26. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C, *et al:* **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2000, **28:**37-40