

# Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma

Tao Shi<sup>1,2</sup>, David Seligson<sup>3</sup>, Arie S Belldegrun<sup>4</sup>, Aarno Palotie<sup>1,3,5,\*</sup> and Steve Horvath<sup>1,2,\*</sup>

<sup>1</sup>Department of Human Genetics, University of California, Los Angeles, CA, USA; <sup>2</sup>Department of Biostatistics, University of California, Los Angeles, CA, USA; <sup>3</sup>Department of Pathology & Laboratory Medicine, University of California, Los Angeles, CA, USA; <sup>4</sup>Department of Urology, University of California, Los Angeles, CA, USA and <sup>5</sup>The Finish Genome Center, University of Helsinki and the Laboratory Department of Helsinki University Central Hospital, Helsinki, Finland

**We describe a novel strategy (random forest clustering) for tumor profiling based on tissue microarray data. Random forest clustering is attractive for tissue microarray and other immunohistochemistry data since it handles highly skewed tumor marker expressions well and weighs the contribution of each marker according to its relatedness with other tumor markers. This is the first tumor class discovery analysis of renal cell carcinoma patients based on protein expression profiles. The tissue array data contained at least three tumor samples from each of 366 renal cell carcinoma patients. The eight tumor markers explore tumor proliferation, cell cycle abnormalities, cell mobility, and the hypoxia pathway. Since the procedure is unsupervised, no clinicopathological data or traditional classifications are used *a priori*. To explore whether the tissue microarray data can be used to identify fundamental subtypes of renal cell carcinoma patients, we first carried out random forest clustering of all 366 patients. By analyzing the tumor markers simultaneously, the procedure automatically detected classes that correspond to clear- vs non-clear cell tumors (demonstration of proof-of-principle). The resulting molecular grouping provides better prediction of survival (logrank  $P=0.000090$ ) than this classical pathological grouping (logrank  $P=0.023$ ). We then sought to extend the class discovery by searching for finer subclasses of *clear cell* patients. The procedure automatically discovered: (a) two classes corresponding to low- and high-grade patients (demonstration of proof-of-principle); (b) a subgroup of long-surviving clear cell patients with a distinct molecular profile and (c) two novel tumor subclasses in low-grade clear cell patients that could not be explained by any clinicopathological variables (demonstration of discovery).**

*Modern Pathology* (2005) 18, 547–557, advance online publication, 29 October 2004; doi:10.1038/modpathol.3800322

**Keywords:** tissue microarray; renal cell carcinoma; random forest clustering; tumor marker; tumor class discovery

The identification of cancer classes has traditionally been based on histomorphology. Recently, DNA microarrays have been used successfully to automatically discover cancer classes through clustering of the expression profiles.<sup>1</sup> It has been shown that many tumors can be clustered into clinically relevant groups based solely on gene expression (mRNA) profiles.

Tissue microarrays have become a widely used tool to screen for protein expression patterns in a

large numbers of tumors.<sup>2</sup> As the number of immunohistochemical marker measurements accumulates, it is natural to ask whether tissue microarray data (protein abundances) can also be used for tumor class discovery. Class discovery in this context entails two challenges: (a) developing algorithms to cluster tumors based on tissue microarray data and (b) determining whether putative classes (clusters) produced by such algorithms are biologically and clinically meaningful.

Most clustering algorithms require as input a dissimilarity measure between tumor samples. We find that dissimilarity measures that work well for DNA microarrays are not optimal for tissue microarrays. There is no reason why they should be: DNA microarray gene expression values are continuous and have a symmetric distribution, while tissue microarray tumor marker expressions are

Correspondence: Assistant Professor S Horvath, PhD, ScD, UCLA Department of Human Genetics, 695 Charles E Young Drive South, Los Angeles, CA 90095-7088, USA.  
E-mail: shorvath@mednet.ucla.edu

\*These authors codirected this work.

Received 26 May 2004; revised 13 September 2004; accepted 15 September 2004; published online 29 October 2004

semicontinuous and often highly skewed (supplement, Fig. Supp1). In this paper, we pioneer the use of the random forest dissimilarity measure<sup>3,4</sup> for the cluster analysis of a renal cell carcinoma tissue microarray data. In the supplement, we show empirically that the random forest dissimilarity is superior to standard dissimilarities used for DNA microarray data.

Renal cell carcinoma, the most common type of kidney cancer in adults, is the 14th leading cause of cancer mortality in the United States. There are five main types of renal cell carcinoma with clear cell being the most common form (70–80%).<sup>5</sup>

There are a number of reports on protein level tumor markers in renal cell carcinoma using tissue microarrays. However, all of these studies analyze less than four markers.<sup>6–11</sup> In this study, we examined a total of eight tumor markers which were reported previously to be involved in the natural history and progression of renal cell carcinoma. To the best of our knowledge, this is the first cluster analysis of renal cell carcinoma patients based on tissue microarray data. The eight markers explore different molecular aspects: tumor proliferation, cell cycle abnormalities, cell mobility, and the hypoxia pathway. Both of the nuclear antigens, Ki67, and p53, a tumor suppressor, are related to cellular proliferation. In renal cell carcinoma, both of them have been shown to be independent predictors of survival.<sup>12</sup> Gelsolin, EpCAM and vimentin may be involved in cell motility and cancer progression. Gelsolin, a member of the actin-binding protein family, has been described as a highly significant indicator of poor prognosis in non-small-cell lung cancer.<sup>13</sup> EpCAM (epithelial cell adhesion molecule) is widely expressed on the surface of many carcinomas.<sup>14,15</sup> Vimentin, an intermediate filament, has previously been identified as an independent predictor of poor prognosis in renal cell carcinoma.<sup>16,17</sup> CA9 and CA12 are members of the carbonic anhydrase family and are critical components of the hypoxia pathway. Decreased expression of CA9 has been shown to predict worse survival.<sup>18</sup> PTEN (phosphatase and tensin homologue deleted from chromosome 10) is a tumor suppressor gene that regulates cellular migration, proliferation and apoptosis.<sup>19</sup> Although PTEN mutation may be a rare event in renal cell carcinoma,<sup>20,21</sup> PTEN deletion has been shown to correlate with poor prognosis.<sup>21</sup>

Our hypothesis was that by analyzing these markers simultaneously, one might be able to (re-)discover biologically and clinically meaningful groups of patients. It is worth emphasizing that random forest clustering is an unsupervised learning method, which aims to find molecular classifications with distinct global expression profiles blinded to clinicopathologic covariates. If the primary goal is to use tumor markers for prediction purposes, a supervised learning approach should be used.

## Materials and methods

### Patients

The tissue samples were collected from a cohort of 366 patients who underwent a radical or partial nephrectomy for renal cell carcinoma at UCLA between 1989 and 2000. The mean age of the patients is 60 years and the male to female ratio is approximately 2:1. Following study protocol (KCP 99–233) approval by the UCLA Institutional Review Board, immunohistochemical studies were performed and clinical data from an established kidney cancer database were reviewed. The tumor samples were histologically subtyped according to the recommendations of the International Union Against Cancer and patients were staged according to the TNM classification.<sup>22</sup> Tumor grade was categorized using Fuhrman grade.<sup>23</sup> Performance status was determined using the Eastern Cooperative Oncology Group Performance Score (ECOG-PS) scale.<sup>24</sup> The primary outcome of interest was disease-specific survival. All the pathology covariates are summarized in Table 1.

### Tissue Array Construction and Immunohistochemistry

A tissue microarray of these 366 renal cell carcinoma patients was constructed and immunohistochemical staining was performed as previously described.<sup>25</sup> Immunostaining was scored by recording the total percentage of tumor cells staining. As discussed below, the same staining score was used for each tumor marker to ensure unbiased results. The arrays contained at least three cores of tumor sample per patient and we arrived at a summary score per patient by forming the mean value. As shown in the frequency plots in Fig. Supp1, the percentage of cells staining of the eight tumor markers are highly skewed, semicontinuous and non-normal.

### Statistical Methods

Our analyses of the data involve the following three general steps: (1) using random forest clustering to group the patients based only on their tumor marker expression profiles; (2) assess the differences between the resultant clusters in terms of their survival distributions and other clinicopathological variables, such as stage, grade etc.; (3) examine the difference in tumor marker expression between the clusters. The statistical methods used in the analyses are described below.

#### Random forest clustering

One major input of a clustering analysis is the dissimilarity measure.<sup>26</sup> We propose to use a random

**Table 1** Patient distribution and summary of survival information for each cluster. *P*-values next to the cross tabulations are Kruskal–Wallis *P*-values, while for survival difference are log-rank *P*-values. ‘NA’ means that the *P*-values cannot be calculated. Integers denote number of patients and percentages are row percentages

| Total no.       | All patients<br>366 |           |          |  | Clear cell patients<br>307 |           |           |          | Clear cell grade 2 patients<br>144 |           |           |          | Clear cell grade 3 patients<br>109 |          |          |          |                    |
|-----------------|---------------------|-----------|----------|--|----------------------------|-----------|-----------|----------|------------------------------------|-----------|-----------|----------|------------------------------------|----------|----------|----------|--------------------|
| Cluster         | 1                   |           | 2        |  | 1                          |           | 2         |          | 1                                  |           | 2         |          | 1                                  |          | 2        |          |                    |
| No. of patients | 327 (89%)           |           | 39 (11%) |  | 248 (81%)                  |           | 59 (19%)  |          | 106 (74%)                          |           | 38 (26%)  |          | 45 (41%)                           |          | 64 (59%) |          |                    |
| TNM stage       |                     |           |          |  |                            |           |           |          |                                    |           |           |          |                                    |          |          |          |                    |
| I               | 111                 | 88 (79%)  | 23 (21%) |  | 79                         | 72 (91%)  | 7 (9%)    |          | 41                                 | 30 (73%)  | 11 (27%)  |          | 11                                 | 7 (64%)  | 4 (36%)  |          |                    |
| II              | 25                  | 19 (76%)  | 6 (24%)  |  | 19                         | 15 (79%)  | 4 (21%)   |          | 12                                 | 9 (75%)   | 3 (25%)   |          | 5                                  | 2 (40%)  | 3 (60%)  |          |                    |
| III             | 59                  | 55 (93%)  | 4 (7%)   |  | 53                         | 46 (87%)  | 7 (13%)   |          | 27                                 | 21 (78%)  | 6 (22%)   |          | 17                                 | 10 (59%) | 7 (41%)  |          |                    |
| IV              | 167                 | 162 (97%) | 5 (3%)   |  | <i>P</i> = 8.05e−07        | 153       | 113 (74%) | 40 (26%) | <i>P</i> = 0.00129                 | 63        | 45 (71%)  | 18 (29%) | <i>P</i> = 0.787                   | 75       | 26 (35%) | 49 (65%) | <i>P</i> = 0.028   |
| Grade           |                     |           |          |  |                            |           |           |          |                                    |           |           |          |                                    |          |          |          |                    |
| 1               | 47                  | 37 (79%)  | 10 (21%) |  | 35                         | 34 (97%)  | 1 (3%)    |          | 0                                  | 0 (NA)    | 0 (NA)    |          | 0                                  | 0 (NA)   | 0 (NA)   |          |                    |
| 2               | 177                 | 155 (88%) | 22 (12%) |  | 144                        | 125 (87%) | 19 (13%)  |          | 144                                | 106 (74%) | 38 (26%)  |          | 0                                  | 0 (NA)   | 0 (NA)   |          |                    |
| 3               | 122                 | 116 (95%) | 6 (5%)   |  | 109                        | 79 (73%)  | 30 (28%)  |          | 0                                  | 0 (NA)    | 0 (NA)    |          | 109                                | 45 (41%) | 64 (59%) |          |                    |
| 4               | 13                  | 13 (100%) | 0 (0%)   |  | <i>P</i> = 0.000624        | 13        | 4 (31%)   | 9 (69%)  | <i>P</i> = 2.74e−07                | 0         | 0 (NA)    | 0 (NA)   | <i>P</i> = NA                      | 0        | 0 (NA)   | 0 (NA)   | <i>P</i> = NA      |
| Metastatic      |                     |           |          |  |                            |           |           |          |                                    |           |           |          |                                    |          |          |          |                    |
| No              | 195                 | 161 (83%) | 34 (17%) |  | 151                        | 133 (88%) | 18 (12%)  |          | 82                                 | 61 (74%)  | 21 (26%)  |          | 33                                 | 19 (58%) | 14 (42%) |          |                    |
| Yes             | 163                 | 159 (98%) | 4 (3%)   |  | <i>P</i> = 4.73e−06        | 149       | 110 (74%) | 39 (26%) | <i>P</i> = 0.00168                 | 61        | 44 (72%)  | 17 (28%) | <i>P</i> = 0.763                   | 71       | 22 (31%) | 49 (69%) | <i>P</i> = 0.0102  |
| ECOG            |                     |           |          |  |                            |           |           |          |                                    |           |           |          |                                    |          |          |          |                    |
| 0               | 141                 | 115 (82%) | 26 (18%) |  | 105                        | 96 (91%)  | 9 (9%)    |          | 56                                 | 43 (77%)  | 13 (23%)  |          | 28                                 | 18 (64%) | 10 (36%) |          |                    |
| 1               | 205                 | 194 (95%) | 11 (5%)  |  | 185                        | 139 (75%) | 46 (25%)  |          | 82                                 | 59 (72%)  | 23 (28%)  |          | 75                                 | 26 (35%) | 49 (65%) |          |                    |
| ≥2              | 16                  | 14 (88%)  | 2 (12%)  |  | <i>P</i> = 0.000622        | 13        | 9 (69%)   | 4 (31%)  | <i>P</i> = 0.000478                | 5         | 3 (60%)   | 2 (40%)  | <i>P</i> = 0.405                   | 6        | 1 (17%)  | 5 (83%)  | <i>P</i> = 0.00296 |
| Clear cell      |                     |           |          |  |                            |           |           |          |                                    |           |           |          |                                    |          |          |          |                    |
| No              | 50                  | 20 (40%)  | 30 (60%) |  | 0                          | 0 (NA)    | 0 (NA)    |          | 0                                  | 0 (NA)    | 0 (NA)    |          | 0                                  | 0 (NA)   | 0 (NA)   |          |                    |
| Yes             | 316                 | 307 (97%) | 9 (3%)   |  | <i>P</i> = 5.5e−34         | 307       | 248 (81%) | 59 (19%) | <i>P</i> = NA                      | 144       | 106 (74%) | 38 (26%) | <i>P</i> = NA                      | 109      | 45 (41%) | 64 (59%) | <i>P</i> = NA      |
| Survival        |                     |           |          |  |                            |           |           |          |                                    |           |           |          |                                    |          |          |          |                    |
| No. of patients | 321                 |           | 39       |  | 244                        |           | 58        |          | 105                                |           | 37        |          | 45                                 |          | 64       |          |                    |
| No. of death    | 155                 |           | 6        |  | 102                        |           | 44        |          | 38                                 |           | 22        |          | 20                                 |          | 50       |          |                    |
| Median survival | 4.0                 |           | >12      |  | 5.6                        |           | 1.2       |          | >12                                |           | 2.7       |          | 5.1                                |          | 1.4      |          |                    |
| 95% CI (lower)  | 2.9                 |           | >12      |  | 4.1                        |           | 0.7       |          | 4.0                                |           | 1.8       |          | 2.7                                |          | 0.7      |          |                    |
| 95% CI (upper)  | 5.3                 |           | >12      |  | <i>P</i> = 9.03e−05        | >12       |           | 2.3      | <i>P</i> = 4.82e−09                | >12       |           | 5.6      | <i>P</i> = 0.0353                  | >12      |          | 2.3      | <i>P</i> = 0.0022  |

forest dissimilarity for tissue microarray data since it has the following theoretical advantages.<sup>4</sup> First, the clustering results do not change when one or more covariates are monotonically transformed since the dissimilarity only depends on the feature ranks. Thus, one does not need to worry about symmetrizing skewed covariate distributions. Second, the random forest dissimilarity weighs the contributions of each covariate on the dissimilarity in a natural way: the more related the covariate is to other covariates, for example the more correlated a protein marker is with other markers, the more it will affect the definition of the random forest dissimilarity. Third, the random forest dissimilarity does not require the user to specify threshold values for dichotomizing tumor expressions. Since the random forest dissimilarity is based on individual tree predictors, which dichotomize the expression values as part of their construction, the random forest dissimilarity automatically dichotomizes the expressions in a principled, data-driven way. Fourth, the random forest dissimilarity naturally accommodates missing values. For a technical description of the random forest dissimilarity consult the supplement, Breiman,<sup>3</sup> Shi and Horvath<sup>4</sup> and a technical report that can be downloaded from <http://www.genetics.ucla.edu/labs/horvath/publications/RFclusteringShiHorvath.pdf>.

The random forest clustering procedure is carried out as follows. The random forest dissimilarity is used to represent each patient as a point in a two-dimensional space with the aid of multidimensional scaling. The distances between the points are used in partitioning around medoids clustering.<sup>26</sup> The number of clusters is chosen by using the partitioning around medoids silhouette plots and inspecting corresponding multidimensional scaling plots.

Computer code and a tutorial that implements random forest clustering in R language (<http://www.r-project.org/>)<sup>27</sup> can be obtained from the following web page: <http://www.genetics.ucla.edu/labs/horvath/kidneypaper/RCC.htm>.

#### *Other statistical methods*

We used several methods for describing the clusters in terms of clinical variables and tumor marker expressions. To test whether variables differed across groups, we used the Kruskal–Wallis test, which is a nonparametric multi-group comparison test. To visualize the survival distributions, we used Kaplan–Meier plots. Log-rank tests were used to test the difference between survival distributions. All *P*-values were two-sided and *P* < 0.05 was considered significant. All statistical analyses were carried out with the freely available software R (<http://www.r-project.org/>).<sup>27</sup>

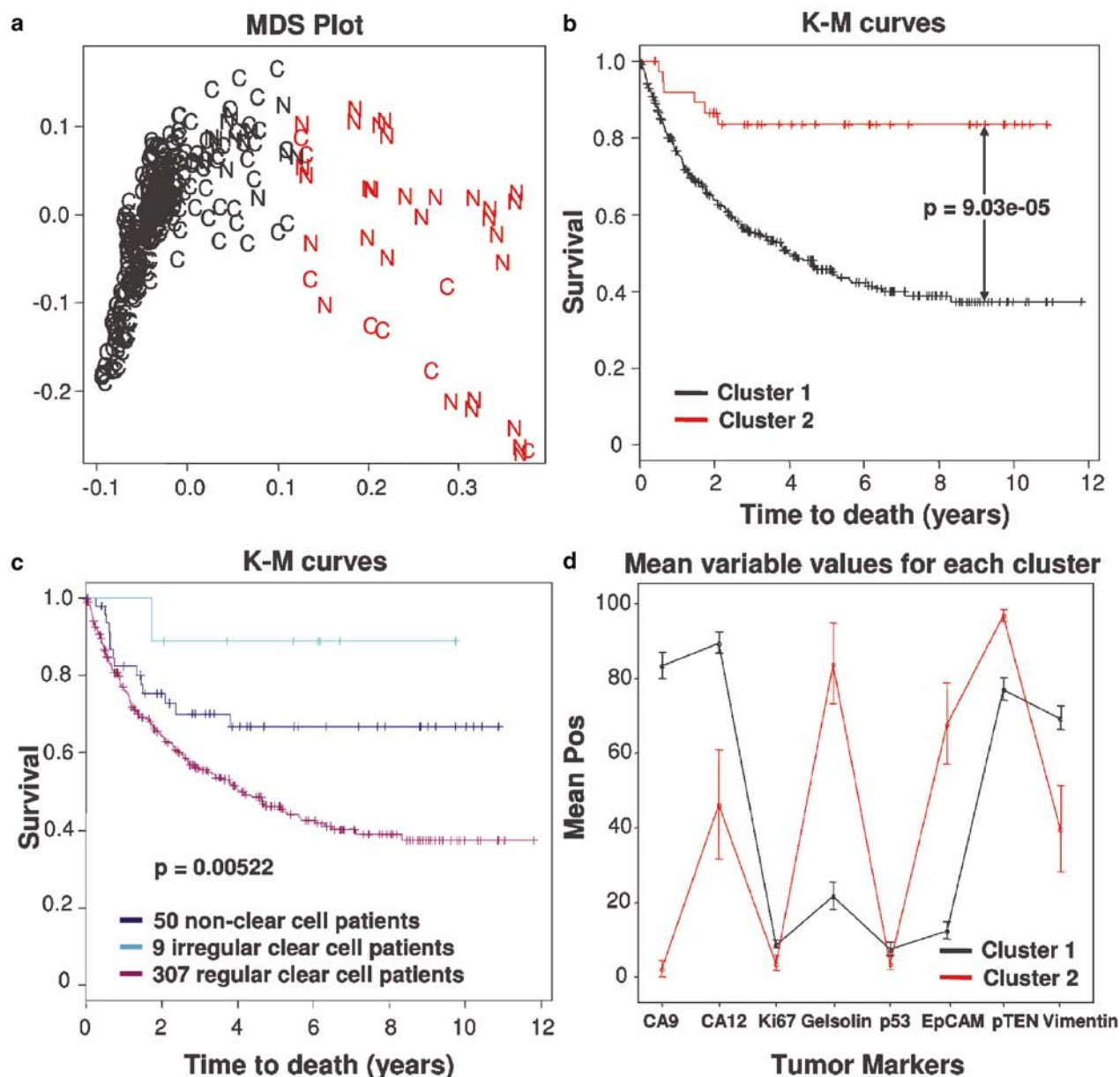
## Results

### Clustering All Renal Cell Carcinoma Patients

To explore whether the tissue microarray data can be used to identify fundamental subtypes of renal cell carcinoma patients, we first carried out random forest clustering of all patients using the staining scores (percent of cells staining) of the eight tumor markers. The patients are depicted as points in two-dimensional multidimensional scaling plots. The distances between the data points reflect the random forest dissimilarities between them. Partitioning around medoids clustering grouped the points (patients) into two clusters with 327 and 39 patients each (Figure 1a).

We related the resultant clusters to commonly used clinicopathological covariates: TNM stage, grade, metastatic status, ECOG (health performance status), renal cell carcinoma subtypes and survival. We find that 97% of the clear cell patients, a renal cell carcinoma subtype known to have a relatively poor prognosis,<sup>28</sup> are in cluster 1, while 60% of the non-clear cell patients are in cluster 2 (Table 1). This difference is highly significant ( $P = 5.5 \times 10^{-34}$ ). This suggests that the clear cell/non-clear cell distinction could have been automatically discovered on the basis of the tumor marker data without previous biological knowledge. We also find that other clinicopathological covariates differ across clusters: TNM stage ( $P = 8.05 \times 10^{-7}$ ), metastasis status ( $P = 4.73 \times 10^{-6}$ ), ECOG ( $P = 0.000622$ ) and grade ( $P = 0.000624$ ) (Table 1). The survival distributions of the patients corresponding to the two clusters are significantly different ( $P = 9.03 \times 10^{-5}$ , Figure 1b). The patients in clusters 1 and 2 have median survival times of 4 and more than 12 years, respectively. The fact that the patients can be grouped into clinically meaningful clusters based only on their tumor marker expression profiles provides indirect empirical evidence that random forest clustering might be a valuable tool for tissue microarray data analysis. In the supplement, we compare random clustering to other widely used clustering methods.

Because we found our strongest cluster association with renal cell carcinoma histology class, it is natural to ask whether the molecular grouping provides better prediction of survival than this classical pathological grouping. When comparing the survival profile of cluster 1 patients to that of cluster 2 patients, we find a highly significant difference ( $P = 9.03 \times 10^{-5}$ ), while we find a less significant difference between clear cell and non-clear cell patients ( $P = 0.0229$ ) (Fig. Supp2). This suggests that, while the molecular grouping tends to delineate clear cell from non-clear cell patients, it provides additional predictive power through associations with other clinicopathological variables and potentially through molecular pathways with no clear association with the variables in our study.



**Figure 1** (a) The 366 renal cell carcinoma patients are visualized using a multidimensional scaling plot based on the random forest dissimilarity. Patients are colored by their cluster membership (black for cluster 1 and red for cluster 2) and labeled by tumor subtypes ('C' for clear cell and 'N' for non-clear cell patients). (b) Kaplan–Meier plots show that patients in the two clusters have very different survival distributions. The curves are colored in the same way as in (a). (c) Kaplan–Meier plots for non-clear cell patients (blue), regular clear cell patients (pink) and irregular clear cell patients (cyan). (d) For each tumor marker, we report the mean expression value in each cluster. The error bars show 95% confidence intervals. The lines are colored in the same way as in (a). For box-plots and *P*-values refer to Fig. Supp2 in the supplement.

The new molecular grouping of the patients can also be used to find certain patient samples, called here 'irregular', that display unexpected molecular profiles. We refer to the clear cell status determined histologically by a pathologist as morphological clear cell status. The clear cell patients in clusters 1 are referred to as 'regular' because the cluster tends to be enriched for these patients and those in cluster 2 as 'irregular' clear cell patients because that cluster

is enriched with non-clear cell patients. In Figure 1c, we plot the Kaplan–Meier estimates of the survivorship functions of the 307 regular, the nine irregular clear cell-, and the 50 non-clear cell patients. The irregular clear cell patients have a distinct survival advantage over regular clear cell patients ( $P=0.025$ ), though the significance is less compared to the survival of non-clear cell patients ( $P=0.22$ ) (Figure 1c), which may be due to the low

sample size. After revisiting the pathology reports, we found that the nine irregular samples came mainly from low-grade (grade <3, nine out of nine), low-stage (stage <3, six out of nine) and nonmetastatic (eight out of nine) patients. When we compare them to the remaining 77 low risk (low-grade, low-stage, and nonmetastatic) clear cell patients, we find that both groups have similar survival distributions (Fig. Supp3). However, the tumor marker expression profiles of the two low-risk groups differ: the nine renal cell carcinoma irregular patients have very low CA9 and Vimentin expression but high Gelsolin expression (Fig. Supp4). When visualizing the nine plus 77 low-risk patients in a multidimensional scaling plot, we find that the nine irregular patients all cluster together (Fig. Supp5). This shows that patient groups with distinctly different molecular profiles may, however, share similar clinicopathologic groupings and outcomes. The utility of molecular classifications in these instances is currently unclear, but speaks to truly different patient populations that otherwise would not be identified.

Next, we examined the tumor marker expression across the two clusters in Figure 1a. In Figure 1d, we plot the mean expression value of each tumor marker for the different clusters. We find that CA9 and CA12 have significantly higher expression in cluster 1 patients than in cluster 2 patients, while Gelsolin and EpCAM have significantly lower expression. We find that CA9, Gelsolin, EpCAM and CA12, are most important for distinguishing the two clusters of patients (corresponding box-plots and Kruskal–Wallis *P*-values can be found in the supplement, Fig. Supp6).

### Clustering Regular Clear Cell Patients

We then sought to extend the class discovery by searching for finer subclasses of the 307 *regular* clear cell patients identified in the previous section. Using random forest clustering, we grouped the 307 patients into two clusters with 248 patients in cluster 1 and 59 patients in cluster 2 (Table 1 and Figure 2a).

When testing whether clinical covariates differed between the two clusters, we find that grade ( $P=2.74e-7$ ), ECOG ( $P=0.000478$ ), TNM stage ( $P=0.00129$ ), and metastatic status ( $P=0.00168$ ) are all significantly different with grade being the most significant (Table 1). We find that 64% of cluster 1 patients but only 34% of cluster 2 patients have a low grade. This suggests that the class discovery approach automatically discovered the distinction between high- and low-grade patients independent of prior biological knowledge. When comparing the survival of cluster 1 patients to that of cluster 2 patients, we find a highly significant difference ( $P=4.82e-9$ ), with cluster 1 patients showing a survival advantage. The median survival times of cluster 1 and 2 patients are 5.6 and

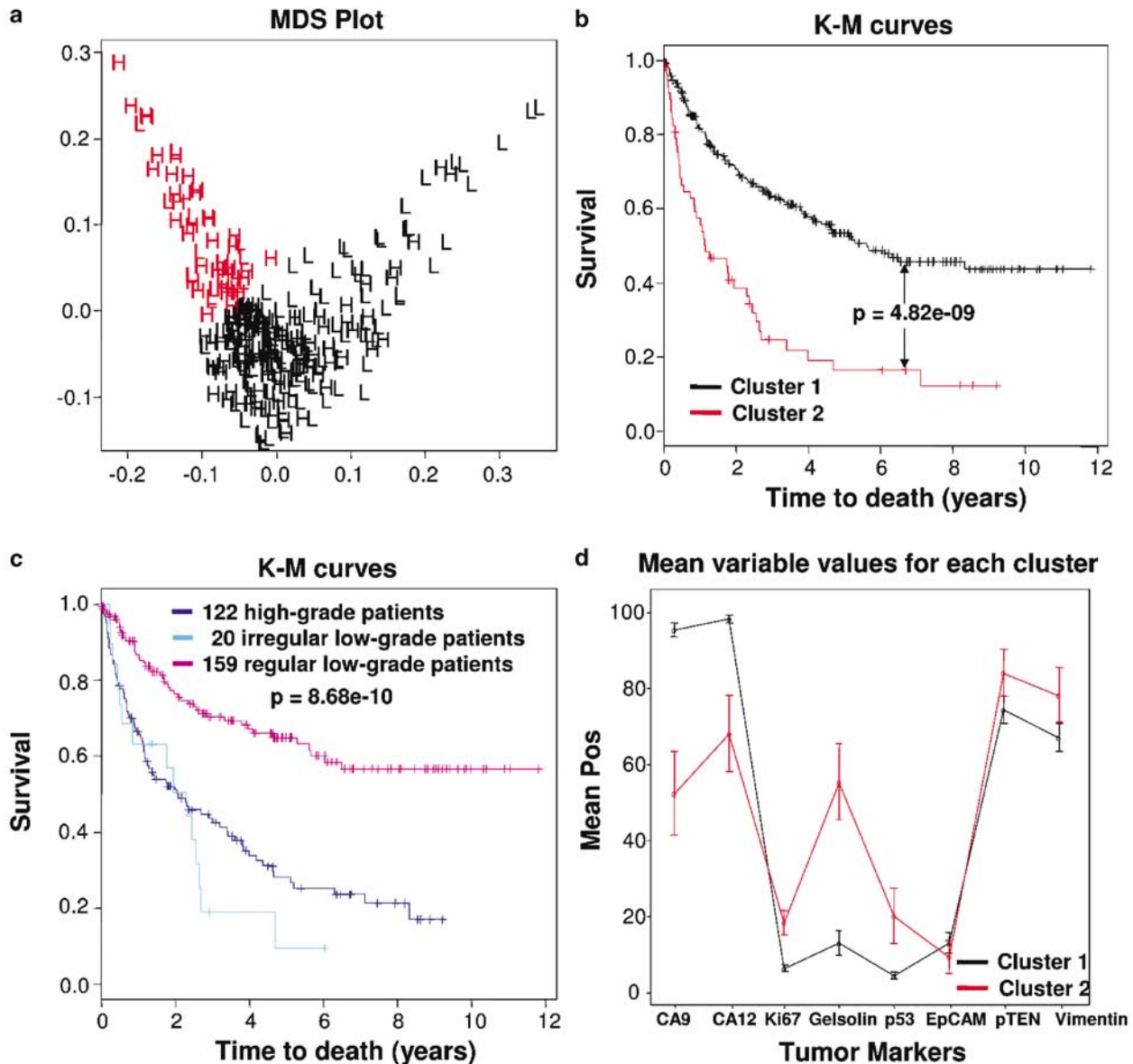
1.2 years, respectively. Since the resultant clusters were most highly associated with tumor grade, we compared that variable to our cluster results. We find a slightly less significant survival difference between low- and high-grade patients ( $P=2.6e-7$ ; the median survival times of the low-grade and high-grade patients are >12 and 2 years, respectively). The corresponding Kaplan–Meier plots can be found in the supplement, Fig. Supp7. As in our first analysis above, we isolated individual tumors that were placed in an unexpected cluster due to their variant molecular profile, calling them again ‘irregular’. Therefore, we refer to low-grade patients in clusters 1 and 2 as regular and irregular low-grade patients, respectively. In Figure 2c, we plot the Kaplan–Meier estimates of the survivorship functions of the 159 regular, the 20 irregular low-grade, and the 122 high-grade patients. The 20 irregular low-grade clear cell patients have a significantly worse survival profile than the 159 regular low-grade clear cell patients ( $P=2.85e-6$ ; the median survival times of the irregular and regular low-grade patients are 2.3 and >12 years, respectively; Figure 2c). We find that 70% of the 20 irregular low-grade clear cell patients are high-stage (group stage >2) patients, while only 55% of the 159 regular low-grade clear cell patients are high-stage patients. This significant difference in stage ( $P=0.045$ ) may explain the difference in survival between regular and irregular low-grade clear cell patients, especially since none of the other clinicopathological covariates are significant. When comparing the 20 irregular low-grade patients to the 87 similar high-stage low-grade patients, we find that their survival profiles are still significantly different ( $P=0.018$ , Fig. Supp8). Therefore, the molecular profile distinguished a low-grade group with poor survival, whose survival is partially explained by enrichment of high-stage cases but other undiscovered mechanisms may be at work.

When examining the tumor marker expressions, we find that all tumor markers except EpCAM differ significantly across the two clusters (corresponding boxplots and *P*-values can be found in the supplement, Fig. Supp9). In particular, CA9 and CA12 have lower, and Gelsolin higher, expression in cluster 2 than in cluster 1 (Figure 2d).

We also clustered the 50 non-clear cell patients, but we did not find meaningful clusters, which may be due to the small sample size.

### Analysis of the Regular Clear Cell Patients with a Fixed Grade

After observing that random forest clustering was able to detect clinically meaningful clusters, we aimed to detect clusters that could not be explained in terms of tumor morphology-based covariates, such as tumor type and grade. Therefore, we



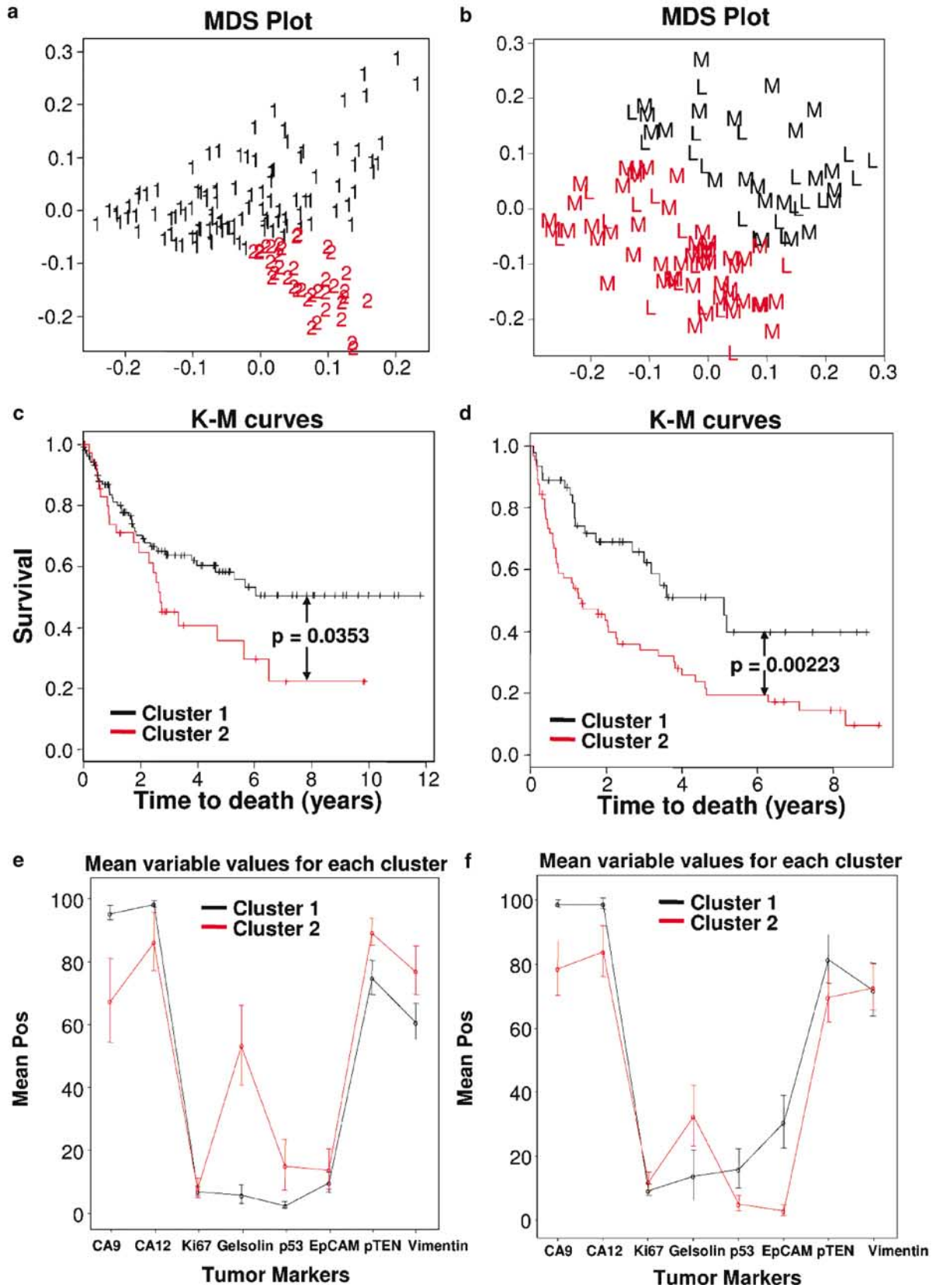
**Figure 2** (a) The 307 regular clear cell renal cell carcinoma patients are visualized using a multidimensional scaling plot based on the random forest dissimilarity. Patients are colored by their cluster membership (black for cluster 1 and red for cluster 2) and labeled by their histological grade ('L' for low- and 'H' for high-grade patients). (b) Kaplan–Meier plots show that patients in the two clusters have very different survival distributions. The curves are colored as in (a). (c) Kaplan–Meier plots for the high grade clear cell patients (blue), regular low-grade clear cell patients (pink) and irregular low grade clear cell patients (cyan). (d) For each tumor marker, we report the mean expression value in each cluster. The error bars show 95% confidence intervals. The lines are colored in the same way as in (a). Box plots and *P*-values refer can be found in Fig. Supp4 of the supplement.

analyzed clear cell patients with a fixed grade. The clear cell patients were comprised of 35 grade-1, 144 grade-2, 109 grade-3, and 13 grade-4 patients.

Random forest clustering groups the 144 grade 2 patients into two clusters (Figure 3a). We find that patients in the two clusters have significantly different survival profiles ( $P=0.035$ ; median survival times for the cluster 1 and 2 patients are >12 and 2.7 years, respectively; Figure 3c). Interestingly, none of the clinicopathological covariates differs significantly across the two clusters even though there are relatively many patients in each cluster

(106 and 38 in clusters 1 and 2, respectively; Table 1). The existence of the two distinct tumor marker expression patterns for grade-2 patients points to tumor marker expression heterogeneity in these patients. When examining the tumor marker expression across the two clusters, we find that CA9 has a significantly higher and Gelsolin a significantly lower expression in cluster 1 patients than in cluster 2 patients (Figure 3e and Fig. Supp10). This suggests that, even within tightly confined morphological classifications, random forest clustering can be used to uncover novel tumor subtypes based on







expression profiles, but this result should be replicated in independent data sets.

Random forest clustering groups the 109 grade-3 patients into two clusters (Figure 3b). We find that cluster 2 is significantly enriched with high stage ( $P=0.028$ ), high ECOG ( $P=0.0030$ ), and metastatic patients ( $P=0.010$ ) (Table 1). As can be expected, cluster 2 patients have lower median survival (1.4 years) than cluster 1 patients (5.1 years) (Figure 3d). The survival difference based on the molecular grouping ( $P=0.0022$ ) was comparable to that seen in the pathology grouping based on ECOG ( $P=0.0021$ , 0 vs >0 ECOG), but was less significant than the pathology grouping based on metastatic status ( $P=3.34e-6$ ) and stage ( $P=0.00077$ , high vs low stage). When examining the expression profiles of the eight tumor markers across the two clusters (Figure 3f), we find that the two clusters are most distinguished by the expression profiles of CA9 and EpCAM. Both markers are highly expressed in cluster 1 ( $P$ -values and box plots can be found in the supplement, Fig. Supp11).

We also clustered the 35 grade-1 and the 13 grade-4 patients (supplement, Figs. Supp12–14) but did not identify meaningful clusters, which may be due to small sample sizes.

## Discussion

We show that tissue microarray data-based class discovery techniques can be used to identify fundamental subtypes of cancer. To the best of our knowledge, this is the first unsupervised analysis of renal cell carcinoma tumors based on protein expression data. A comparison of unsupervised and supervised results can be found in the supplement.

Tissue microarrays are a tumor marker validation technique that aims to validate relatively few tumor markers on many tumor samples. In contrast, DNA microarrays and proteomics assays probe many genes on relatively few samples. Thus, these techniques are complementary and address different research aims. The main road for identifying tumor classes will be to probe many (thousands of) genes using DNA microarrays and proteomics assays since more genes means more information. But this paper provides evidence that a less traveled, a less obvious road, can also lead to the discovery of tumor classes. We show that tumor marker validation data can be used to find tumor classes, especially if powerful data mining methods are used. In the supplement,

we provide some empirical evidence that random forest clustering outperforms other standard clustering approaches used for DNA microarrays.

While unsupervised analyses have not been used to analyze protein expression data in renal cell carcinoma, several unsupervised analyses of renal cell carcinoma samples based on mRNA expression data have been reported in the literature.<sup>29–32</sup> It is interesting that the eight tumor markers in our study yield results that are consistent with those found by using thousands of mRNA level gene expression values. In particular, using different clustering methods, all of the DNA microarray studies observe distinct global gene expression signatures associated with clear cell- and non-clear cell renal cell carcinomas. In addition, our results coincide with the findings of Takahashi *et al*<sup>29,30</sup> that (a) there are two subgroups of clear cell renal cell carcinoma with significantly different survival outcomes, and (b) that the low-risk (better surviving) group contains more low-grade patients than the high-risk group.

In this study, we measured the tumor marker expressions by the percent of positively staining cells. This staining score is a continuous, undichotomized variable, ranging from 0 to 100%. It is standard practice in *supervised* analyses to dichotomize tumor marker expressions for ease of interpretation and reproducibility. But, we caution against using external threshold values for dichotomizing expressions in *unsupervised* analyses since continuous variables may contain additional predictive information when compared to dichotomized variables. In addition, using undichotomized staining scores may be particularly relevant in the future when semiautomated or automated methods for assessing staining scores become available. To allow for comparisons across institutions, standardized tumor marker staining and scoring protocols should be established.

The fact that the random forest method was able to create clinically well defined, meaningful classes using the molecular signature of only eight protein-level markers provides indirect evidence that the method works well on real data; the main groupings generated were frequently associated with strongly predictive conventional variables, such as tumor subtype and grade. Using the method we were able to discover novel molecularly defined patient groups that might not have been isolated using traditional clinicopathological data. These novel subtypes of cancer will need to be

**Figure 3** Multi-dimensional scaling plots (top), Kaplan–Meier curves (middle) and parallel coordinate plots (bottom) of tumor marker expression in each cluster of the grade 2 (left, **a**, **c** and **e**) and grade 3 (right, **b**, **d** and **f**) patients. (**a**) Multi-dimensional scaling plot of the 144 grade 2 clear cell patients. Patients are labeled and colored by their cluster membership. (**b**) Multidimensional scaling plot of the 109 grade 3 clear cell patients. Patients are colored by their cluster membership (black for cluster 1 and red for cluster 2) and labeled by their metastatic status ('L' for localized and 'M' for metastatic patients). (**c**, **d**) Kaplan–Meier curves by cluster for grade 2 and grade 3 patients, respectively. The curves are colored in the same way as in (**a**) and (**b**). (**e**, **f**) For each tumor marker, we report the mean expression value in each cluster. The error bars show 95% confidence intervals. The lines are colored in the same way as before. Box plots and  $P$ -values can be found in Fig. Supp5 and Fig. Supp6 of the supplement.

validated across different institutions and technological platforms.

## Acknowledgements

This work is supported by the UCLA IGERT Bioinformatics Program funded by NSF DGE 9987641 (TS) and by the Jonsson Comprehensive Cancer Center (JCCC) NIH 2 P30 CA16042-29 (DS). We thank Mervi Eeva, Sheila Tze, and Hong Yu for their help in generating the data.

## References

- Golub TR, Slonim DK, Tamayo P, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531–537.
- Kononen J, Bubendorf L, Kallioniemi A, *et al*. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844–847.
- Breiman L. Random forests. *Machine Learning* 2001; 45:5–32.
- Shi T, Horvath S. Using random forest similarities in unsupervised learning: applications to microarray data In: Atlantic Symposium on Computational Biology and Genome Informatics (CBGI'03); 2003; Cary, NC, USA. The Association of Intelligent Machinery: Durham, NC, USA, 2003.
- Jemal A, Murray T, Samuels A, *et al*. Cancer statistics, 2003. *CA Cancer J Clin* 2003;53:5–26.
- Langner C, Ratschek M, Rehak P, *et al*. Steroid hormone receptor expression in renal cell carcinoma: an immunohistochemical analysis of 182 tumors. *J Urol* 2004;171(2 Part 1):611–614.
- Jacobsen J, Grankvist K, Rasmuson T, *et al*. Expression of vascular endothelial growth factor protein in human renal cell carcinoma. *BJU Int* 2004;93:297–302.
- Langner C, Ratschek M, Rehak P, *et al*. Expression of MUC1 (EMA) and E-cadherin in renal cell carcinoma: a systematic immunohistochemical analysis of 188 cases. *Mod Pathol* 2004;17:180–188.
- Hotakainen K, Ljungberg B, Haglund C, *et al*. Expression of the free beta-subunit of human chorionic gonadotropin in renal cell carcinoma: prognostic study on tissue and serum. *Int J Cancer* 2003;104: 631–635.
- Moch H, Schraml P, Bubendorf L, *et al*. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *Am J Pathol* 1999;154:981–986.
- Hedberg Y, Ljungberg B, Roos G, *et al*. Expression of cyclin D1, D3, E, and p27 in human renal cell carcinoma analysed by tissue microarray. *Br J Cancer* 2003;88:1417–1423.
- Rioux-Leclercq N, Turlin B, Bansard J, *et al*. Value of immunohistochemical Ki-67 and p53 determinations as predictive factors of outcome in renal cell carcinoma. *Urology* 2000;55:501–505.
- Shieh DB, Godleski J, Herndon II JE, *et al*. Cell motility as a prognostic factor in Stage I nonsmall cell lung carcinoma: the role of gelsolin expression. *Cancer* 1999;85:47–57.
- Shetye J, Christensson B, Rubio C, *et al*. The tumor-associated antigens BR55-2, GA73-3 and GICA 19-9 in normal and corresponding neoplastic human tissues, especially gastrointestinal tissues. *Anticancer Res* 1989;9:395–404.
- Riethmuller G, Schneider-Gadicke E, Schlimok G, *et al*. Randomised trial of monoclonal antibody for adjuvant therapy of resected Dukes' C colorectal carcinoma. German Cancer Aid 17-1A Study Group. *Lancet* 1994;343:1177–1183.
- Moch H, Schraml P, Bubendorf L, *et al*. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *Am J Pathol* 1999;154: 981–986.
- Sabo E, Miselevich I, Bejar J, *et al*. The role of vimentin expression in predicting the long-term outcome of patients with localized renal cell carcinoma. *Br J Urol* 1997;80:864–868.
- Bui MH, Seligson D, Han KR, *et al*. Carbonic anhydrase IX is an independent predictor of survival in advanced renal clear cell carcinoma: implications for prognosis and therapy. *Clin Cancer Res* 2003;9: 802–811.
- Steck PA, Pershouse MA, Jasser SA, *et al*. Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers. *Nat Genet* 1997;15:356–362.
- Alimov A, Li C, Gizatullin R, *et al*. Somatic mutation and homozygous deletion of PTEN/MMAC1 gene of 10q23 in renal cell carcinoma. *Anticancer Res* 1999;19:3841–3846.
- Velickovic M, Delahunt B, McIver B, *et al*. Intragenic PTEN/MMAC1 loss of heterozygosity in conventional (clear-cell) renal cell carcinoma is associated with poor patient prognosis. *Mod Pathol* 2002;15: 479–485.
- Sobin LH, Fleming ID. TNM Classification of Malignant Tumors, fifth edition (1997). Union Internationale Contre le Cancer and the American Joint Committee on Cancer. *Cancer* 1997;80:1803–1804.
- Fuhrman SA, Lasky LC, Limas C. Prognostic significance of morphologic parameters in renal cell carcinoma. *Am J Surg Pathol* 1982;6:655–663.
- Oken MM, Creech RH, Tormey DC, *et al*. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol* 1982;5:649–655.
- Kim HL, Seligson D, Liu XL, *et al*. Using protein expressions to predict survival in clear cell renal carcinoma. *Clin Cancer Res* 2004;10:5464–5471.
- Kaufman L, Rousseeuw PJ. Finding Groups in Data: an Introduction to Cluster Analysis. Wiley: New York, 1990.
- Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat* 1996;5:299–314.
- Cheville JC, Lohse CM, Zincke H, *et al*. Comparisons of outcome and prognostic features among histologic subtypes of renal cell carcinoma. *Am J Surg Pathol* 2003;27:612–624.
- Takahashi M, Yang XJ, Sugimura J, *et al*. Molecular subclassification of kidney tumors and the discovery of new diagnostic markers. *Oncogene* 2003;22: 6810–6818.
- Takahashi M, Rhodes DR, Furge KA, *et al*. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. *Proc Natl Acad Sci USA* 2001;98:9754–9759.

- 31 Boer JM, Huber WK, Sultmann H, *et al*. Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Res* 2001;11:1861–1870.
- 32 Young AN, Amin MB, Moreno CS, *et al*. Expression profiling of renal epithelial neoplasms: a method for tumor classification and discovery of diagnostic molecular markers. *Am J Pathol* 2001;158:1639–1651.

Supplementary Information accompanies the paper on Modern Pathology website (<http://www.nature.com/mpath>).