

# Final Progress Report for Research Projects Funded by Health Research Grants

Instructions: Please complete all of the items as instructed. Do not delete instructions. Do not leave any items blank; responses must be provided for all items. If your response to an item is “None”, please specify “None” as your response. “Not applicable” is not an acceptable response for any of the items. There is no limit to the length of your response to any question. Responses should be single-spaced, no smaller than 12-point type. The report **must be completed using MS Word**. Submitted reports must be Word documents; they should not be converted to pdf format. Questions? Contact Health Research Program staff at 717-783-2548.

1. **Grantee Institution:** Carnegie Mellon University
2. **Reporting Period (start and end date of grant award period):** June 1, 2012–May 31, 2014
3. **Grant Contact Person (First Name, M.I., Last Name, Degrees):** Rhonda Kloss
4. **Grant Contact Person’s Telephone Number:** 412-268-1015
5. **Grant SAP Number:** 4100059192
6. **Project Number and Title of Research Project:** 01 - Automated Biomarker Identification for Cancer Detection and Prognosis
7. **Start and End Date of Research Project:** June 1, 2012-May 31, 2014
8. **Name of Principal Investigator for the Research Project:** Robert F. Murphy, Ph.D.
9. **Research Project Expenses.**

9(A) Please provide the total amount of health research grant funds spent on this project for the entire duration of the grant, including indirect costs and any interest earned that was spent:

\$ 926,572.14

9(B) Provide the last names (include first initial if multiple individuals with the same last name are listed) of **all** persons who worked on this research project and were supported with health research funds. Include position titles (Principal Investigator, Graduate Assistant, Post-doctoral Fellow, etc.), percent of effort on project and total health research funds expended for the position. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project	Cost
Murphy, Robert	Principal Investigator	8% Yr 1 & 2	35,143
Rohde, Gustavo	Co-Investigator	8% Yr 1 & 2	21,520
Anil Parwani	Co-Investigator	10% Yr 1 & 2	34,000
Ozolek, John	Co-Investigator	10% Yr 1 & 2	9,583
Tosun, Akif Burak	Post Doctoral Fellow	60% Yr 1 & 2	89,787
Naik, Armaghan	Post Doctoral Fellow	0% yr1 / 25% yr 2	15,000
Buck, Taraz	Post Doctoral Fellow	0% yr 1 / 33% yr2	12,000
Kumar, Aparna	Graduate Student	100% Yr 1 & 2	135,808
Liu, Chi	Graduate Student	0% yr 1/ 25% yr 2	9,300
Chen, Cheng	Graduate Student	0% yr 1/ 20% yr 2	6,198
Cao-Berg, Ivan	Senior Research Programmer	0% yr 1 / 45% yr2	24,750
Bakal, Jennifer	Research Programmer	0% yr1 / 12% yr 2	6,345
Wilson, Jonathan	Research Associate	42% yr 1 /100% yr 2	43,650

9(C) Provide the names of **all** persons who worked on this research project, but who *were not* supported with health research funds. Include position titles (Research Assistant, Administrative Assistant, etc.) and percent of effort on project. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project
None		

9(D) Provide a list of **all** scientific equipment purchased as part of this research grant, a short description of the value (benefit) derived by the institution from this equipment, and the cost of the equipment.

Type of Scientific Equipment	Value Derived	Cost
4 Power Edge R420	For developing, testing and running algorithms on both Human Protein Atlas images and UPMC images	17,910
Hitachi Ultrastar ZTB SATA III 7.2K	For transporting large volumes of images from UPMC to Carnegie Mellon	1,922
Omnyx Scanner, HP Server, IDP Software, Pathologist Workstation, Histologist Workstation	For imaging stained tissue slides	136,387

**10. Co-funding of Research Project during Health Research Grant Award Period.** Did this research project receive funding from any other source during the project period when it was supported by the health research grant?

Yes \_\_\_\_\_ No X \_\_\_\_\_

If yes, please indicate the source and amount of other funds:

### 11. Leveraging of Additional Funds

11(A) As a result of the health research funds provided for this research project, were you able to apply for and/or obtain funding from other sources to continue or expand the research?

Yes \_\_\_\_\_ No X \_\_\_\_\_

If yes, please list the applications submitted (column A), the funding agency (National Institutes of Health—NIH, or other source in column B), the month and year when the application was submitted (column C), and the amount of funds requested (column D). If you have received a notice that the grant will be funded, please indicate the amount of funds to be awarded (column E). If the grant was not funded, insert “not funded” in column E.

Do not include funding from your own institution or from CURE (tobacco settlement funds). Do not include grants submitted prior to the start date of the grant as shown in Question 2. If you list grants submitted within 1-6 months of the start date of this grant, add a statement below the table indicating how the data/results from this project were used to secure that grant.

A. Title of research project on grant application	B. Funding agency (check those that apply)	C. Month and Year Submitted	D. Amount of funds requested:	E. Amount of funds to be awarded:
None	<input type="checkbox"/> NIH <input type="checkbox"/> Other federal (specify: _____) <input type="checkbox"/> Nonfederal source (specify: _)		\$	\$

11(B) Are you planning to apply for additional funding in the future to continue or expand the research?

Yes X \_\_\_\_\_ No \_\_\_\_\_

If yes, please describe your plans:

Drs. Rohde and Ozolek, in conjunction with Dr. Peter Shaw of the Children's Hospital of Pittsburgh, plan on utilizing some of the results presented in our final report as preliminary evidence for future grant applications on topics related to the prognosis of several childhood cancers.

**12. Future of Research Project.** What are the future plans for this research project?

We plan to continue analyzing the data acquired as part of this project. For example, the segmented nuclei database will be re-analyzed to remove potentially erroneous segmentations. Dr. Murphy plans to explore whether combinations of proteins will provide improved classification accuracy. Dr. Rohde plans on recomputing the classification results reported using more accurate (and more computationally expensive) image approximations, which are necessary for the transport-based morphometry approach employed. In addition, Dr. Rohde also plans to compare the results obtained with his transport-based methodology to several other approaches utilizing the standard numerical feature descriptors often employed.

**13. New Investigator Training and Development.** Did students participate in project supported internships or graduate or post-graduate training for at least one semester or one summer?

Yes  No

If yes, how many students? Please specify in the tables below:

	Undergraduate	Masters	Pre-doc	Post-doc
Male				3
Female			1	
Unknown				
<b>Total</b>			<b>1</b>	<b>3</b>

	Undergraduate	Masters	Pre-doc	Post-doc
Hispanic				
Non-Hispanic			1	3
Unknown				
<b>Total</b>			<b>1</b>	<b>3</b>

	Undergraduate	Masters	Pre-doc	Post-doc
White				3
Black				
Asian				
Other			1	
Unknown				
<b>Total</b>			<b>1</b>	<b>3</b>

**14. Recruitment of Out-of-State Researchers.** Did you bring researchers into Pennsylvania to carry out this research project?

Yes  No

If yes, please list the name and degree of each researcher and his/her previous affiliation:

Dr. Akif Burak Tosun (Ph.D.) was recruited from Bilkent University in Turkey for this project.

**15. Impact on Research Capacity and Quality.** Did the health research project enhance the quality and/or capacity of research at your institution?

Yes  No

If yes, describe how improvements in infrastructure, the addition of new investigators, and other resources have led to more and better research.

This project formed the major part of the Ph.D. thesis research of Aparna Kumar, who anticipates defending her thesis in Computational Biology early in the fall. She did the primary work on the location biomarkers project. She is currently doing a summer internship at Omnyx in Pittsburgh and may take a regular position there at the end of the summer.

The Omnyx scanner that was purchased through this project provides an important infrastructure component for future digital pathology projects.

**16. Collaboration, business and community involvement.**

16(A) Did the health research funds lead to collaboration with research partners outside of your institution (e.g., entire university, entire hospital system)?

Yes  No

If yes, please describe the collaborations:

This project led to a new and useful collaboration between Carnegie Mellon University, the University of Pittsburgh and Omnyx, Inc.

16(B) Did the research project result in commercial development of any research products?

Yes  No

If yes, please describe commercial development activities that resulted from the research project:

A provisional patent application has been submitted for the method described in Ozolek et al. (2014).

16(C) Did the research lead to new involvement with the community?

Yes \_\_\_\_\_ No X \_\_\_\_\_

If yes, please describe involvement with community groups that resulted from the research project:

**17. Progress in Achieving Research Goals, Objectives and Aims.**

List the project goals, objectives and specific aims (as contained in the grant agreement). Summarize the progress made in achieving these goals, objectives and aims for the period that the project was funded (i.e., from project start date through end date). Indicate whether or not each goal/objective/aim was achieved; if something was not achieved, note the reasons why. Describe the methods used. If changes were made to the research goals/objectives/aims, methods, design or timeline since the original grant application was submitted, please describe the changes. Provide detailed results of the project. Include evidence of the data that was generated and analyzed, and provide tables, graphs, and figures of the data. List published abstracts, poster presentations and scientific meeting presentations at the end of the summary of progress; peer-reviewed publications should be listed under item 20.

This response should be a DETAILED report of the methods and findings. It is not sufficient to state that the work was completed. Insufficient information may result in an unfavorable performance review, which may jeopardize future funding. If research findings are pending publication you must still include enough detail for the expert peer reviewers to evaluate the progress during the course of the project.

Health research grants funded under the Tobacco Settlement Act will be evaluated via a performance review by an expert panel of researchers and clinicians who will assess project work using this Final Progress Report, all project Annual Reports and the project's strategic plan. After the final performance review of each project is complete, approximately 12-16 months after the end of the grant, this Final Progress Report, as well as the Final Performance Review Report containing the comments of the expert review panel, and the grantee's written response to the Final Performance Review Report, will be posted on the CURE Web site.

**There is no limit to the length of your response. Responses must be single-spaced below, no smaller than 12-point type. If you cut and paste text from a publication, be sure symbols print properly, e.g., the Greek symbol for alpha ( $\alpha$ ) and beta ( $\beta$ ) should not print as boxes ( $\square$ ) and include the appropriate citation(s). DO NOT DELETE THESE INSTRUCTIONS.**

## **Project Overview**

The goal of this project is to carry out translational research to further develop two related new cancer diagnostic methods for which proof-of-concept has been demonstrated. Carnegie Mellon University has intellectual property consisting of algorithms for (1) determining the subcellular location of marker proteins from immunohistochemistry images and identifying proteins that change subcellular location between normal and tumor tissue, and (2) robustly identifying changes in nuclear morphology that can distinguish normal from cancerous tissues in several pathologies in the liver. The research seeks to build on these results to create commercially viable diagnostic products that can be marketed by Omnyx, a pioneering digital pathology company headquartered in Pittsburgh.

The first specific aim is to perform translational research studies to determine the value of automated image analysis technology developed at Carnegie Mellon University for providing enhanced diagnostic and prognostic information in adult prostate and pediatric/adult liver tumors. These studies will be done using paraffin-embedded tissue selected from tissue banks to cover a range of tumor grades, initial diagnoses and clinical outcome. The images will be analyzed to determine the prognostic value of the phenotypes measured by the two technologies. Additional image analysis methods will be developed as needed to improve upon the existing technologies. The second specific aim is to interface the automated technology with the Omnyx digital pathology platform and develop an appropriate and efficient clinician interface to the technology in order to facilitate the ability of the clinician to integrate its outputs into the diagnostic process.

## **Summary of Research Completed**

### *Aim 1*

#### *Improvement of algorithms for detecting Location Biomarkers*

Over the two years of the project, we have significantly improved and done extensive further testing on our initial algorithms for identifying location biomarkers. This work is under revision for submission to the Proceedings of the National Academy of Sciences. The processing pipeline has been streamlined and made more efficient. Additions include optimization of identification of regions for feature calculation, and improved quantification of expression level differences. Using images from the Human Protein Atlas, we have measured the ability of each candidate location marker to distinguish between normal and cancerous tissue in four tissues (breast, prostate, liver and urinary bladder), and shown that a number of the markers have accuracies over 80% and some over 90% for distinguishing normal from cancerous tissue. Some of these results are presented in Figure 1 and Table 1.

We have also added a new capability, the ability to identify biochemical pathways that contain significant numbers of proteins that change location in cancer. These are listed in Table 2. It is important to note that the names of the pathways are based on the context in which they were originally described, and that they may play a role in more than one process. Identification of their potential involvement in cancer may help elucidate such roles.

### *Improvement of algorithms for detecting morphology differences*

Over the two years of the project, we have improved and expanded our capabilities for segmenting and analyzing the morphology of nuclei from histopathology images under a variety of stains. Improvements on the image segmentation front have been made by rendering the related numerical optimization methods more robust with respect to initialization parameters. In addition, a graphical user interface has been added to the package so as to render the training and testing process (necessary for segmenting different stains) more straightforward and less time consuming. We have tested the code with a variety of image stains (H&E, Feulgen, Agnor, fluorescence, silver, etc.).

We have also added new capabilities for analyzing nuclear morphology to our currently existing software package. More specifically, we have implemented a pipeline for performing patient-based blind cross validation utilizing a variety of well-known classifiers, including: linear Fisher discriminant analysis, penalized linear discriminant analysis, linear and nonlinear support vector machines, as well as K-nearest neighbor methods, including kernel versions. These are now set up to work with either common-place numerical features extracted from images of nuclei, or using our linearized optimal transport framework for measuring distances between nuclei. Finally, scripts allowing for one to utilize the newly setup computer cluster to run such cross validation jobs in parallel have been set up.

While images were being collected for the studies below, we sought to test the effectiveness of the transport-based image analysis pipeline using alternative datasets. We utilized already imaged data of Feulgen stained thyroid specimens to test whether the segmentation and nuclear morphometry pipeline outlined above could be used to effectively differentiate a variety of benign and malignant cancers using a large cohort of patients. During the course of this research, we experimented with several classification methods for single nuclei, in combination with voting strategies for sets of nuclei for a single patient. The results of this effort were recently published (Ozolek, et al., 2014) and are summarized in Table 3 below.

### *Reading whole slide images from Omnyx scanners*

We have produced software to export whole slide images from the native Omnyx format to both compressed and uncompressed Jpeg2000 images, and to read these images on our analysis system (the images are approximately 200 gigabytes so this is not a trivial task). We now have Matlab scripts and C code for loading and processing large images sequentially (by loading single image regions at one time).

### *Identification of staining targets - liver*

The following staining probes or antibodies have been identified for use in the liver study; the list includes nuclear stains, existing liver biomarkers and new potential location biomarkers that we have identified using the methods described above. The markers are Hematoxylin and eosin, Feulgen, Glutamine synthetase, L-FABP (liver fatty acid binding protein), CRP (C-reactive protein), HSP70 (heat shock protein 70), Beta-catenin, Glypican-3, ki67, DKC1 (dyskerin), NDUFAF1 (NADH dehydrogenase 1 alpha subcomplex, assembly factor 1), NPM1 (nucleophosmin), DEK (DEK oncogene), IRX6 (Iroquois homeobox 6), and TJP1 (tight junction protein 1).

#### *Identification of staining targets - prostate*

We have identified the following list of markers to measure for the prostate study; the list includes both existing prostate cancer biomarkers and new potential location biomarkers that we have identified using the methods described above. The markers are AMACR, ki67, AR, COX-2, TMPRSS2-ERG, Akt-1, PSCA, EGR-1, beta-catenin, FOLH1, HOX3, PSMAL, SOX-4, and Cyclin B1, TMEM194a, DHAH100S, RASGRF2, TMOD3, and MAVS.

#### *Identification of subjects*

We identified tissue blocks from patients with adult and pediatric liver lesions for use in this study. They are itemized below. Note that the number of subjects projected in the original proposal was incorrect; it was an initial estimate that had not been corrected based on the cost per case and the budget available. It should have been 100 subjects, not 300.

We have identified appropriate case specific paraffin-embedded prostate tissue blocks using an honest broker system at the Health Science Tissue Bank. 11 cases of normal donor prostate (NDP), 37 cases of normal tissue adjacent to prostatic adenocarcinoma (NAC), 15 cases of benign prostatic hyperplasia (BPH), 35 cases of high-grade prostatic intraepithelial neoplasia (HGPIN), 103 cases of primary prostatic adenocarcinoma (PCa), and 36 cases of metastatic prostatic adenocarcinoma (Mets) have been selected.

#### *Preparation of slides and image collection*

For the liver project, we obtained 186 terabytes of whole slide images, for 86 subjects and 15 antibodies plus traditional hematoxylin/eosin staining.

For the prostate project, we have not yet obtained images from Dr. Parwani's group. The slides are still being scanned and we will receive the images when they are collected.

#### *Results from classification of liver images using location biomarkers*

Four types of adult lesions were analyzed: Dysplastic nodules (DN, 13 cases), Focal nodular hyperplasia (FNH, 23 cases), Hepatocellular carcinoma (HCC, 16 cases), and Macroregenerative nodules (MRN, 14 cases). For beta-catenin, CRP, DEK, DKC, GS, HSP70, KI67, FABP, NDUFAF1, NPM1, and TJP, the accuracies for distinguishing the four classes all ranged between 0.375 and 0.495 (for reference, an accuracy of 0.457 was obtained using Feulgen staining alone). Two proteins showed accuracies above 50%, IRX6 and Glypcan-3. The confusion matrices for them are shown below.

Antibody=IRX6; Accuracy = 0.55

True/pred	DN	FNH	HCC	MRN	
DN		0.69	0.15	0.07	0.09
FNH		0.05	0.89	0.02	0.04
HCC		0.30	0.34	0.08	0.28
MRN		0.22	0.19	0.05	0.54

Antibody=Glypican-3; Accuracy = 0.62

True/pred	DN	FNH	HCC	MRN	
DN		0.70	0.11	0.09	0.10
FNH		0.01	0.99	0.00	0.00
HCC		0.25	0.14	0.45	0.16
MRN		0.38	0.15	0.15	0.31

Three types of pediatric lesions were examined: Fetal hepatoblastoma (FHB, 6 cases), Hepatocellular adenoma (HCS, 4 cases), and Focal nodular hyperplasia (FNH, 3 cases). For CRP and L-FABP, accuracies ranged from 0.383-0.425. For beta-catenin, HSP70, IRX6, KI67, NDUFAF1, NPM1 and TJP!, accuracies ranged 0.513-0.63. These were all below the accuracy for Feulgen staining alone, 0.663. For DEK, DKC, GS, and Glypican-3, the accuracies were in the range of 0.675-0.717. The confusion matrices for the best two are shown below.

Antibody=DEK; Accuracy = 0.72

True/pred	DN	FNH	HCC	
DN		0.56	0.33	0.11
FNH		0.18	0.70	0.13
HCC		0.06	0.05	0.89

Antibody=Glypican-3; Accuracy = 0.70

True/pred	DN	FNH	HCC	
DN		0.80	0.10	0.10
FNH		0.29	0.39	0.33
HCC		0.03	0.05	0.93

*Results from classification of liver images using morphology*

Table 4 shows results obtained to date using the liver images obtained through this project. These were for 8 diagnostic challenges selected by Dr. JA Ozolek. They were selected due to their being common clinical decisions in current practice. For each diagnostic challenge, we report the average classification accuracy, as well as Cohen’s kappa, a statistic meant to evaluate how far the agreement is from chance (0 being pure chance, and 1 being complete agreement). With the exception of the Normal vs. Abnormal challenge, our preliminary results are good. A potential reason behind the poor results for Normal vs. Abnormal is that, the distribution of the data between normal and abnormal cases is unbalanced.

## *Aim 2*

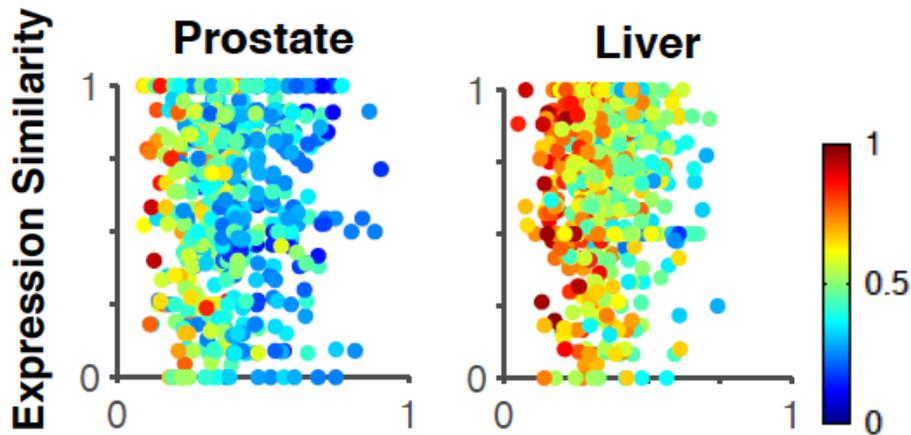
### *Software for display of location biomarkers results to clinicians*

Our location biomarkers software automatically selects regions from whole slide IHC images that are in focus and show antibody staining. These are used for classifying the sample as normal or abnormal. The issue then becomes how to communicate these results to the clinician. Recently, we have created an initial version of software for displaying the selected regions along a spectrum from most normal to most abnormal. Clicking on each image will display the region in the original whole slide image.

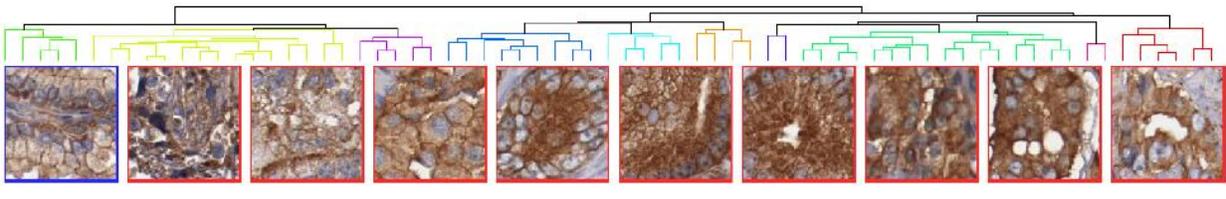
The idea behind this display is to first identify the primary “axis” in the feature space that differentiates normal regions from cancerous regions (for a given tissue). This is learned on training images for which the diagnosis is known. Features can then be calculated from regions of test images (regions from a patient image for which diagnosis is desired) and converted to a position along this axis. This allows all test regions to be displayed in order of their “normality.” An example display using clustering is shown in Figure 2. Because of the importance that we place on this project, we will complete development of this software with support from internal Carnegie Mellon sources.

### *Software for display of morphology results to clinicians*

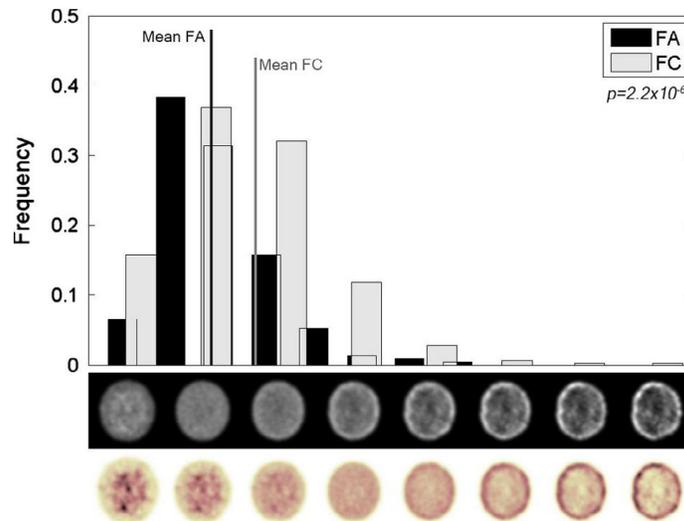
In (Ozolek, et al., 2014) we describe a method for visualizing the differences between two cell populations (e.g. benign vs. malignant). The method works by combining our previous work on penalized linear discriminant analysis (Wang, Mo, Ozolek, & Rohde, 2011), and linear optimal transport (Wang, Slepcev, Ozolek, Basu, & Rohde, 2013) (Basu, Kolouri, & Rohde, 2014). It modifies the approach presented earlier in (Wang, Slepcev, Ozolek, Basu, & Rohde, 2013) by selecting the regularization parameter of the penalized LDA technique we described in (Wang, Mo, Ozolek, & Rohde, 2011) by regression with an exponential function, as we described in (Basu, Kolouri, & Rohde, 2014). During experimentation with the thyroid H&E images we found this approach to be more robust with respect to our previous work. A representative result is shown in Figure 3 below. The source code for this software is available at <http://www.andrew.cmu.edu/user/gustavor/software.html>. A graphical user interface allowing pathologists to utilize the software more seamlessly is the subject of current development, using internal funds.



**Figure 1.** Distinguishing between intensity and location changes. The system was used to calculate the p-values that a given protein changed its location between normal and cancer tissue. In parallel, whether the intensity of that protein had changed was determined by calculating the number of images that had different intensity annotations between the two disease states. Lastly, some of the images for each protein were used to train a classifier to distinguish the two states. The color of each point indicates the classification accuracy for separating *test* normal and cancer images that were not used for training. Note that a number of proteins with low p-values (high likelihood that they changed location) were able to distinguish normal and cancerous images with high accuracies.



**Figure 2.** Ordering regions by location change progression. The regions of all images for a single protein are clustered into a binary hierarchical tree with optimal leaf ordering. The tree is cut at 10 clusters and leaves contained in each cluster are indicated by color. The region closest to the mean of each cluster is displayed below the tree. These results are for the protein TMOD3 in prostate. The image outlined in blue is from normal tissue, the tumor images are in red. Note that the images reveal progressive movement from the plasma membrane on the left (including the normal image) to a relatively homogenous cytoplasmic location.



**Figure 3** - Nuclei distribution histograms to distinguish between follicular adenoma (FA) and follicular carcinoma (FC). The rows of images of nuclei beneath the histogram bins are the normalized grayscale (upper) and colorized (lower) visual representations of nuclei along the optimal transportation pathway (geodesic) that best discriminates between FA and FC.

**Table 1** - Location biomarker rankings. For each antibody in each tissue, normal and cancer feature distributions were compared to calculate a p-value as in Figure 1. The ten proteins with the lowest p-values are shown for each tissue (after removing proteins that changed intensity). Classification accuracies for distinguishing normal and cancer are also shown.

LIVER LOCATION BIOMARKER RANKING

Rank	Ensembl	Gene Name	HPA Ab	p-value	Accuracy
1	ENSG00000100577	GSTZ1	HPA004701	0.078	0.907
2	ENSG00000060339	CCAR1	HPA007856	0.191	0.889
3	ENSG00000160789	LMNA	CAB004022	0.195	0.593
4	ENSG00000137806	NDUFAF1	HPA039933	0.2	0.722
5	ENSG00000112699	GMD5	HPA031528	0.2	0.741
6	ENSG00000107821	KAZALD1	HPA011800	0.205	0.833
7	ENSG00000185838	GNB1L	HPA034628	0.217	0.685
8	ENSG00000130826	-	HPA000447	0.218	0.648
9	ENSG00000181163	NPM1	CAB012983	0.232	0.944
10	ENSG00000197746	PSAP	HPA004426	0.239	0.796

PROSTATE LOCATION BIOMARKER RANKING

Rank	Ensembl	Gene Name	HPA Ab	p-value	Accuracy
1	ENSG00000151304	SRFBP1	HPA042737	0.0902	0.852
2	ENSG00000181163	NPM1	CAB012983	0.103	0.37
3	ENSG00000197746	PSAP	HPA004426	0.132	0.5
4	ENSG00000048828	FAM120A	HPA019734	0.134	0.704
5	ENSG00000029725	RABEP1	HPA024691	0.149	0.704
6	ENSG00000137154	RPS6	CAB004027	0.149	0.759
7	ENSG00000183098	GPC6	HPA017671	0.164	0.537
8	ENSG00000135094	SDS	HPA039230	0.173	0.778
9	ENSG00000138835	RGS3	CAB032507	0.179	0.481
10	ENSG00000116667	C1orf21	HPA026831	0.189	0.204

**Table 2.** Pathways with significant location change but no significant expression change. Pathway p-values were calculated using individual protein p-values from the pipeline for the analysis set, and using p-values from gross pathologist annotations. For each type of p-value, those pathways whose p-value was lower than the Bonferroni-Holm threshold are marked (the threshold depends on the number of nodes in the pathway). These pathways show significant change in localization of their component proteins.

Pathway	Loc:Pipe	Exp:Pipe	Loc:Ann	Exp:Ann
<i>Liver</i>				
Chagas disease American trypanosomiasis	*		*	*
NOD like receptor signaling pathway	*		*	*
Long term depression	*			*
Valine, leucine and isoleucine degradation	*			*
Fatty acid elongation	*			*
Butanoate metabolism	*			
Thyroid cancer	*			
<i>Prostate</i>				
Amino sugar and nucleotide sugar metabolism	*			*
Histidine metabolism	*			

**Table 3.** Summary of classification performance comparisons between LOT and Feature based approaches.

Comparisons	LOT		Feature	
	Accuracy (%)	Cohen's Kappa	Accuracy (%)	Cohen's Kappa
FA vs. FC	100	1.00	87	0.74
FA vs. FVPC	100	1.00	97	0.93
FC vs. WIFC	100	1.00	76	0.45
FVPC vs. FC	87	0.67	90	0.78
FA vs. NG	100	1.00	84	0.67
FC vs. NG	100	1.00	81	0.60

Table 4– Classification results for liver dataset using morphology.

Comparisons	LOT	
	Accuracy (%)	Cohen's Kappa
Normal vs. Abnormal	60	0.01
Benign (MRN, FNH, HCA) vs. Malignant (DN, HCC)	82	0.63
(MRN, FNH) vs. HCA	90	0.65
DN vs. HCC	68	0.38
MRN vs. FNH	88	0.75
Normal vs. (FNH, MRN)	81	0.29
FHB vs. Normal (Kids only)	89	0.73
FHB vs. Normal (All )	92	0.83

### Bibliography

Basu, S., Kolouri, S., & Rohde, G. (2014). Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proc Natl Acad Sci* , 111, 3448-3453.

Cheng , C., Wang , W., Ozolek, J., & Rohde, G. (2013). A flexible and robust approach for segmenting cell nuclei from 2D microscopy images using supervised learning and template matching. *Cytometry A* , 83 (5), 495-507.

Ozolek, J., Tosun, A., Wang, W., Chen, C., Kolouri, S., Basu, S., Huang, H., Rohde, G.K. (2014). Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Medical Image Analysis* , 18 (5), 772-780.

Rohde, G., Ribeiro, A., Dahl, K., & Murphy , R. (2008). Deformation-based nuclear morphometry: Capturing nuclear shape variation in HeLa cells. 2008; 73A:. *Cytometry A* , 73, 341-350.

Wang, W., Mo, Y., Ozolek , J., & Rohde, G. (2011). Penalized Fisher discriminant analysis and its application to image-based morphometry. *Pattern Recognition Letters* , 32 (15), 2128-2135.

Wang, W., Ozolek, J., & Rohde, G. (2010). Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry Part A* , 77 (5), 485-494.

Wang, W., Slepcev, D., Ozolek, J., Basu, S., & Rohde, G. (2013). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int J Comput Vision* , 101, 254-269.

**18. Extent of Clinical Activities Initiated and Completed.** Items 18(A) and 18(B) should be completed for all research projects. If the project was restricted to secondary analysis of clinical data or data analysis of clinical research, then responses to 18(A) and 18(B) should be “No.”

18(A) Did you initiate a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

Yes

No

18(B) Did you complete a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

Yes

No

**If “Yes” to either 18(A) or 18(B), items 18(C) – (F) must also be completed.** (Do NOT complete 18(C-F) if 18(A) and 18(B) are both “No.”)

18(C) How many hospital and health care professionals were involved in the research project?

\_\_\_\_\_ Number of hospital and health care professionals involved in the research project

18(D) How many subjects were included in the study compared to targeted goals?

\_\_\_\_\_ Number of subjects originally targeted to be included in the study

\_\_\_\_\_ Number of subjects enrolled in the study

**Note:** Studies that fall dramatically short on recruitment are encouraged to provide the details of their recruitment efforts in Item 17, Progress in Achieving Research Goals, Objectives and Aims. For example, the number of eligible subjects approached, the number that refused to participate and the reasons for refusal. Without this information it is difficult to discern whether eligibility criteria were too restrictive or the study simply did not appeal to subjects.

As described in item 17, the original number of cases expected for the liver study was an error.

18(E) How many subjects were enrolled in the study by gender, ethnicity and race?

Gender:

\_\_\_\_\_ Males

\_\_\_\_\_ Females

\_\_\_\_\_ Unknown

Ethnicity:

\_\_\_\_\_ Latinos or Hispanics

\_\_\_\_\_ Not Latinos or Hispanics

\_\_\_\_\_ Unknown

Race:

- American Indian or Alaska Native  
 Asian  
 Blacks or African American  
 Native Hawaiian or Other Pacific Islander  
 White  
 Other, specify: \_\_\_\_\_  
 Unknown

18(F) Where was the research study conducted? (List the county where the research study was conducted. If the treatment, prevention and diagnostic tests were offered in more than one county, list all of the counties where the research study was conducted.)

**19. Human Embryonic Stem Cell Research.** Item 19(A) should be completed for all research projects. If the research project involved human embryonic stem cells, items 19(B) and 19(C) must also be completed.

19(A) Did this project involve, in any capacity, human embryonic stem cells?

- Yes  
 No

19(B) Were these stem cell lines NIH-approved lines that were derived outside of Pennsylvania?

- Yes  
 No

19(C) Please describe how this project involved human embryonic stem cells:

**20. Articles Submitted to Peer-Reviewed Publications.**

20(A) Identify all publications that resulted from the research performed during the funding period and that have been submitted to peer-reviewed publications. Do not list journal abstracts or presentations at professional meetings; abstract and meeting presentations should be listed at the end of item 17. **Include only those publications that acknowledge the Pennsylvania Department of Health as a funding source** (as required in the grant agreement). List the title of the journal article, the authors, the name of the peer-reviewed publication, the month and year when it was submitted, and the status of publication (submitted for publication, accepted for publication or published.). Submit an electronic copy of each publication or paper submitted for publication, listed in the table, in a PDF version 5.0.5 (or greater) format, 1,200 dpi. Filenames for each publication should include the number of the research project, the last name of the PI, and an abbreviated title of the publication. For example, if you submit two publications for Smith (PI for Project 01), one

publication for Zhang (PI for Project 03), and one publication for Bates (PI for Project 04), the filenames would be:

- Project 01 – Smith – Three cases of isolated
- Project 01 – Smith – Investigation of NEB1 deletions
- Project 03 – Zhang – Molecular profiling of aromatase
- Project 04 – Bates – Neonatal intensive care

If the publication is not available electronically, provide 5 paper copies of the publication.

**Note:** The grant agreement requires that recipients acknowledge the Pennsylvania Department of Health funding in all publications. Please ensure that all publications listed acknowledge the Department of Health funding. If a publication does not acknowledge the funding from the Commonwealth, do not list the publication.

Title of Journal Article:	Authors:	Name of Peer-reviewed Publication:	Month and Year Submitted:	Publication Status (check appropriate box below):
1. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning	Ozolek, J., Tosun, A., Wang, W., Chen, C., Kolouri, S., Basu, S., Huang, H., Rohde, G.K.	Medical Image Analysis	12/2013	<input type="checkbox"/> Submitted <input type="checkbox"/> Accepted <input checked="" type="checkbox"/> Published
2. Automated Analysis of Immunohistochemistry Images Identifies Candidate Location Biomarkers for Cancers	Kumar, A., Rao, A., Bhavani S., Newberg, J.Y., Murphy, R.F.	Proc. Natl. Acad. Sci. U.S.A.	8/2014	<input checked="" type="checkbox"/> Submitted <input type="checkbox"/> Accepted <input type="checkbox"/> Published

20(B) Based on this project, are you planning to submit articles to peer-reviewed publications in the future?

Yes   X   No \_\_\_\_\_

If yes, please describe your plans:

At least two manuscripts describing the classification studies (one for each technologies) will be submitted.

**21. Changes in Outcome, Impact and Effectiveness Attributable to the Research Project.**

Describe the outcome, impact, and effectiveness of the research project by summarizing its impact on the incidence of disease, death from disease, stage of disease at time of diagnosis, or other relevant measures of outcome, impact or effectiveness of the research project. If there were no changes, insert “None”; do not use “Not applicable.” Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

This translational project has further developed two technologies for automated digital pathology, and demonstrated that they provide the ability to classify/distinguish different subtypes of cancers. There is significant potential impact on diagnosis of disease severity, but it has not yet been demonstrated in a prospective study.

**22. Major Discoveries, New Drugs, and New Approaches for Prevention Diagnosis and Treatment.** Describe major discoveries, new drugs, and new approaches for prevention, diagnosis and treatment that are attributable to the completed research project. If there were no major discoveries, drugs or approaches, insert “None”; do not use “Not applicable.” Responses must be single-spaced below, and no smaller than 12-point type. **DO NOT DELETE THESE INSTRUCTIONS.** There is no limit to the length of your response.

The project demonstrated that IRX6, DEF and Glypican-3 are valuable as biomarkers for distinguishing different liver lesions. The project also developed new methods for communicating results from automated analyses to clinicians.

**23. Inventions, Patents and Commercial Development Opportunities.**

23(A) Were any inventions, which may be patentable or otherwise protectable under Title 35 of the United States Code, conceived or first actually reduced to practice in the performance of work under this health research grant? Yes \_\_\_\_\_ No  X

If “Yes” to 23(A), complete items a – g below for each invention. (Do NOT complete items a - g if 23(A) is “No.”)

- a. Title of Invention:
- b. Name of Inventor(s):
- c. Technical Description of Invention (describe nature, purpose, operation and physical, chemical, biological or electrical characteristics of the invention):
- d. Was a patent filed for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?  
Yes \_\_\_\_\_ No \_\_\_\_\_

If yes, indicate date patent was filed:

- e. Was a patent issued for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?  
Yes \_\_\_\_\_ No \_\_\_\_\_

If yes, indicate number of patent, title and date issued:

Patent number:

Title of patent:

Date issued:

- f. Were any licenses granted for the patent obtained as a result of work performed under this health research grant? Yes \_\_\_\_\_ No \_\_\_\_\_

If yes, how many licenses were granted? \_\_\_\_\_

- g. Were any commercial development activities taken to develop the invention into a commercial product or service for manufacture or sale? Yes \_\_\_\_\_ No \_\_\_\_\_

If yes, describe the commercial development activities:

23(B) Based on the results of this project, are you planning to file for any licenses or patents, or undertake any commercial development opportunities in the future?

Yes  X  No \_\_\_\_\_

If yes, please describe your plans:

We plan to offer commercial licenses for both technologies through Carnegie Mellon, most likely on a non-exclusive licensing basis (as an add-on “app” for digital pathology systems).

**24. Key Investigator Qualifications.** Briefly describe the education, research interests and experience and professional commitments of the Principal Investigator and all other key investigators. In place of narrative you may insert the NIH biosketch form here; however, please limit each biosketch to 1-2 pages. *For Nonformula grants only – include information for only those key investigators whose biosketches were not included in the original grant application.*

This is a nonformula grant. Biosketches were provided in grant application.