Article

Development of a Single Molecule Counting Assay to Differentiate Chromophobe Renal Cancer and Oncocytoma in Clinics

Khaled Bin Satter [1], Zach Ramsey [2], Paul M. H. Tran [1], Diane Hopkins [1], Gregory Bearden [1], Katherine P. Richardson [1], Martha K. Terris [3], Natasha M. Savage [2], Sravan K. Kavuri [2] and Sharad Purohit [1,4,5,*]

[1] Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Augusta University, 1120, 15th St., Augusta, GA 30912, USA; fbinsatter@augusta.edu (K.B.S.); ptran@augusta.edu (P.M.H.T.); dhopkins@augusta.edu (D.H.); gbearden@augusta.edu (G.B.); katrichardson@augusta.edu (K.P.R.)

[2] Department of Pathology, Medical College of Georgia, Augusta University, 1120 15th St., Augusta, GA 30912, USA; zramsey@augusta.edu (Z.R.); nsavage@augusta.edu (N.M.S.); skavuri@augusta.edu (S.K.K.)

[3] Department of Urology, Medical College of Georgia, Augusta University, 1120 15th St., Augusta, GA 30912, USA; mterris@augusta.edu

[4] Department of Obstetrics and Gynecology, Medical College of Georgia, Augusta University, 1120 15th St., Augusta, GA 30912, USA

[5] Department of Undergraduate Health Professionals, College of Allied Health Sciences, Augusta University, 1120 15th St., Augusta, GA 30912, USA

[*] Correspondence: spurohit@augusta.edu

Simple Summary: We previously reported a gene signature, "Chromophobe-Oncocytoma Gene Signature" (COGS), to differentiate chromophobe renal cell carcinoma from oncocytoma. Current clinical workflow with histology and immunohistochemistry can fail to distinguish these two renal cancers. We have evaluated the potential of COGS genes to classify these two renal tumors, on a single-molecule counting platform by measuring the expression of the COGS genes in archival tissue at Augusta University Medical Center. We show the expression level difference of the COGS signature in these tumors is able to classify these cancers accurately using machine learning. The assay has the potential to be instituted in clinically complex cases to differentiate these renal tumors.

Abstract: Malignant chromophobe renal cancer (chRCC) and benign oncocytoma (RO) are two renal tumor types difficult to differentiate using histology and immunohistochemistry-based methods because of their similarity in appearance. We previously developed a transcriptomics-based classification pipeline with "Chromophobe-Oncocytoma Gene Signature" (COGS) on a single-molecule counting platform. Renal cancer patients (*n* = 32, chRCC = 17, RO = 15) were recruited from Augusta University Medical Center (AUMC). Formalin-fixed paraffin-embedded (FFPE) blocks from their excised tumors were collected. We created a custom single-molecule counting code set for COGS to assay RNA from FFPE blocks. Utilizing hematoxylin-eosin stain, pathologists were able to correctly classify these tumor types (91.8%). Our unsupervised learning with UMAP (Uniform manifold approximation and projection, accuracy = 0.97) and hierarchical clustering (accuracy = 1.0) identified two clusters congruent with their histology. We next developed and compared four supervised models (random forest, support vector machine, generalized linear model with L2 regularization, and supervised UMAP). Supervised UMAP has shown to classify all the cases correctly (sensitivity = 1, specificity = 1, accuracy = 1) followed by random forest models (sensitivity = 0.84, specificity = 1, accuracy = 1). This pipeline can be used as a clinical tool by pathologists to differentiate chRCC from RO.

Keywords: kidney neoplasms; chromophobe; oncocytoma; supervised machine learning; classification

1. Introduction

Chromophobe renal cell carcinoma (chRCC) and renal oncocytoma (RO) are renal tumor types, each generally termed as oncocytic renal neoplasms [1] that arise from collecting ducts and constitute about 8–13% of all renal tumors [1,2]. ChRCC is a slow-growing malignant tumor that, once metastasized, has a poor prognosis [3]. RO is a benign tumor that does not require treatment [3]. Due to prognostic differences, it is essential to differentiate these tumors to guide treatment. Both of these tumors are diagnosed using hematoxylin and eosin (H&E) and immunohistochemistry (IHC) staining [4]. Histologically these tumors are similar; however, features such as binucleate or multinucleated cells, nuclear wrinkling, perinuclear clearing, presence of mitotic figures, and cytoplasmic invaginations are distinctive to chRCC and can help differentiate chRCC from RO. However, given substantial morphological overlap, especially in biopsies, the distinction between the two may be challenging despite the aforementioned distinctive morphologic characteristics [4]. Due to their similarity in histology and lack of reliable IHC markers, the distinction between chRCC and RO can be challenging.

Immunohistochemical detection of CK7 is used in laboratories with variable sensitivity and specificity [4–6]. Other studies showed that the inclusion of CD117, vimentin, and S100A1 in clinical diagnosis distinguishes chRCC from RO [4,6,7]. Several recent studies [7–9] have suggested using genetic and molecular assays as well as deep learning-based image detection to classify these tumors. Throughout the literature, smaller sample sizes have been the main limitation leading to variation in marker sensitivity and specificity [6,8–10]. In an effort to resolve interobserver variability and misdiagnosis, comprehensive guidelines and additional markers for renal tumor differentiation are needed.

We previously reported the Chromophobe-Oncocytoma-related gene signature (COGS) and a bioinformatics pipeline that can differentiate chRCC from RO and be implemented as a clinical tool [11]. Here we present our results on a retrospective validation of the gene signature at Augusta University Medical Center (AUMC). We developed a diagnostic workflow using a single-molecule counting assay to use formalin-fixed paraffin-embedded (FFPE) samples to differentiate these tumors. We demonstrate the robustness of the gene-signature and bioinformatics pipeline for the classification of chRCC from RO with unsupervised and supervised models.

2. Materials and Methods

2.1. Human Subjects and Study Design

We identified patients (*n* = 82) who attended AUMC between 2003 and 2018. These patients underwent tumor excision or needle biopsies to diagnose renal masses identified in abdominal images. The inclusion criterion was a histological diagnosis of chRCC or RO. Core/fine needle biopsy samples were excluded. Power analysis showed that six samples per group were required to identify gene expression differences between chRCC and RO for COGS genes ($\beta = 1.65$, $1 - \beta = 0.8$, $\alpha = 0.05$). Based on inclusion criteria.

We retrieved archived formalin-fixed paraffin-embedded (FFPE) tissues diagnosed as chRCC (*n* = 17) or RO (*n* = 15) from the Department of Pathology at AUMC. The tissue samples (*n* = 32) had been surgically excised from renal tumors and embedded in paraffin after fixing with formalin (FFPE); 5μm thick sections were taken for histological diagnosis (Figure 1).

Three AUMC Pathologists examined the selected sample set and identified the tumor region on FFPE blocks while also evaluating the matching H&E slides.

Demographic and clinical data were collected using the Augusta University Cancer Registry and validated through the patients' electronic health records. The study was conducted according to the Declaration of Helsinki (1997, revised in 2013) and approved by the Institutional Review Board at Augusta University (Biomarkers and Therapeutics in Cancer, 611205-49).
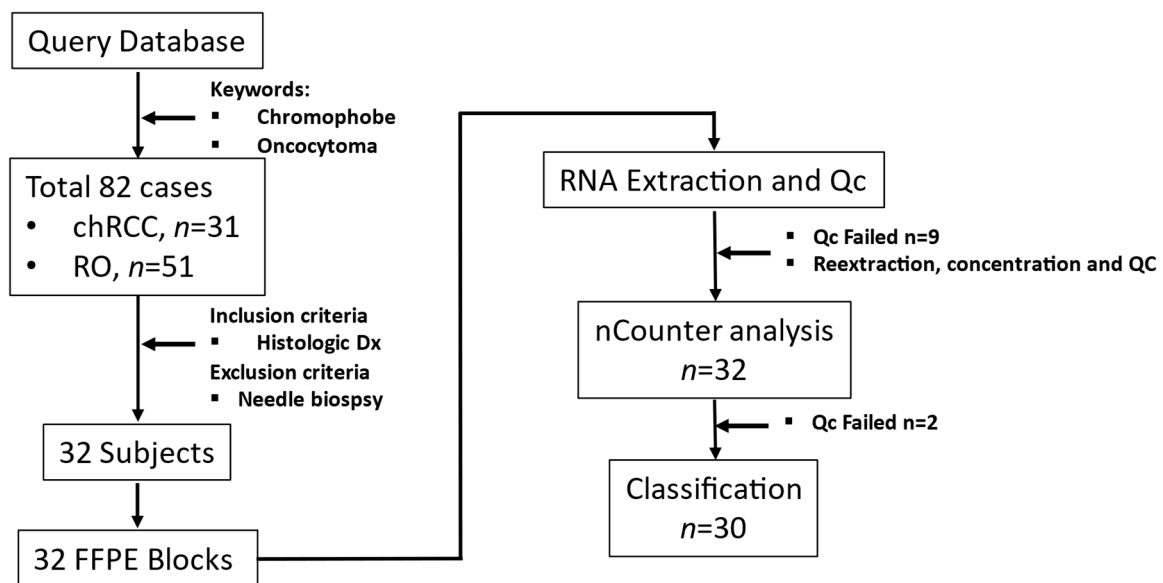
Figure 1. Study workflow from sample collection to RNA quantitation.

2.2. Coring, RNA Extraction, and Quality Control

Tumor-rich region marked by the pathologist was cored using a 2 mm biopsy punch (Integra Miltex, York, PA, USA). Paraffin was removed from the tissue core by treatment with Citrisolve (Fisher Scientific, Piscataway, NJ, USA) at 65 °C. The tissue was then homogenized in 10 mM Tris HCl-EDTA (pH 7.4) by mechanical disruption. According to the manufacturer's instructions, RNA from lysates was extracted using an RNA extraction kit designed for FFPE tissue (RNEasy, Qiagen, Hilden, Germany). RNA quality, quantity, and concentration were checked using Qbit and TapeStation 2200 (Supplemental Figure S1).

2.3. Quantification of Gene Expression Data

Custom nCounter assay consisting of gene expression code set for 33 genes (including three housekeeping genes) was synthesized (NanoString Technology Inc., Seattle, WA, USA) [12]. Two hundred nanograms of RNA/sample were hybridized into the reporter probe. Data capture was performed on a NanoString nCounter Digital Analyzer (NanoString Technology Inc., Seattle, WA, USA) and exported as reporter code count (RCC) files. These RCC files were analyzed using nSolver analysis software (Version 4.0.70, NanoString Technology Inc., Seattle, WA, USA) for quality control purposes (imaging, binding density, positive spike in control, and limit of detection).

The output from nSolver software was read into R for data pre-processing. The data was initially normalized by their concentration, followed by background thresholding. It was normalized first by the geometric mean of the code set's internal positive controls (Supplemental Figure S2A,B) and then the geometric mean of the housekeeping genes included in our assay (*HPRT1*, *LDHA*, and *TBP*) (Supplemental Figure S2C,D). The data was then log2-transformed and subjected to our machine learning models described below. We calculated the optimum cut point for each gene by calculating sensitivity plus specificity at each percentile. The value at which sensitivity plus specificity is maximum is called the optimum cut point. This data was then used as the "test" dataset in the machine learning models.

2.4. Discovery Data, Train-Validation-Test Set, Supervised Models

COGS was developed from a meta-analysis dataset from multiple GEO studies containing 53 chRCC and 36 RO arrays [11]. This dataset was split randomly by 80:20 into train and validation sets for unsupervised and supervised learning with the caret R package [13]. "Test" dataset contains the entire NanoString dataset for COGS genes, 30 samples

(2 samples failed QC for low count across all COGS genes) with the clinical diagnosis of 16 chRCC and 14 RO cases.

Unsupervised learning models were developed with UMAP and Hierarchical clustering. UMAP was implemented as a feature extraction algorithm to develop two components from COGS expression in the training dataset [14]. As UMAP is a stochastic algorithm, it was checked with 100 iterations by visual inspection of cluster formation, each with a different random seed. The hierarchical model was developed using the "Euclidean" distance and "ward.D2" clustering method. For the supervised UMAP model, a supervised embedding was developed using the train set to identify the internal structure of the training dataset. This supervised embedding was developed as a form of feature engineering and classification models were built on that new space. Test data was passed to this embedding for the classification of the samples.

Multiple supervised models (random forest, generalized linear model with L2 regularization, and support vector machine) were trained on COGS training data [11] with three repeats of 10-fold cross-validation and tested on Test data. These models were selected based on their performance in previous gene expression studies in the literature [15–17]. The discovery data introduced a UMAP manifold for supervised UMAP models and tested on NanoString data. NanoString data's UMAP projection was used to cluster the data. A contingency matrix was created for all models' results and performance metrics (Sensitivity, specificity, positive and negative predictive value) were calculated. The *p*-value for the metrics was estimated by bootstrap distribution where applicable. Classification metrics, i.e., sensitivity, specificity, positive predictive value, negative predictive value, and accuracy, were calculated by constructing a confusion matrix for each model's true positive, false positive, true negative, and false-negative rate.

2.5. Statistical Methods

All statistical analyses were performed using the R language and environment for statistical computing (v4.1.2; R Foundation for Statistical Computing, Indianapolis, IN, USA). Software packages were used for cutpointr, caTools, UMAP, ComplexHeatmap, Caret, and tidy models in the R environment. Continuous data are presented as mean and standard deviation, and differences in the group means were tested using Student's t-test. Categorical variables are presented as count and percentages and differences in count data were evaluated using Chi-square tests. All *p*-values were two-sided and a $p < 0.05$ was considered significant.

3. Results

3.1. Human Subject Demographics

The demographic information on subjects diagnosed with chRCC ($n = 17$) and RO ($n = 15$) is presented in Table 1. The mean age of onset of chRCC and RO was 57.8 (SD = 8.51) and 64.3 (SD = 13.9) years. Patients with RO patients significantly are older than chRCC (*p*-value = 0.003). There is a higher number of females in RO; overall, there were no significant differences in the distribution of males and females in both tumor types (Table 1). All chRCC tumors were staged. The majority of the cases were Stage I (64%), whereas Stage II and III were 17% each. RO cases were not staged. The race breakdown for chRCC was 70% Caucasian and 29% African American, while RO was 60% Caucasian and 33% African American.

Table 1. Clinical and Demographic information for chRCC and RO cases in the AUMC cohort.

Variable	chRCC [a]	RO [b]	Overall	*p*-Value
Stage, *n*(%)				
1	11(64.7%)	NA	11(34.4%)	
2	3(17.6%)	NA	3(9.4%)	
3	3(17.6%)	NA	3(9.4%)	
Age at Diagnosis, Mean (SD)	57.8(14.7)	71.5(8.51)	64.3(13.9)	0.003 *
Race, *n* (%)				
AA	5(29.4%)	5(33.3%)	10(31.3%)	0.55 **
C	12(70.6%)	9(60%)	21(65.6%)	
Other	0	1(6.7%)	1(3.1%)	
Gender, *n* (%)				
F	8(47.1%)	3(20%)	11(34.4%)	0.11**
M	9(52.9%)	12(80%)	21(65.6)	
Tumor size (mean)	5.15(2.4)	4.68(4.05)	4.93(3.22)	0.696 **

[a] chRCC: chromophobe renal cell carcinoma, [b] RO: renal oncocytoma cases are not staged, * Student's *t*-test, ** Chi-Square test.

3.2. Univariate Analysis

We performed a univariate AUC analysis for differentiating chRCC and RO with count difference. The sensitivity was 1 for nine genes (*AQP6, NDUFS1, MAP4K3, HOOK2, ESRP1, ELMO3, BSPRY, PRDX3,* and *LIMS1*) with a median of 0.94. Seven genes have a specificity of 1 (*AQP6, AP1M2, ITGB3, LRFN5, RSPO3, SPINT2,* and *LSR*) and a median of 0.93. The minimum AUC was 0.5 (*LAMA1*) with a median of 0.93 and a maximum of 1 (*AQP6*). Genes with the highest difference in AUC analysis were *AQP6* (AUC = 1.00), *BSPRY* (AUC = 0.99), *HOOK2* (AUC = 0.99), *SPINT2* (AUC = 0.99), *ESPR1* (AUC = 0.99), *ITGB3* (AUC = 0.98). The maximum is 1 (*AQP6*) for accuracy, and the median is 0.90. Amongst all the COGS genes, *AQP6* has the highest sensitivity, specificity, AUC, and accuracy. The full table is presented in Table 2.

Table 2. Univariate analysis of chRCC and RO NanoString Data. All values were calculated at the optimum cut point between chRCC and RO count data.

Gene	Optimal Cut Point	Accuracy	AUC	Sensitivity	Specificity
AP1M2	10.16	0.97	0.96	0.94	1
AQP6	13.24	1	1	1	1
ATP2C1	11.38	0.87	0.94	0.94	0.79
BSPRY	9.2	0.97	0.99	1	0.93
CLDN8	12.57	0.93	0.93	0.94	0.93
DNAI3	6.51	0.9	0.93	0.86	0.94
ELMO3	9.32	0.97	0.97	1	0.93
ESRP1	9.94	0.97	0.99	1	0.93
HOOK2	9.02	0.97	0.99	1	0.93
ITGB3	9.97	0.97	0.98	0.93	1
KCNG3	7.62	0.8	0.83	0.86	0.75
KIDINS220	9.8	0.9	0.92	0.93	0.88
KRT7	8.28	0.9	0.96	0.94	0.86
LAMA1	5.63	0.6	0.5	0.36	0.81
LIMS1	9.26	0.97	0.98	1	0.94
LRFN5	9.55	0.8	0.88	0.63	1
LSR	10.43	0.93	0.96	0.88	1
MANEA	8.66	0.9	0.9	0.86	0.94
MAP4K3	8.62	0.93	0.98	1	0.88
MSH2	8.37	0.67	0.54	0.86	0.5

Table 2. *Cont.*

Gene	Optimal Cut Point	Accuracy	AUC	Sensitivity	Specificity
NDUFS1	11.33	0.9	0.91	1	0.81
PLCL1	10.78	0.87	0.9	0.93	0.81
PLCL2	11.18	0.73	0.75	0.56	0.93
PNPT1	8.68	0.87	0.82	0.86	0.88
PRDX3	11.94	0.97	0.96	1	0.94
RSPO3	8.83	0.83	0.88	0.69	1
S100A1	7.12	0.87	0.88	0.93	0.81
SOCS1	9.32	0.93	0.98	0.94	0.93
SPINT2	13	0.97	0.99	0.94	1
SUCLA2	10.83	0.87	0.85	0.86	0.88

AUC: Area under the curve.

3.3. Unsupervised and Supervised Machine Learning Models

As single markers can fail to differentiate chRCC and RO, we implemented multivariate models using unsupervised learning. We implemented UMAP and hierarchical clustering to see the combined markers' ability to differentiate these tumors. UMAP analysis on NanoString count data showed two clusters, and 29/30 samples were correctly identified (Supplemental Figure S3). Then we developed a supervised UMAP model using COGS discovery data as the training set by setting parameters to n_neighbors = 15, epoch = 200, min_dist = 0.1, init = "spectral", metric = "Euclidean" (Figure 2A) [11]. This trained model was tested on the AUMC cohort. This trained UMAP model created two clusters congruent with their histology (30/30 samples are correctly classified) (Figure 2B). Hierarchical clustering showed 100% congruency with their histological classification (Figure 2C).
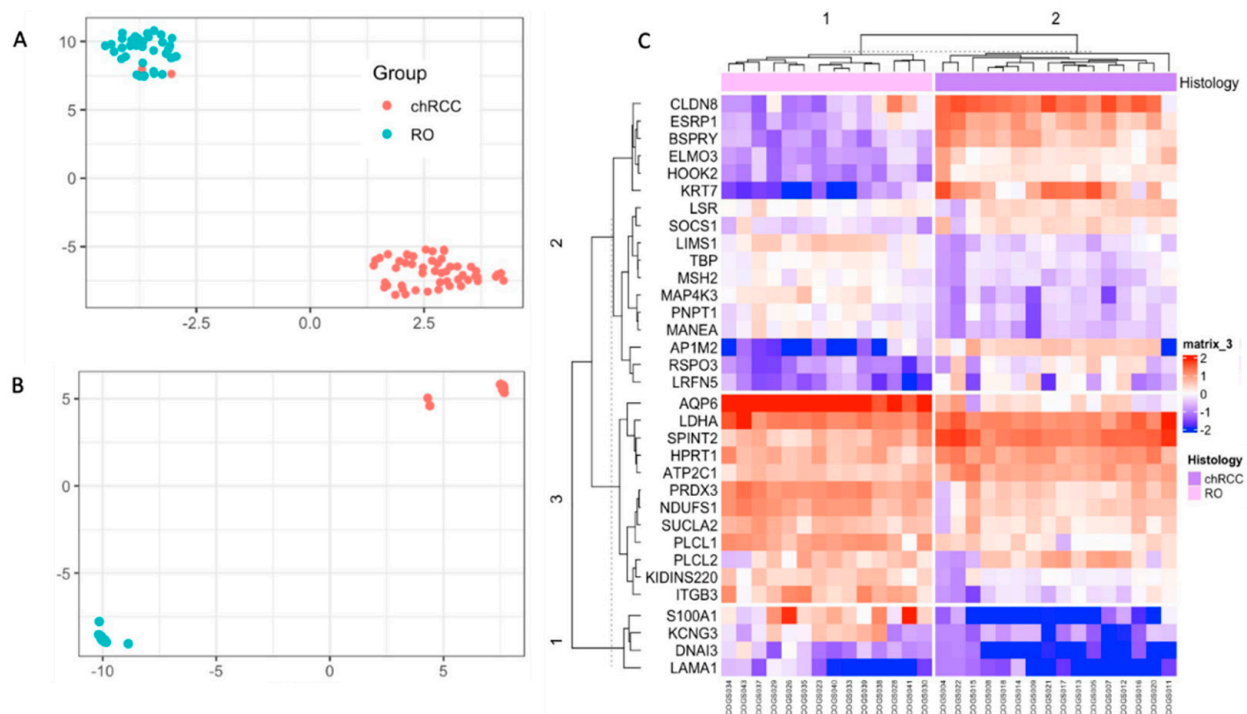

Figure 2. Unsupervised classification of the NanoString data with UMAP and Hierarchical Clustering. (**A**): UMAP of discovery data showing two clusters were used to train the supervised UMAP model. (**B**): UMAP projection of the NanoString data using the trained UMAP model from A. (**C**): Hierarchical clustering showing two clusters for the samples, congruent with their histological classification.

Although unsupervised models are a powerful tool for identifying data structure in high dimensional space, they are not ideal for testing a new sample. Therefore, we compared four supervised learning models, viz., random forest, support vector machine, and a generalized linear model with L2 regularization trained with COGS' discovery data with 10-fold cross-validation with three repeats. The four models were compared (Table 3) by sensitivity, specificity, positive predictive value, negative predictive value, and accuracy.

Table 3. Supervised Model metrics for classification of chRCC and RO.

Metric	Random Forest	SVM	GLM	Supervised UMAP
Sensitivity	0.84	0.76	0.88	1
Specificity	1	1	1	1
Accuracy	0.93	0.83	0.9	1
95% CI	0.78–0.99	0.65–0.94	0.73–0.98	-
p-value	4.74×10^{-5}	0.07659	0.001066	0
PPV	1	1	1	1
NPV	0.87	0.64	0.78	1

SVM: Support Vector Machine, GLM: Generalized Linear Model, UMAP: Uniform Manifold Approximation and Projection, PPV: Positive Predictive Value, NPV: Negative Predictive Value.

Supervised UMAP showed to be most accurate in classifying the chRCC and RO based on their COGS expression (Sensitivity = 1.0, Specificity = 1.0, PPV = 1.0, NPV = 1.0 and Accuracy = 1.0). As Supervised UMAP did not have any misclassification, the *p*-value is 0, and the 95% confidence interval cannot be calculated. Random forest models performed better than the other SVM and GLMNet models. For random forest models, a minimum of 250 trees were needed to achieve an accuracy of 1 (Figure 3A). A sample tree is presented in Figure 3B. Although support vector machines are popular in classification analysis, they did not reach statistical significance in our dataset (Figure 3C, Table 3). Coefficients for the generalized linear model with ridge regression are presented in Figure 3D. All the supervised models were used to create a confusion matrix, and sensitivity, specificity, positive predictive value, negative predictive value, and accuracy were calculated (Table 3). All the *p*-values were calculated by accuracy over no information rate for the models.

3.4. H&E Scoring between Pathologists

Three independent AUMC pathologists evaluated H&E for the selected sample set (*n* = 32). Two of the three pathologists had one discrepancy in diagnosis from the original clinical diagnosis, while the third pathologist had two discrepancies. We developed a linear mixed model with clinical diagnosis as the fixed effects and H&E and scorer as the random effects. The model showed that H&E accounts for 0.918 (*p*-value <0.001) of the variation in the diagnosis of these cases. The student t-test between the clinical diagnosis and pathologist impression or between the pathologist impression did not show any difference (*p*-value > 0.05). All the raw data is provided in Supplemental Table S1.
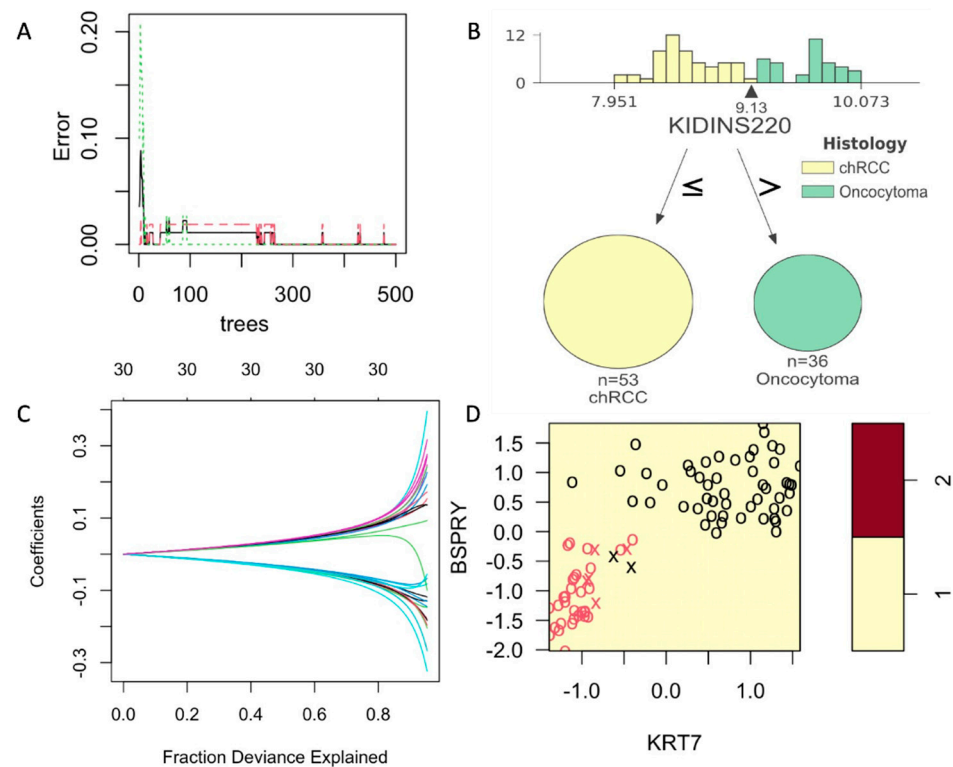
Figure 3. Supervised models with training on COGS discovery data and tested on NanoString data. (**A**): Figure 3: Development and comparison of four supervised models. Random Forest models showing a minimum of 250 trees can achieve the error = 0. (**B**): A sample tree showing the distribution of gene *KIDINS220* expression on COGS discovery data with the arrow pointing at the best cutoff value. (**C**): A sample support vector model with *KRT7* and *BSPRY*. (**D**): Generalized linear model with Ridge regression showing the fraction explained by the COGS genes as covariates. All these metrics show the results on NanoString data (AUMC cohort).

4. Discussion

In this study, we tested the potential of COGS to differentiate chRCC and RO renal tumor types (Supplemental Table S2). We also developed a single-molecule counting assay in a single institutional cohort. Our univariate analysis demonstrates that the majority of the genes in the COGS signature shows expression level difference between the tumor types and can serve as potential biomarkers to differentiate these tumors (Table 2). We showed that the COGS profile between these tumors is distinct in unsupervised models (UMAP and hierarchical clustering) (Figure 2B,C). To classify future samples, we developed and compared four different supervised models (random forest, support vector machine, generalized linear model, and supervised UMAP) (Figure 3A,D) and identified supervised UMAP outperforming all the other models in classifying chRCC and ROs (Table 3). This workflow can be implemented in clinical settings to quantify and correctly classify future samples suspected of chRCC or RO.

The current clinical workflow to differentiate these tumors is based on H&E impressions and certain IHC markers. H&E with or without IHC suffers in sensitivity and specificity because there is a lacking of definitive features on H&E or the overlapping expression of the immunohistochemical markers. This is even more difficult in biopsy cases where there is limited tissue. CK7 is the most widely used marker specific for chRCC with variable sensitivity (0.73–1.0) and specificity (0.84–1.0) [18,19]. S100A1 is the specific IHC marker for RO with variable sensitivity (0.87–1.0) and specificity (0.7–1.0) [20,21]. The variability in these markers can leave up to 10% of cases to be misdiagnosed by traditional clinical workflow. IHC staining can be subjective and requires optimization as the reading is dependent on the quality of the staining [22]. In comparison, gene expression workflows

follow a strict protocol, have high quality-control measures, and require a smaller amount of tissue. In this study, we used a 2 mm punch biopsy core (~3–4 mm in length) to collect tumor tissue which is a smaller amount of tissue than core biopsies used in the current clinical workflow [23]. Our workflow requires 72 h from tissue to RNA quantification, which is comparable to pathologist evaluation, IHC processing, and optimization. Our NanoString validation pipeline correctly identified 100% of the cases analyzed in this study. This limits the possibility of inconclusive biopsy results, repeat biopsies [24], as well as inadequate treatment for chRCC or unnecessary surgical intervention for RO cases. This workflow can overcome the obstacles such as inconclusive biopsy results or interobserver variability between pathologists with the integration of supervised learning. This will significantly improve the patient's prognosis and lower the economic burden as it will lower unnecessary surgical intervention, i.e., partial or total nephrectomy. *AQP6*, *HOOK2*, *AP1M2*, *ESPR1*, and *CLDN* from the COGS panel were already suggested in the previous studies as potential biomarkers [25,26]. We report *AQP6, AP1M2, LIMS1, PRDX3, SPINT2, BSPRY, ELMO3, ESPR1, HOOK2, ITGB3* (sensitivity > 0.93, specificity > 0.93, AUC > 0.96, and accuracy > 0.93) as top-performing genes in our study and can serve as potential transcriptomic biomarkers between these tumors.

A limitation of using bulk transcriptomics is that the expression profile includes tumor cells and stromal cells, immune cells, etc. Future studies are needed to identify the differences between tumor cells themselves in chRCC and RO and incorporate stroma and immune-specific genes into the signature. Patients with RO were older in age compared to chRCC as the peak age of detection of RO is the seventh decade, whereas peak chRCC detection is the sixth decade [27]. Another limitation was due to the FFPE RNA being highly fragmented, which resulted in some genes not showing the observed differences in our discovery data, such as LAMA1 and MSH2. To work around the fragmentation issues, multivariate models were implemented as they account for such variation due to technical or biological reasons. Although SVM models are one of the most well-understood machine learning algorithms, in this study, they did not reach statistical significance. This could have stemmed from the smaller sample size (*n* = 30) collected from a single institution. Therefore, an additional study is needed to evaluate these findings in a larger sample set collected from multiple sites.

The strength of this study is the direct clinical application. The supervised model training was on bulk transcriptomic data from multiple GEO studies previously validated in microarray and RNASeq platform, which are now translated into the NanoString platform. Therefore, quantifying the expression of COGS genes assayed on any platform can be conducted and followed up by the implementation of the machine learning models. This workflow can be performed directly on biopsies by extracting RNA and quantifying it using the NanoString platform. Our method has shown an overall greater accuracy (Supervised UMAP, accuracy = 1.0) in comparison to the current clinical workflow [18,20,21,28].

We have shown that the single-molecule counting on the nCounter platform working in conjunction with our machine-learning pipeline can be a viable clinical assay for differentiation of chRCC from RO. This approach has been shown to predict therapeutic outcomes in breast cancer [29], diagnosis of small B-cell lymphoid neoplasm [30] cancer, and classification of gliomas [31].

5. Conclusions

We report a pipeline that uses gene expression profiling by single-molecule counting to differentiate chRCC from RO, which can be an important clinical tool in renal cancer diagnosis.

References

1. Moch, H.; Ohashi, R. Chromophobe Renal Cell Carcinoma: Current and Controversial Issues. *Pathology* **2021**, *53*, 101–108. [CrossRef] [PubMed]
2. Nicolau, C.; Antunes, N.; Paño, B.; Sebastia, C. Imaging Characterization of Renal Masses. *Medicina* **2021**, *57*, 51. [CrossRef] [PubMed]
3. Atkins, M.; Choueiri, T. Epidemiology, Pathology, and Pathogenesis of Renal Cell Carcinoma—UpToDate. Available online: https://www.uptodate.com/contents/epidemiology-pathology-and-pathogenesis-of-renal-cell-carcinoma (accessed on 2 March 2021).
4. Williamson, S.R.; Gadde, R.; Trpkov, K.; Hirsch, M.S.; Srigley, J.R.; Reuter, V.E.; Cheng, L.; Kunju, L.P.; Barod, R.; Rogers, C.G.; et al. Diagnostic Criteria for Oncocytic Renal Neoplasms: A Survey of Urologic Pathologists. *Hum. Pathol.* **2017**, *63*, 149–156. [CrossRef] [PubMed]
5. Tan, P.H.; Cheng, L.; Rioux-Leclercq, N.; Merino, M.J.; Netto, G.; Reuter, V.E.; Shen, S.S.; Grignon, D.J.; Montironi, R.; Egevad, L.; et al. Renal Tumors: Diagnostic and Prognostic Biomarkers. *Am. J. Surg. Pathol.* **2013**, *37*, 1518–1531. [CrossRef]
6. Carvalho, J.C.; Wasco, M.J.; Kunju, L.P.; Thomas, D.G.; Shah, R.B. Cluster Analysis of Immunohistochemical Profiles Delineates CK7, Vimentin, S100A1 and C-Kit (CD117) as an Optimal Panel in the Differential Diagnosis of Renal Oncocytoma from Its Mimics. *Histopathology* **2011**, *58*, 169–179. [CrossRef]
7. Zhang, L.; Henske, E.P. Chromophobe Renal Cell Carcinoma: New Genetic and Metabolic Insights. *Urol. Oncol.* **2020**, *38*, 678–681. [CrossRef]

8. Baghdadi, A.; Aldhaam, N.A.; Elsayed, A.S.; Hussein, A.A.; Cavuoto, L.A.; Kauffman, E.; Guru, K.A. Automated Differentiation of Benign Renal Oncocytoma and Chromophobe Renal Cell Carcinoma on Computed Tomography Using Deep Learning. *BJU Int.* **2020**, *125*, 553–560. [CrossRef]
9. Brennan, K.; Metzner, T.J.; Kao, C.-S.; Massie, C.E.; Stewart, G.D.; Haile, R.W.; Brooks, J.D.; Hitchins, M.P.; Leppert, J.T.; Gevaert, O. Development of a DNA Methylation–Based Diagnostic Signature to Distinguish Benign Oncocytoma From Renal Cell Carcinoma. *JCO Precis. Oncol.* **2020**, *4*, 1141–1151. [CrossRef]
10. Gowrishankar, B.; Przybycin, C.G.; Ma, C.; Nandula Subhadra, V.; Rini, B.; Campbell, S.; Klein, E.; Chaganti, R.S.K.; Magi-Galluzzi, C.; Houldsworth, J. A Genomic Algorithm for the Molecular Classification of Common Renal Cortical Neoplasms: Development and Validation. *J. Urol.* **2015**, *193*, 1479–1485. [CrossRef]
11. Satter, K.B.; Tran, P.M.H.; Tran, L.K.H.; Ramsey, Z.; Pinkerton, K.; Bai, S.; Savage, N.M.; Kavuri, S.; Terris, M.K.; She, J.-X.; et al. Oncocytoma-Related Gene Signature to Differentiate Chromophobe Renal Cancer and Oncocytoma Using Machine Learning. *Cells* **2022**, *11*, 287. [CrossRef]
12. Geiss, G.K.; Bumgarner, R.E.; Birditt, B.; Dahl, T.; Dowidar, N.; Dunaway, D.L.; Fell, H.P.; Ferree, S.; George, R.D.; Grogan, T.; et al. Direct Multiplexed Measurement of Gene Expression with Color-Coded Probe Pairs. *Nat. Biotechnol.* **2008**, *26*, 317–325. [CrossRef] [PubMed]
13. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; R Core Team; et al. *Caret: Classification and Regression Training*; Astrophysics Source Code Library: Online, 2020.
14. Basic UMAP Parameters—Umap 0.5 Documentation. Available online: https://umap-learn.readthedocs.io/en/latest/parameters.html (accessed on 17 August 2021).
15. Roy, S.; Kumar, R.; Mittal, V.; Gupta, D. Classification Models for Invasive Ductal Carcinoma Progression, Based on Gene Expression Data-Trained Supervised Machine Learning. *Sci. Rep.* **2020**, *10*, 4113. [CrossRef] [PubMed]
16. Piles, M.; Fernandez-Lozano, C.; Velasco-Galilea, M.; González-Rodríguez, O.; Sánchez, J.P.; Torrallardona, D.; Ballester, M.; Quintanilla, R. Machine Learning Applied to Transcriptomic Data to Identify Genes Associated with Feed Efficiency in Pigs. *Genet. Sel. Evol.* **2019**, *51*, 10. [CrossRef] [PubMed]
17. Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genom. Proteom.* **2018**, *15*, 41–51. [CrossRef]
18. Zhou, J.; Yang, X.; Zhou, L.; Zhang, P.; Wang, C. Combined Immunohistochemistry for the "Three 7" Markers (CK7, CD117, and Claudin-7) Is Useful in the Diagnosis of Chromophobe Renal Cell Carcinoma and for the Exclusion of Mimics: Diagnostic Experience from a Single Institution. *Dis. Markers* **2019**, *2019*, 4708154. [CrossRef]
19. Wu, S.L.; Kothari, P.; Wheeler, T.M.; Reese, T.; Connelly, J.H. Cytokeratins 7 and 20 Immunoreactivity in Chromophobe Renal Cell Carcinomas and Renal Oncocytomas. *Mod. Pathol.* **2002**, *15*, 712–717. [CrossRef]
20. Kuroda, N.; Kanomata, N.; Yamaguchi, T.; Imamura, Y.; Ohe, C.; Sakaida, N.; Hes, O.; Michal, M.; Shuin, T.; Lee, G.-H. Immunohistochemical Application of S100A1 in Renal Oncocytoma, Oncocytic Papillary Renal Cell Carcinoma, and Two Variants of Chromophobe Renal Cell Carcinoma. *Med. Mol. Morphol.* **2011**, *44*, 111–115. [CrossRef]
21. Rocca, P.C.; Brunelli, M.; Gobbo, S.; Eccher, A.; Bragantini, E.; Mina, M.M.; Ficarra, V.; Zattoni, F.; Zamò, A.; Pea, M.; et al. Diagnostic Utility of S100A1 Expression in Renal Cell Neoplasms: An Immunohistochemical and Quantitative RT-PCR Study. *Mod. Pathol.* **2007**, *20*, 722–728. [CrossRef]
22. Kim, S.-W.; Roh, J.; Park, C.-S. Immunohistochemistry for Pathologists: Protocols, Pitfalls, and Tips. *J. Pathol. Transl. Med.* **2016**, *50*, 411–418. [CrossRef]
23. Caoili, E.M.; Davenport, M.S. Role of Percutaneous Needle Biopsy for Renal Masses. *Semin. Interv. Radiol.* **2014**, *31*, 20–26. [CrossRef]
24. Leveridge, M.J.; Finelli, A.; Kachura, J.R.; Evans, A.; Chung, H.; Shiff, D.A.; Fernandes, K.; Jewett, M.A.S. Outcomes of Small Renal Mass Needle Core Biopsy, Nondiagnostic Percutaneous Biopsy, and the Role of Repeat Biopsy. *Eur. Urol.* **2011**, *60*, 578–584. [CrossRef] [PubMed]
25. Wu, H.; Fan, L.; Liu, H.; Guan, B.; Hu, B.; Liu, F.; Hocher, B.; Yin, L. Identification of Key Genes and Prognostic Analysis between Chromophobe Renal Cell Carcinoma and Renal Oncocytoma by Bioinformatic Analysis. *Biomed Res. Int.* **2020**, *2020*, 4030915. [CrossRef] [PubMed]
26. Rohan, S.; Tu, J.J.; Kao, J.; Mukherjee, P.; Campagne, F.; Zhou, X.K.; Hyjek, E.; Alonso, M.A.; Chen, Y.-T. Gene Expression Profiling Separates Chromophobe Renal Cell Carcinoma from Oncocytoma and Identifies Vesicular Transport and Cell Junction Proteins as Differentially Expressed Genes. *Clin. Cancer Res.* **2006**, *12*, 6937–6945. [CrossRef] [PubMed]
27. Ng, K.L.; Rajandram, R.; Morais, C.; Yap, N.Y.; Samaratunga, H.; Gobe, G.C.; Wood, S.T. Differentiation of Oncocytoma from Chromophobe Renal Cell Carcinoma (RCC): Can Novel Molecular Biomarkers Help Solve an Old Problem? *J. Clin. Pathol.* **2014**, *67*, 97–104. [CrossRef] [PubMed]
28. Wobker, S.E.; Williamson, S.R. Modern Pathologic Diagnosis of Renal Oncocytoma. *J. Kidney Cancer VHL* **2017**, *4*, 1–12. [CrossRef]
29. Krug, K.; Jaehnig, E.J.; Satpathy, S.; Blumenberg, L.; Karpova, A.; Anurag, M.; Miles, G.; Mertins, P.; Geffen, Y.; Tang, L.C.; et al. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell* **2020**, *183*, 1436–1456.e31. [CrossRef]

30. Zhang, W.; Ao, Q.; Guan, Y.; Zhu, Z.; Kuang, D.; Li, M.M.Q.; Shen, K.; Zhang, M.; Wang, J.; Yang, L.; et al. A Novel Diagnostic Approach for the Classification of Small B-Cell Lymphoid Neoplasms Based on the NanoString Platform. *Mod. Pathol.* **2022**, *35*, 632–639. [CrossRef]

31. Tran, P.M.H.; Tran, L.K.H.; Satter, K.B.; Purohit, S.; Nechtman, J.; Hopkins, D.I.; dos Santos, B.; Bollag, R.; Kolhe, R.; Sharma, S.; et al. Retrospective Validation of a 168-Gene Expression Signature for Glioma Classification on a Single Molecule Counting Platform. *Cancers* **2021**, *13*, 439. [CrossRef]