Robust Classification of Renal Cell Carcinoma Based on Gene Expression Data and Predicted Cytogenetic Profiles

Kyle A. Furge,1 Kerry A. Lucas,1 Masayuki Takahashi,2,3 Jun Sugimura,2,4 Eric J. Kort,2 Hiro-omi Kanayama,3 Susumu Kagawa,5 Philip Hoekstra,6 John Curry,6 Ximing J. Yang,5 and Bin T. Teh3

1 Bioinformatics Special Program and 2 Laboratory of Cancer Genetics, Van Andel Research Institute, Grand Rapids, Michigan; 3 Department of Urology, School of Medicine, The University of Tokyo, Tokyo, Japan; 4 Department of Urology, School of Medicine, Iwate Medical University, Morioka, Japan; 5 Department of Pathology, University of Chicago Cancer Research Center, Chicago, Illinois; and 6 Division of Urology, Spectrum Health Hospital, Grand Rapids, Michigan

ABSTRACT

Renal cell carcinoma (RCC) is a heterogeneous disease that includes several histologically distinct subtypes. The most common RCC subtypes are clear cell, papillary, and chromophobe, and recent gene expression profiling studies suggest that classification of RCC based on transcriptional signatures could be beneficial. Traditionally, however, patterns of chromosomal alterations have been used to assist in the molecular classification of RCC. The purpose of this study was to determine whether it was possible to develop a classification model for the three major RCC subtypes that utilizes gene expression profiles as the bases for both molecular genetic and cytogenetic classification. Gene expression profiles were first used to build an expression-based RCC classifier. The RCC gene expression profiles were then examined for the presence of regional gene expression biases. Regional expression biases are genetic intervals that contain a disproportionate number of genes that are coordinately up- or down-regulated. The presence of a regional gene expression bias often indicates the presence of a chromosomal abnormality. In this study, we demonstrate an expression-based classifier can distinguish between the three most common RCC subtypes in 99% of cases (n = 73). We also demonstrate that detection of regional expression biases accurately identifies cytogenetic features common to RCC. Additionally, the in silico-derived cytogenetic profiles could be used to classify 81% of cases. Taken together, these data demonstrate that it is possible to construct a robust classification model for RCC using both transcriptional and cytogenetic features derived from a gene expression profile.

INTRODUCTION

There are >32,000 new cases of renal cell carcinoma (RCC) diagnosed each year in the United States, accounting for 3% of all adult malignancies (1). RCC is a clinicopathologically heterogeneous disease, traditionally subdivided into clear cell, granular cell, papillary, chromophobe, spindle cell, cystic, and collecting duct carcinoma subtypes based on morphological features according to the WHO International Histological Classification of Kidney Tumors (2). Clear cell RCC is the most common adult renal neoplasm, representing 70% of all renal neoplasms, and is thought to originate in the proximal tubules. Papillary RCC accounts for 10–15%, chromophobe RCC 4–6%, collecting duct carcinoma <1%, and unclassified lesions 4–5% of RCC. Spindle RCC, also called sarcomatoid RCC, is characterized by prominent spindle cell features and is thought to represent the high-grade end of all subgroups. Granular cell RCC, which is no longer considered a subtype in classification systems used at present, can often be reclassified into other subtypes (3).

To better understand the molecular mechanisms of RCC, several cytogenetic and molecular genetic approaches have been used to identify the genetic alterations underlying this disease (4). Conventional G-banding, loss of heterozygosity, and comparative genomic hybridization have all contributed to consistent and reliable correlation between chromosomal alterations and histological subtypes, leading to a proposed molecular subclassification model for RCC (5, 6). In this model, clear cell RCCs are characterized by deletion of chromosome 3p or a gain of 5q combined with deletions of two or more of chromosomes 6q, 8p, 9p, or 14q; papillary RCCs are characterized by a gain of two or more of chromosomes 3q, 7, 8, 12, 16, 17, or 20 and no 5p loss; finally, chromophobe RCCs are characterized by loss of two or more of chromosomes 1, 2, 6, 10, 13, or 17. In addition, some evidence suggests that a gain of chromosome 5q in clear cell tumors may lead to a more favorable outcome, whereas a loss of 14q may lead to a less favorable outcome.

More recently, several groups have used gene expression microarray analysis to profile the major RCC histological subtypes: papillary, clear cell, and chromophobe (7–10). Discovery-based hierarchical clustering has suggested that these three histological subtypes have distinct gene expression profiles. In addition, it has been demonstrated that gene expression profiling may distinguish between two clinically distinct subtypes of clear cell RCC: one with favorable outcome and one with poor outcome (11). In view of the long-established correlation between cytogenetics and histology and the emerging correlation between gene expression profiling and clinical parameters, it would be beneficial to develop methods that could generate both cytogenetic and gene expression information for the same set of samples. Transcriptional and cytogenetic data often yield complementary information; however, both types of data are often not routinely generated because of resource limitations (e.g., tissue availability, time, and cost).

Gene expression profiling studies have recently demonstrated that changes in DNA copy number can significantly influence gene expression values (12–17). If a genomic region is amplified, frequently a disproportionate number of genes that map to amplified region show increased expression compared to cytogenetically normal regions. Likewise, if a genomic region is lost, a disproportionate number of genes within the region show relatively decreased expression. Conceptually, therefore, it may be possible to infer cytogenetic information from gene expression profiles by identifying these regions of expression bias. At least two methods to computationally identify regional gene expression biases have been described, including comparative genomic microarray analysis (CGMA) and positional effect profiling (13, 15, 17, 18). Use of CGMA to infer cytogenetic profiles has been verified in a relatively large study (n = 98) of hepatocellular carcinoma (19). Notably, the cytogenetic profiles derived by CGMA were within the margin of error of cytogenetic profiles produced by comparative genomic hybridization.

In the present study, we examined the feasibility of using gene expression profiling data to build classification models for RCC in two ways: (a) we used the gene expression data directly to build an

4117
expression-based classification model; and (b) we identified regional expression biases to predict cytogenetic features. The in silico-derived cytogenetic profiles were used as surrogates for cytogenetic profiles derived from molecular based technologies. We demonstrate that we can construct a robust classification model for the three most common types of RCC that uses both gene-expression and inferred cytogenetic data. In addition, we demonstrate that identification of regional expression biases by the CGMA approach can provide accurate approximations of cytogenetic abnormalities that may assist in diagnosis and prognosis of RCC. The rapid generation of cytogenetic information from existing gene expression data will be useful in developing a blended classification model of RCC that incorporates pathological, cytogenetic, and gene expression-based components.

MATERIALS AND METHODS

Microarray Analysis

Microarray Data Sets. Normalized, log₂-transformed, two-color gene expression profiles for 81 RCC samples were generated as described previously (8). For each sample, a set of gene expression ratios (R) was generated from gene expression intensities derived from tumor tissue (T) and adjacent noncancerous tissue (N) such that R = log₂(T/N). Gene expression data for an additional 33 RCC and adjacent noncancerous tissue samples were obtained from the Stanford Microarray Database (20). In this data set, the tumor and adjacent non-tumor samples were compared with a pooled-cell-line reference. To allow direct comparison of tumor with normal expression values, new gene expression ratios (T/N) were estimated. Log₂-transformed noncancerous tissue ratios (N/U) were subtracted from the log₂-transformed tumor ratios (T/U) for each gene such that log₂(T/N) = log₂(T/U) − log₂(N/U) (15). If a RCC sample did not have a corresponding noncancerous sample, the global mean of the noncancerous tissue gene expression ratios was used.

To compare data from different sources, we used sequence comparisons to map all probe sequences to predicted Ensembl v10.2 genes (21). Expression values from multiple probes that mapped to the same gene were condensed by averaging. Ensembl gene identifiers were used as the common identifiers across the data sets. Overall, the data sets contained common expression values for 10,462 predicted Ensembl genes.

Comparative Genomic Microarray Analysis. Included in the Ensembl gene annotations are chromosomal mapping locations at base-pair resolution. To identify regional gene expression biases, we segregated gene expression values into sets based on chromosomal arm mapping. For each set, a binomial test was applied such that for n nonzero expression values, r gene expression values were scored as “up” if the log₂(T/N) value was positive and (n − r) values were scored as “down” if the log₂(T/N) was negative. In regions of significant bias (a ≤ 0.004), a summary statistic (z) for each region was computed by use of the normal approximation to the binomial distribution such that z = (2r − n)/√(n). Therefore, a positive z-statistic indicates a significant positive expression bias (i.e., genomic gain), and a negative z-statistics indicates the presence of a negative expression bias (i.e., genomic loss). The set of z-statistics was plotted as a heat map to identify and summarize the regional expression biases (22).

Classification of RCC Samples

Classification Based on Gene Expression Profiles. Sample classification was performed with the R environment according to the nearest-shrunken-centroid method as described previously (23, 24).7 The only modification to the referenced procedure was that missing values were imputed to the overall sample mean rather than the 4th nearest neighbor. Briefly, based on a training set of gene expression profiles, a subset of genes that could best discriminate between the sample grouping were identified. In our case, 10-fold cross-validation of the training set determined that a set of centroids consisting of 1018 genes (“shrunken” using a threshold = 3) gave the greatest classification accuracy for the training set (data not shown). Twenty-four of the 60 genes (40%) previously identified by Takahashi et al. (8) were also identified as discriminators by this approach. For classification, each test sample was compared with a consensus clear cell, papillary, or chromophobe expression profile (centroid) constructed from the training samples. A list of the classifying genes can be found in the Supplementary Data section.

Classification Based on Predicted Cytogenetic Profiles. An unweighted voting scheme was used to classify samples based on predicted cytogenetic abnormalities. Predicted deletions of chromosomes 3p, 8p, and 14q and amplification of chromosome 5q each provided a vote for the clear cell type (n = 4); chromosome losses for the p and q arms of chromosomes 1, 2, 6, 10, 13, and 17 each provided a vote for the chromophobe type (n = 12); chromosome gains for the p and q arms of chromosomes 7, 12, 16, and 17 each provided a vote for the papillary type (n = 8). A sample was classified based on which RCC subtype received the highest percentage of votes. Samples were not classified if they received no votes or received an equal percentage of votes for more than one type.

RESULTS

Classification of RCC Based on Gene Expression and Predicted Cytogenetic Profiles. Although gene expression-based classification schemes hold promise to assist in determining the diagnosis and prognosis of RCC, cytogenetic profiling has traditionally been used to assist in determining the RCC subtype. To determine whether it is possible to infer cytogenetic abnormalities by detecting regional expression biases, we obtained gene expression profiles for clear cell, papillary, and chromophobe subtypes generated in a previous study and organized them by hierarchical clustering (8). As described, the gene expression profiles formed distinct clusters that corresponded to tumor histological subtype (Fig. 1A). In parallel, cytogenetically abnormal regions were predicted by scanning the gene expression profiles for regional expression biases, using the CGMA approach (see “Materials and Methods”). In these samples, identification of regional expression biases corresponded well with traditional cytogenetic profiles of RCC (Fig. 1). Frequent upward gene expression biases were observed for chromosomes 7, 16, and 17 in the papillary samples, and frequent downward biases were frequently identified for chromosomes 1, 2, 6, 10q, and 17q in the chromophobe samples. For the clear cell samples, frequent downward expression biases were identified for chromosome 3p, and a frequent upward expression bias was identified for chromosome 5q. This approach did not predict a deletion of chromosome 13q, a frequent event in chromophobe RCC. Additionally, this approach predicted frequent amplifications of chromosomes 20q and 14q, events not commonly observed in previous cytogenetic studies of chromophobe RCC. Overall, however, the cytogenetic profiles predicted by CGMA agreed well with published cytogenetic abnormalities commonly attributed to RCC (6).

It is worth noting that the Pap G51, Pap G43, and Pap G56 samples, although classified as papillary in this study, were initially diagnosed as clear cell RCCs (data not shown). Two lines of evidence suggested that these samples should be classified as papillary: (a) the gene expression profiles from these samples clustered with other papillary-sample-derived gene expression profiles (Fig. 1); and (b) the pattern of regional expression biases (amplifications of chromosomes 7, 16, and 17) suggested that these samples have cytogenetic changes that are characteristics of papillary tumors. On the basis of this evidence, we sent, in a blinded manner, histological data from these samples to an additional pathologist specializing in urological pathology (Dr. Yang). This pathologist concluded these samples were indeed papillary RCCs (see Supplementary Data). These cases demonstrate the value of combining both gene expression and cytogenetic profiles to aid in RCC diagnosis.

Classification of RCC Based on Gene Expression and Predicted Cytogenetic Profiles. On the basis of these observations, we constructed a classifier that uses both gene expression profiles and predicted cytogenetic changes to assist in determining whether a partic-

---

ular RCC sample is clear cell, chromophobe, or papillary type. Rather than use hierarchical clustering to classify the RCC gene expression profiles (Fig. 1), we used a more robust class prediction method, termed prediction analysis of microarrays. Prediction analysis of microarrays is a variant of a nearest-centroid classifier that includes an automated gene selection step (23). To develop the classifier, gene expression profiles for 81 RCC generated by Takahashi et al. (8) were divided into two nonoverlapping sets, a training set that consisted of 41 samples and a test set that contained 40 samples. The training set contained 31 clear cell, 3 papillary, and 7 chromophobe gene expression profiles and was used to build the gene expression classifier. We then selected a subset of 1018 genes that could best discriminate between the sample grouping, based on the training profiles (see “Material and Methods”). The test set, which consisted of 29 clear cell, 2 papillary, and 9 chromophobe samples, was then applied to the trained RCC classifier. The expression-based RCC classifier correctly classified 100% of the test samples (Fig. 2).

Although the gene expression-based classifier was accurate in determining the RCC subtype, we also wanted to determine whether predicted cytogenetic profiles could complement the expression-based classifications. For each of the 40 samples in the test set, we generated a predicted cytogenetic profile by identifying regional expression biases (Fig. 1A). A voting-based classifier (see “Materials and Methods”) was constructed to classify samples based on their inferred cytogenetic profiles. For example, if a sample contained 75% of the cytogenetic abnormalities normally attributed to the clear cell type and only 10% and 15% of those abnormalities commonly associated with papillary and chromophobe, respectively, then the sample would be classified as clear cell. On the basis of the predicted cytogenetic profiles, the cytogenetic classifier correctly identified 19 of 29 (66%) clear cell samples, 2 of 2 (100%) papillary samples, and 7 of 9 (78%) chromophobe samples (Fig. 2A). Seven of the samples could not be classified because no significant regional expression biases could be
detected, and two samples received an equal percentage of votes for more than one class. Clear cell sample CC.J12 was classified as papillary because of predicted gains in chromosomes 7q, 12p, and 12q, features characteristic of papillary RCC. Reevaluation of the pathology confirmed the clear cell diagnosis; we therefore are presently investigating the significance of these unusual predicted gains. Noteworthy are the Pap.G51 and Pap.G43 samples; both were correctly identified as papillary by both the gene expression and cytogenetics-based classifiers, confirming their eventual diagnosis.

**Robust Classification of RCC Samples.** To determine whether the accuracy of this classification model was specific to the data produced by Takahashi et al. (8), we obtained gene expression profiles for an additional 33 RCC samples (26 clear cell, 4 papillary, and 3 chromophobe) described in Higgins et al. (10) and applied them to the expression-based classifier. The expression-based classifier correctly predicted 26 of 26 (100%) clear cell samples, 3 of 4 (75%) papillary samples, and 3 of 3 (100%) chromophobe samples (Fig. 2B). The papillary sample Pap.SKT030 was classified as clear cell. This discrepancy is likely due to the lack of representative papillary samples in the initial training set (n = 3). When we included all papillary samples from the dataset of Takahashi et al. (Ref. 8; n = 5), this papillary sample was correctly classified (data not shown). In addition, the cytogenetics-based classifier correctly predicted 32 of 33 samples (97%; Fig. 2B). It is worth noting that although the expression-based classifier initially misclassified the Pap.SKT030 sample, it was correctly classified by the pattern of regional expression biases.

Overall, the cytogenetics-based classifier supported the expression-based classifications in 60 of 73 (82%) cases. The classifiers disagreed in 3 of 73 cases (4%), and the cytogenetics classifier was uninformative in 11 of 73 cases (15%). Taken together, these data suggest that the combination of gene expression-based and cytogenetics-based classification models could allow robust classification the three major types of RCC from a single gene expression profile.

**Identification of Prognostic Factors from Predicted Cytogenetic Profiles.** We had previously observed that a significant molecular difference exists between clear cell RCCs that were isolated from patients with good overall 5-year survival versus patients who had poor overall 5-year survival (11). We therefore reexamined the gene expression profiles of these samples to determine whether an underlying cytogenetic event was associated with high stage/poor survival. As described previously, two subclusters emerged that largely discriminated between tumor stage and overall survival when the gene expression profiles were organized by hierarchical clustering of the (Fig. 1B). Cytogenetic abnormalities were also predicted by detecting regional expression biases. Again, CGMA identified features that are commonly found in clear cell RCC, including a frequent deletion of chromosome 3p and frequent amplification of chromosome 5q. A test of equal proportions was used to determine whether a significant cytogenetic abnormality existed that could distinguish between the clinically distinct groups. A predicted chromosomal loss of 14q was identified as an indicator of high stage/poor outcome (α = 0.04). This outcome agrees with previously published comparative genomic hybridization studies that identified chromosome 14q as a negative prognostic indicator. It has also been reported that a gain of chromosome 5q indicates a more favorable outcome (25). In our dataset a chromosome 5q gain was not significant indicator of outcome. This discrepancy may reflect our relatively small sample set and/or that previously published comparisons were based on longer overall survival rather than defined 5-year survival.

**DISCUSSION**

RCC is a heterogeneous group of cancers and histologically has been subdivided into clear cell, papillary (types 1 and 2), chromophobe, collecting duct, and unclassified subtypes. As such, the diagnosis of RCC can occasionally be challenging if, for example, the tumor contains mixed types of tissues or when the tumor does not resemble any of the known types (i.e., unclassified). Many genetic and chromosomal studies characterizing each subtype of tumor have been performed, making RCC one of the best genetically characterized tumors. As such, specific genetic and chromosomal changes have been correlated with each major RCC subtype. Recently, gene expression profiling, using cDNA, and other microarray technologies have been used to identify the molecular signatures of kidney tumors. Gene expression profiling relies on expression information from hundreds or even thousands of genes that may reflect the biology of the tumors. Molecular genetic and transcriptional information is often complementary; it would therefore be ideal to have data from both sources when examining associated clinicopathological features of RCC.

Although technically possible, from our experience extracting both RNA (for expression profiling) and DNA (for cytogenetic profiling) from the same piece of tissue at the same time (e.g., Trizol extraction) does not yield DNA of optimum quality to perform either conventional or array comparative genomic hybridization (26). To generate high-quality data from a sample, DNA and RNA are often obtained during separate extraction steps. From a practical point of view, this procedure is often laborious, costly, and consumes a larger amount of tissue. With the availability of advanced molecular tools such as genomics, transcriptomics, and proteomics, conservation of tissue is important. This is particularly true for small-sized tumors and/or biopsy samples, which often do not contain enough tissue to allow multiple extractions. In this study, we describe and validate a computational approach that can quickly, inexpensively, and accurately add cytogenetic profiling information to gene expression data. More importantly, we demonstrate that gene expression and predicted cytogenetic data are highly correlated and that their association with clinicopathological features therefore adds additional significance.

One potential application of combined gene expression data and cytogenetic information is to identify, based on their expression level, candidate tumor-suppressor genes or oncogenes from the chromosomal regions of interests. For example, from the inferred cytogenetic profiles of the papillary RCC, there are potential candidate oncogene(s) located in chromosomes 16 and 17 (Fig. 1). In clear cell RCC, chromosome 5q may harbor an important oncogene(s), whereas chromosome 14 may harbor poor-outcome-related genes. Further genetics studies, such as mutation analysis or fluorescent in situ hybridization studies, will be required to establish the identification of these RCC-related genes.

Although inconsistencies between separate gene expression profiling studies have been alluded to, we found extremely good agreement between the results generated by Takahashi et al. (8) and those generated by Higgins et al (10). Notably, the expression-based classifier was trained with data solely generated by Takahashi et al. (8), and data generated by Higgins et al. (10) were directly applied to it. The high percentage of accurately classified samples demonstrates that considerable agreement exists between these data sets. These results, in combination with initial studies suggesting that expression profiling can identify clinical differences between clear cell RCCs, make us optimistic that once appropriate clinical follow-up data can be collected, we will be able to construct more detailed classification models that will use a combination of expression profiles and inferred cytogenetics to predict other clinical parameters, such as patient outcome.
REFERENCES

Robust Classification of Renal Cell Carcinoma Based on Gene Expression Data and Predicted Cytogenetic Profiles

Kyle A. Furge, Kerry A. Lucas, Masayuki Takahashi, et al.

Cancer Res 2004;64:4117-4121.

Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/64/12/4117

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2004/06/25/64.12.4117.DC1

Cited articles
This article cites 21 articles, 12 of which you can access for free at:
http://cancerres.aacrjournals.org/content/64/12/4117.full.html#ref-list-1

Citing articles
This article has been cited by 9 HighWire-hosted articles. Access the articles at:
/content/64/12/4117.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.