Automated analysis of protein subcellular location in immunohistochemistry images for cancer diagnosis

Aparna Kumar
January 2015
Revised May 2021

CMU-CB-21-101

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Dr. Robert F. Murphy (chair)
Dr. Russell Schwartz
Dr. Chakra Chennubhotla
Dr. Gustavo Rohde
Dr. John Ozolek

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Research reported in this thesis was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under award number T32EB009403, the National Institute of General Medical Sciences under award number R01GM075205, and the Pennsylvania Department of Health Commonwealth Universal Research Enhancement (CURE) program under award number 4100059192.

Keywords: Immunohistochemistry, pathology, automated image analysis, cancer, protein subcellular location, biomarkers, liver lesions

Abstract

Protein subcellular location and compartmentalization play an important role in regulating cellular processes. Protein mislocalization alters cell signaling and is observed in diverse diseases (Hung and Link 2011). Drug resistance can occur when proteins are mislocalized to the cytoplasm and nucleus, suggesting that the measurement of protein location can help clinicians personalize therapies and diagnose disease. Here, two projects explore how automatically quantitating subcellular location from pathology images can be used be in diagnostics and for understanding disease. 1) We developed an automated pipeline to compare the subcellular location of proteins between two sets of immunohistochemistry images. We used the pipeline to compare images of healthy and tumor tissue from the Human Protein Atlas, ranking hundreds of proteins in breast, liver, prostate and bladder based on how much their location was estimated to have changed. The performance of the system was evaluated by determining whether proteins previously known to change location in tumors were ranked highly. We present a number of new candidate location biomarkers for each tissue. Further we identified biochemical pathways that are enriched in proteins that change location. We confirmed some previously implicated pathways and we report new pathways previously unassociated with cancer to have changed. 2) We extended the IHC pipeline to process full slide images. Using the pipeline we explored how measuring changes in protein subcellular location can aid in identifying adult and pediatric liver lesions. Our results indicate that most of the time single protein measurements are poor markers for the lesions. Next we explored lesion.specific protein signatures for identifying diseases. Given our dataset we found a signature set of proteins that can successfully identify liver lesions in adult and pediatric populations with perfect accuracy. Finally we report two new proteins that aid in classifying the lesions when used as part of a signature protein set.

Acknowledgements

I would like to thank my thesis advisor Dr. Robert F. Murphy for the opportunity to pursue this degree and to be a member of the Murphy lab. The knowledge, experience and lessons I have learned are invaluable and have allowed me to become a stronger researcher and individual.

I would also like to thank my committee members for their involvement and guidance: Dr. Gustavo Rohde, Dr. John Ozolek and Dr. Chakra Chennubhotla.

In addition I would like to specially thank all of the professors, TAs, teachers and friends who have impacted my time in graduate school.

I would like to thank the current and past members of the Murphy lab, specifically my fantastic officemates Taraz Buck and Greg Johnson.

I would like to thank all of my friends and family for their support throughout my time in Pittsburgh, and especially to Ermine and Ceren for the non-stop laughs and fond memories.

Table of Contents

Chapter 1

Introduction

The field of pathology is currently undergoing a revolutionary change and digital pathology is quickly advancing field. The goal is to develop computational tools to aid in more accurate and objective pathology diagnoses. Robust computational methods are allowing for the discovery of new cancer markers that were previously unappreciated. In classic biology and translational medicine protein subcellular location is acknowledged to play an important role in cellular regulation. Misregulation is reported to play a role in the development of disease. Protein subcellular location has not been well studied in the context of disease; a system wide quantitative study of differential location patterns is necessary to improve our understanding. This thesis presents an automated computational pipeline for analyzing protein subcellular location in immunohistochemistry images. The pipeline is used to 1) identify potential location biomarkers from thousands of IHC images, and 2) used to assess how protein location can be used to discriminate adult and pediatric liver lesions.

Background

Challenges in cancer diagnoses: the need for cancer-specific markers

Cancer diagnosis is challenging due to the heterogeneity of the disease across the population. Cancers arising from the same tissue may be driven by different mutations and different types of aberrant signaling, effectively making each cancer a unique disease. The presence of specific proteins and other biological characteristics, called biomarkers, can indicate the types of mutations and regulation patterns in the tissue and ultimately allow clinicians to make a diagnosis.

One of the most frequently used and reliable methods to diagnose cancer is to

perform a biopsy at the site of question. During a biopsy a physician collects a small sample of the tissue. The tissue is prepared into an immunohistochemistry (IHC) slide by a histologist, and finally a pathologist performs a differential diagnosis by visually comparing and scoring the differences in growth patterns and the chemical composition, compared to the normal tissue.

Immunohistochemistry

Histology is the study of microscopic anatomy of plants and animals on a cellular and tissue level, while histopathology is the microscopic study of diseased tissue. Immunohistochemistry (IHC) is the mounting of tissue on a glass slide where antigens of interest are detected by protein-specific antibodies.

IHC is an excellent source of protein location in a tissue, and also in a cell. In IHC images the tissue morphology, tissue architecture, protein expression and spatial organization are preserved. Unlike other protein profiling experiments, IHC image analysis samples are not homogenized giving the data an additional dimension, location.

Tissue sections must be carefully prepared to retain the structure. First the tissue is fixed with paraformaldehyde to preserve the structure. Next the tissue is sliced or sectioned. It is paraffin embedded so the specimen can be handled without compromising the quality. The paraffin-tissue block is slice with a microtome at 4-40um. The slices are mounted on a glass slide and dehydrated with alcohol. The mounted tissue is dyed depending on the regions and proteins of interest. Hematoxylin is used to dye the nuclei of the cells in the tissue. Specific proteins and antigens are detected with antibodies. Next, the antibodies are detected with a secondary Ab that binds the immunoglobulin of the primary. The secondary is tagged with horseradish peroxidase (HRP), which reacts with diaminobenzidine (DAB)to give a brown stain (Ramos-Vara and Miller 2014) (Coons AH Creech HJ 1941).

Antibodies

Protein detecting antibodies are monospecific; they are designed to target the same antigen. Monospecific antibodies can either be polyclonal or monoclonal. Polyclonal antibodies are created by injecting a peptide fragment of interest into an animal. A secondary immune response is triggered in the animal and antibodies can be isolated from the animal serum. Monoclonal antibodies are formed from creating a cell line from the spleen of the animal. The cell line is grown in culture and monoclonal antibodies are harvested from the media. While monoclonal antibodies are selected to be very specific and bind to a single epitope, polyclonal antibodies are less expensive to manufacture and can be equally effective at detecting the antigen, making them a good detection method for tissue samples.

The images used in this thesis, 1) Human Protein Atlas (HPA) and 2) UPMC liver lesion data were produced from tissue sections stained with monospecific antibodies.

Tissue microarrays

Tissue microarrays (TMAs) are a high throughput screening format based on tissue sections (Fowler, Man et al. 2011). They are made up of small punches from paraffin embedded tissue. A malignant breast tissue TMA contains punch biopsies from hundreds of different breast cancer patients on a single controlled platform. One of the biggest advantages of a TMA is a side-by-side controlled comparison of sections of tissue from multiple sources. In addition the small punches require a small amount of reagents, making the system cost effective. Tissue screens would otherwise be unfeasible with single core slides. TMAs can be used for immunofluorescence, immunohistochemistry, in situ hybridization, and conventional histology staining. TMAs are particularly useful for screening antibodies for diagnostic or research purposes.

The HPA images used in Chapter 2 of this thesis were collected from TMAs. The controlled platform allowed for consistent sample preparation and image collection. This allowed us to do a simple cross analysis of tissue sections stained with different antibodies and from different tissue sources.

Scoring immunohistochemistry

Once TMAs are processed using IHC they must be scored. Manual scoring of TMAs by pathologists has been a bottleneck for tissue-based studies and drug screens. Visual examination to assess changes is a difficult and time-consuming task. When pathologists score TMAs multiple experts are usually involved and final scores are decided by a consensus. Controversial cases are discussed to minimize subjectivity and inter and intra observer variability (Arihiro, Umemura et al. 2007) (Scolyer, Shaw et al. 2003).

There can be significant variation when experts quantitate from day to day and facility to facility. This can lead to different scores for tissue sections in research and diagnostic purposes (Ghaznavi, Evans et al. 2013) (Bauer, Schoenfield et al. 2013). Computationally aided diagnosis can provide an objective, quantitative assessment of ambiguous slides (Gurcan, Boucheron et al. 2009).

Automated analysis of pathology images

With the advent of high throughput acquisition technologies like tissue microarrays and automated slide scanners, the computerized analysis of tissue images have made scoring and analysis feasible. In addition, computational analysis has been used to create objective quantitative diagnostic and research tools for pathology images.

One of the first reports on automated analysis for pathology was in 1969 by Mawdesley-Thomas and Healey (Mawdesley-Thomas and Healey 1969) on a

bioassay for the irritant effects of sulfur dioxide vapor by using image analysis to quantitate globlet cells from rats.

Since then many automated methods have been developed to quantitate and score TMAs. Some companies have reported quantitative software systems that measure expression in cells (Mulrane, Rexhepaj et al. 2008), and some reports mention automated quantitation of protein subcellular location and expression (Camp, Chung et al. 2002). Automated protein profiling using cell microarrays has also been reported (Stromberg, Bjorklund et al. 2007). Further automated methods have been applied to tissue specific TMAs for evaluation (Amin, Srinivasan et al. 2014). Similar methods have been used for molecular profiling of tumors (Kononen, Bubendorf et al. 1998), (Kallioniemi, Wagner et al. 2001). Such systems have increased the efficiency of screening by TMAs.

Automated analysis and quantitation have also been used in diagnostic and research applications. For example automated analysis has been used to identify and quantify macrovesicular steatosis in human livers (Nativ, Chen et al. 2014), for scoring DDS-induced colitis in mice (Kozlowski, Jeet et al. 2013), to quantitate kidney damage in rodent models (Klapczynski, Gagne et al. 2012), and for HER-2 status classification (Dobson, Conway et al. 2010). Reports include a variety of approaches including fractal methods to identify colon cancer (Esgiar, Naguib et al. 2002). Studies have shown that quantitative software can detect changes in disease states that are missed by visual inspection (Guillaud, Adler-Storthz et al. 2005) (Beck, Sangoi et al. 2011) supporting the development for accurate, quantitative and unbiased image analysis methods.

Automated methods have been developed for research applications as well, for example TMARKER that counts cells and quantitates stains (Schuffler, Fuchs et al. 2013), quality assurance tests (Webster, Simpson et al. 2011), and applications to identify connectomes (Kleinfeld, Bharioke et al. 2011).

Thus improving the state of the art in automated image analysis systems will contribute to further developments in high throughput analysis, diagnostic and research systems.

Protein subcellular location

The subcellular location of a protein defines its interacting partners and in turn defines the role the protein can play in signaling cascades and the effect it can have on the system. Subcellular location and translocation play an important role in the regulation and timing of cellular processes. Reports have shown that protein translocation activates cellular processes suggesting that compartmentalization is an important aspect of cellular regulation (Htun, Barsony et al. 1996) (Edgington and Futcher 2001) (O'Neill, Kaffman et al. 1996) (Andreadi, Noble et al. 2012) (Lau, Parisien et al. 2000). Activation of a protein upon translocation is fast and efficient and it provides the cell with an energy inexpensive method for controlling gene function, as opposed to protein degradation and regeneration. An example, pho4p is shuttled into and out of nucleus based on phosphate availability (Hung and Link 2011). Some steroid receptors are subject to a regulatory cycle involving conditional nucleo-cytoplasmic shuttling (Pratt 1992).

Regulation by compartmentalization is complex and forcing translocation in the absence of native stimuli will not necessarily activate function (Geda, Patury et al. 2008). Therefore a proteome-wide comparison to find differentially localized proteins could reveal new disease mechanisms related to compartmentalization.

Some proteins are known to change function when transported to new locations. For instance HMGB1, a chromatin protein gains cytokine function when it is transported out of the nucleus (Muller, Ronfani et al. 2004). Nuclear EGFR is observed in cancer and is associated with cell proliferation and drug resistance, and when it translocates to the nucleus EGFR is involved in DNA repair (Hung and Link 2011).

Some proteins have been found to mislocalize in disease states. Changes in protein subcellular location have been linked to functions that drive disease and protein mislocalization is known to occur in many diseases as diverse as Alzheimer's disease, kidney stones and cancer. In cardiac muscle cells protein mislocalization can be critical to the development of disease, specifically changes in location can alter degradation processes and lead to aberrant cellular activity (Lyon, Lange et al. 2013). Many other diseases and processes are also regulated by changes in protein location (Shelton, Chock et al. 2005) (Dai, Wei et al. 2007) (Hara, Agrawal et al. 2005) (Maulik, Engelman et al. 1999) (Maulik, Engelman et al. 1999) (Song and Lee 2003).

Location and cancer

In cancer reports show that the extent of localization in the nucleus can be used to predict patient prognosis. For example tissue microarray analysis of beta-catenin in colorectal cancer shows nuclear phospho-beta-catenin is associated with a better prognosis (Chung, Provost et al. 2001). The nuclear expression levels of NFκB in prostate lymph node metastases predict patient prognosis (Ismail, Lessard et al. 2004). Phospho-beta-catenin subcellular distribution in invasive breast carcinomas predicts phenotype and the clinical outcome of patients (Nakopoulou, Mylona et al. 2006). The cytoplasmic FOXO3a, p21, p27 and the nuclear EGFR are correlated with poor prognosis across many cancers and their mislocalization has been correlated with specific drug resistances (Hung and Link 2011).

The discovery of more proteins that undergo oncogenesis-associated changes in subcellular location could potentially improve disease diagnosis in conjunction with traditional protein expression markers. Further, discovering proteins that mislocate in the disease state may give new insight into changes driving disease. Such changes will go undetected by experiments measuring only expression levels.

Here we investigate location biomarkers, proteins that undergo changes in subcellular location that are indicative of disease. To discover such biomarkers, we have developed an automated pipeline to compare the subcellular location of proteins between sets of immunohistochemistry images. The pipeline also classifies and learns protein signatures to distinguish disease states.

Our lab previously described an automated system for recognizing major subcellular patterns in IHC images (Newberg and Murphy 2008). A set of 57 texture and nuclear overlap features at different levels of resolution were selected to distinguish eight subcellular location classes with high accuracy. Preliminary results to identify proteins that change location in various cancers were presented (Glory, Newberg et al. 2008, Newberg and Murphy 2008) however the performance on a larger collection of proteins with mixed patterns and pattern variation was found to be significantly lower compared to the 16 marker proteins used for training.

Statement of the problem

The subcellular location of a protein is an important property and changes in location are linked to changes in regulation and disease. One of the first steps in understanding subcellular location change is to identify proteins that relocate and under what conditions. An objective screening and measuring process would allow us to identify changing candidates.

Here we will specifically address how subcellular location can be measured from IHC images and we will perform a computational screening of location changes to identify location markers for cancer. We will extend this work to begin to understand how location changes can affect pathways and systems as a whole.

The initial hypothesis is that subcellular location changes occur across the proteome and that these changes can be quantified with computer vision methods. We

hypothesize that proteins do not change location independently but that parts of pathways and protein complexes translocate and are linked to changes in cellular regulation. Further, these location changes marker disease states and we can test the accuracy of using these proteins as markers.

Our second hypothesis was by measuring the differences in protein subcellular location and expression together we would be able to discriminate different disease states and find the smallest optimal set of proteins that can classify tissues.

Summary

Changes in the expression of proteins are often associated with oncogenesis, and are frequently used as cancer biomarkers. Changes in the subcellular localization of proteins have been less frequently investigated. We present for the first time a large-scale quantitative analysis of protein location across the proteome. The analysis pipeline uses state of the art computer vision methods to provide diagnostic applications and biological insight.

In chapter 2, we describe a robust pipeline for identifying proteins whose subcellular location undergoes statistically significant changes in cancers of four tissues: breast, liver, prostate and bladder. We used the pipeline to compare images of healthy and tumor tissue from the Human Protein Atlas, ranking hundreds of proteins in breast, liver, prostate and bladder based on how much their location is estimated to have changed. The performance of the system was evaluated by determining whether proteins previously known to change location in tumors were ranked highly. We present a number of candidate location biomarkers for each tissue, for some of which have been associated with cancer, and biochemical pathways enriched for proteins that translocate. The analysis technology is anticipated to be useful for discovering new location biomarkers and also for enabling automated analysis of biomarker distributions as an aid to determining diagnosis and prognosis.

Chapter 3 presents a study on liver lesions. Hepatic lesions range in severity from benign to malignant where the most malignant cases are fatal. Premalignant and malignant liver lesions usually require a biopsy however diagnosis of the tissue can be challenging because different lesions can have similar morphological appearances (Isaacs 2007), however the best treatment for each lesion can be very different (Isaacs 2007) (Litten and Tomlinson 2008). The need to improve liver lesion characterization is immediate. Here we focus on classifying lesions in two age groups, adult and pediatric. The pediatric analysis group consists of three liver lesions: normal liver (nl), fetal hepatoblastoma (FHB) and well-differentiated hepatocellular carcinoma (WDHCC). The adult analysis group consists of 5 lesions: dysplastic nodules (DN), focal nodular hyperplasia (FNH), hepatocellular adenoma (HCA), hepatocellular carcinoma (HCC), macroregenerative nodules (MRN). In this chapter we extend our pipeline to process full slide images and we apply it to quantitate protein subcellular location and expression in images. We construct a series of classifiers to learn the optimal signature of proteins necessary to discriminate lesion types. We show that some lesions can be distinguished with proteins that are currently not used as markers in the clinic.

Chapter 2

Automated Analysis of Immunohistochemistry Images Identifies Candidate Location Biomarkers for Cancers

This chapter describes joint work with Aparna Kumar, Arvind Rao, Santosh Bhavani, Justin Y. Newberg, Robert F. Murphy and is modified from the published manuscript entitled "Automated Analysis of Immunohistochemistry Images Identifies Candidate Location Biomarkers for Cancers" (Kumar, Rao et al. 2014). This chapter refers to supplemental sections or datasets of the online paper.

Abstract

Molecular biomarkers are changes measured in biological samples that reflect disease states. Such markers can help clinicians identify types of cancer or stages of progression, and they can guide in tailoring specific therapies. Many efforts to identify biomarkers consider genes that mutate between normal and cancerous tissues or changes in protein or RNA expression levels. Here we define *location biomarkers,* proteins that undergo changes in subcellular location that are indicative of disease. To discover such biomarkers, we have developed an automated pipeline to compare the subcellular location of proteins between two sets of immunohistochemistry images. We used the pipeline to compare images of healthy and tumor tissue from the Human Protein Atlas, ranking hundreds of proteins in breast, liver, prostate and bladder based on how much their location was estimated to have changed. The performance of the system was evaluated by determining whether proteins previously known to change location in tumors were ranked highly. We present a number of candidate location biomarkers for each tissue, and identify biochemical pathways that are enriched in proteins that change location. The analysis technology is anticipated to be useful not only for discovering new location biomarkers but also for enabling automated analysis of biomarker distributions as an aid to determining diagnosis.

Significance Statement

Changes in the expression of proteins are often associated with oncogenesis, and are frequently used as cancer biomarkers. Changes in the subcellular localization of proteins have been less frequently investigated. In this paper, we describe a robust pipeline for identifying those proteins whose subcellular location undergoes statistically significant changes in cancers of four tissues, and also for identifying biochemical pathways that are enriched for proteins that translocate. Future investigation of these proteins and pathways may provide new insight into oncogenesis. Further, the analysis pipeline is expected to be useful for assessing disease type and severity in a clinical setting.

Introduction

Our understanding of the number and types of changes that occur in various cancers is continuously growing. Previous work to discover proteins that vary significantly between normal and cancer cells has used techniques such as microarray profiling, next generation sequencing, antibody arrays and proteomic profiling (Kononen, Bubendorf et al. 1998, Khan, Saal et al. 1999, Mardis and Wilson 2009, Leung, Diamandis et al. 2012). These studies have led to the discovery of proteins (termed expression biomarkers) whose expression levels mark different disease states. However for some proteins, the extent of localization in the nucleus can be used to predict patient prognosis; β-catenin (Chung, Provost et al. 2001) and NFκB (Lessard, Karakiewicz et al. 2006) are examples. The discovery of more proteins that undergo oncogenesis-associated changes in subcellular location (which we term location biomarkers) could potentially improve disease diagnosis in conjunction with traditional protein expression markers. Further, discovering proteins that relocate in the disease state may give new insight into changes driving disease, and that changes would go undetected by measuring only expression.

Immunohistochemistry (IHC) studies are a major source of data on protein expression and location. Most such studies use visual examination to assess changes, a difficult and time-consuming task. With the advent of high throughput acquisition technologies like tissue microarrays and automated slide scanners, computerized analysis of tissue images is highly desirable and studies have shown that quantitative software can detect changes in disease states that are missed by visual inspection (Guillaud, Adler-Storthz et al. 2005). Methods for analyzing changes in expression and pattern are well established in cultured cells (Shariff, Kangas et al. 2010) but histological images are typically more difficult to analyze because cellular heterogeneity and the closely packed organization of cells lead to significant cell segmentation challenges. Several projects have been initiated to build workflows that process IHC images (Lejeune, Jaen et al. 2008, Matos, Trufelli et al. 2010). Most of this work has been focused on quantitating differences in protein abundance between normal and cancer tissue. However, as discussed above, differences in subcellular protein locations could also be critical both for understanding and diagnosing disease. Thus there is a strong need for systems that can analyze protein subcellular location in IHC images.

We have previously described an automated system for recognizing major subcellular patterns in IHC images (Newberg and Murphy 2008), and presented preliminary results on using that system to identify proteins that change location in various cancers (Glory, Newberg et al. 2008). These studies used a subset of the extensive collection of IHC images in the Human Protein Atlas (HPA) (Uhlen, Bjorling et al. 2005). However, we have found that the performance on a larger collection of proteins with more pattern variation was significantly lower compared to the 16 marker proteins used in our previous study. We therefore sought to develop a system that can identify potential location biomarkers using new approaches without explicit classification. Using images from the HPA, we show that our system can identify proteins with altered subcellular location directly from tissue images and anticipate that approaches such as this may significantly contribute to diagnosis, treatment and monitoring of cancers.

Results

Our analysis pipeline (Fig. 1) consists of five steps.

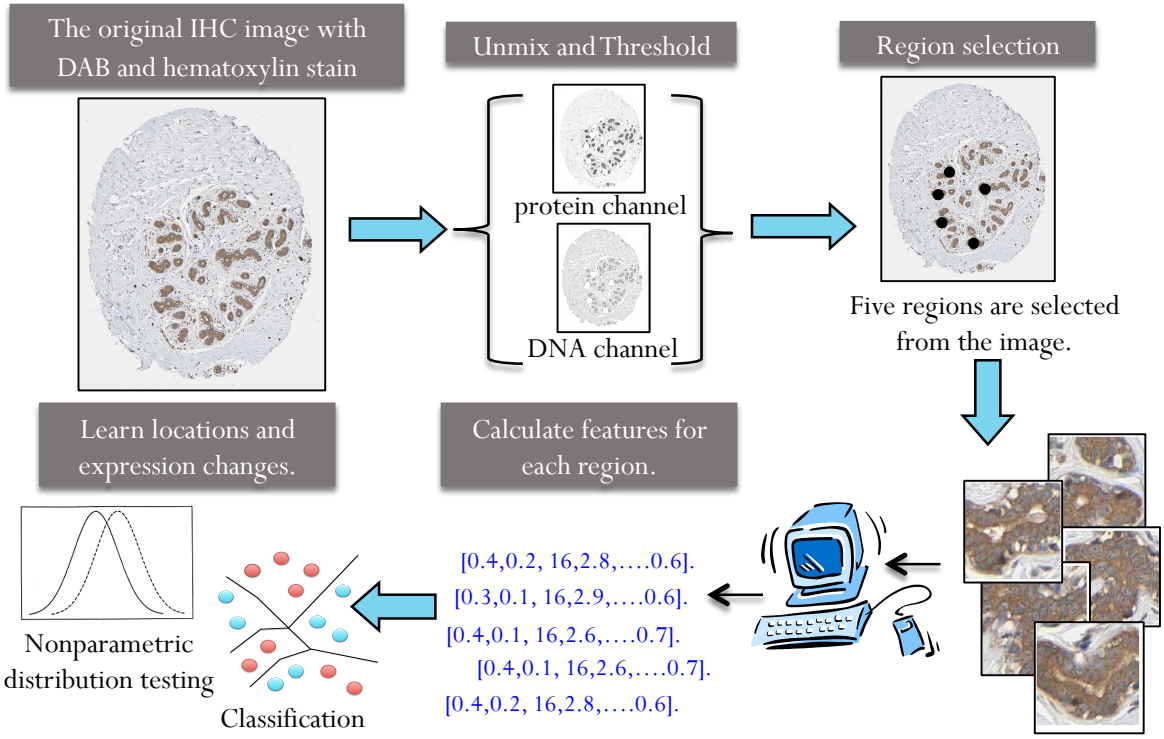


Fig. 1. Overview of the location biomarker discovery pipeline. Images with strong or moderate antibody staining were selected. Linear unmixing was used to separate each image into two composite images representing the DNA and protein stains as previously described (11). Regions were selected by convolving the protein image with a low pass filter and selecting the highest points as region centers. Fifty-seven numerical features were calculated to describe the pattern in each region. The non-parametric FR test was used to calculate a p-value and determine whether the null hypothesis, that the features from the normal and cancer image come from the same distribution, should be rejected. The non-parametric WW test was used to calculate a p-value to measure how likely the two sets of images are to come from the same expression distribution. A nearest neighbor classifier was also used to determine the ability of each antibody to distinguish normal and cancer images.

(i) Selecting a set of proteins for analysis guided by staining levels. For a given tissue, we selected antibodies from the HPA whose staining intensity was annotated as moderate or strong, and whose staining quantity was annotated as greater than 75%. Due to tissue specific expression and variations in staining, the proteins identified (referred to as the analysis set) were different for each tissue.

(ii) Separating the DNA and protein components of each image by unmixing the hematoxylin and diamino-benzidine stains. The HPA images are collected as RGB images in which the two stains appear as purple and brown, respectively. The intensity derived from each stain is therefore a combination of the intensities from the three RGB channels. We unmixed the spectra to give separate images reflecting mainly DNA and protein content (Newberg and Murphy 2008).

(iii) Selecting regions of each image with the highest protein expression, under the assumption that the highest stained regions would be less likely to contain connective tissue, stroma and other non-cellular regions.

iv) Calculating features to describe the localization patterns in each region (Newberg and Murphy 2008).

(v) First, estimating the probability that a given protein's localization pattern differs between the two conditions. The nonparametric Friedman-Rafsky (FR) test was used to calculate a p-value for the null hypothesis that the sets of regions from normal images and from cancer images show the same pattern. Second, estimating the probability that a given protein's level of expression differs between the two conditions. Expression p-values were calculated using the Wald Wolfowitz method to test the null hypothesis that the level of expression in the regions from the normal and cancer images came from the same distribution. These calculations were done for 35 random samplings of images, giving very high repeatability of the

results (see Methods). Finally, calculating a classification accuracy for separating normal and cancer images by using protein location information.

We applied this pipeline to images from the HPA for four tissues: breast, liver, prostate, and bladder (the results are contained in Dataset S1). After running the pipeline for each tissue, the proteins were sorted by their location p-values to obtain a ranking by extent of subcellular location change. Representative images of the top three hits for each tissue are shown in Fig. 2.

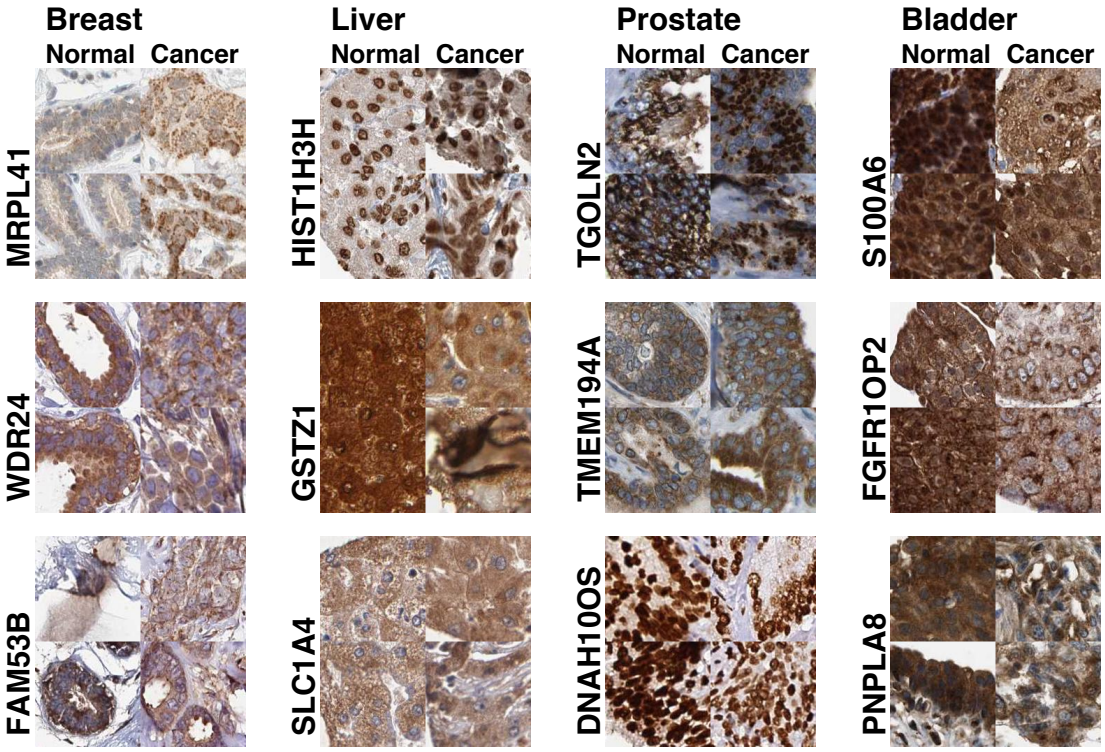


Fig. 2. Example images from top ranked potential location biomarkers. The three proteins with the lowest location p-values are shown for each tissue (without considering expression level). The two regions closest to the two centroids found from k-means clustering (k=2) for the normal and cancer feature distributions are

displayed for each of the top hits. Note that some of the top hits may have been detected due to expression changes.

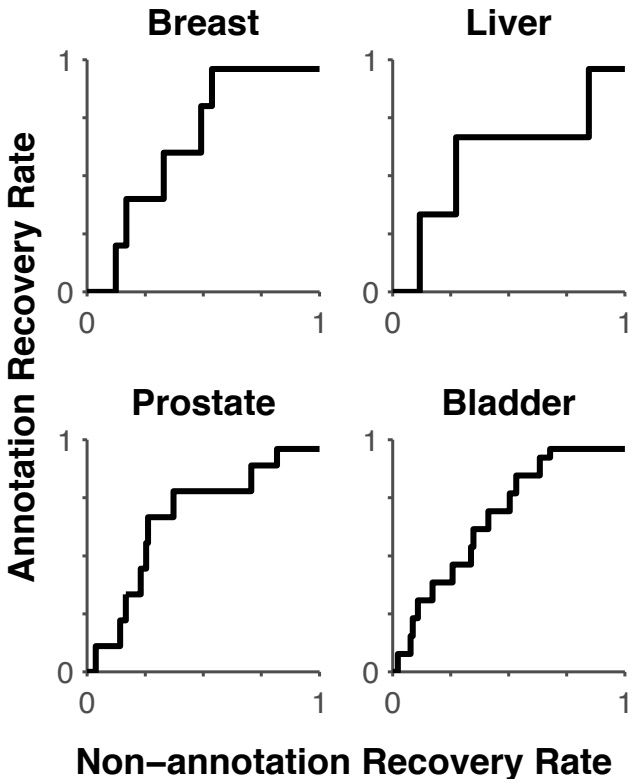


Fig. 3. Ability of the system to detect known location biomarkers. ROC curves were constructed for each tissue by determining how many true positives and false positives were found as a threshold on the p-value was varied. The validation set for a given tissue consisted of those proteins from the analysis set that were annotated as having a different location between the normal and cancer images. Note that some of the false positives may actually be positives that were not present in the validation set.

Testing using known location biomarkers

We expected that proteins known to change location in cancer would be ranked high on this list. To test this, we constructed validation sets using pathologists' annotations of the gross subcellular location provided in HPA: 1) nuclear, 2) cytoplasm and plasma membrane, 3) nuclear, cytoplasm and plasma membrane, 4) none. The validation set for a given tissue consisted of those proteins from the analysis set for that tissue that had different location annotations between the normal and cancer images (see Materials and Methods). Treating these as true positives, we constructed receiver-operating characteristic (ROC) curves in which a threshold on the p-value at which a protein was considered positive was varied (Fig. 3). The area under these curves is a measure of how well our test finds the true positives. If the validation markers were the only proteins expected to change location, and if the system performed perfectly, the area under these curves should be one. However, we expect some of the proteins ranked highly by p-value may be actual location biomarkers even if they are not in the validation set. For example, proteins may undergo a change in location that was not captured by the gross location annotations used to define true positives. Thus we do not expect even a very good discovery system to give values near one. The areas under the ROC curves for breast, liver, prostate and bladder were 0.67, 0.59, 0.67, 0.68, respectively. These are all significantly above 0.5, the area expected for random performance.

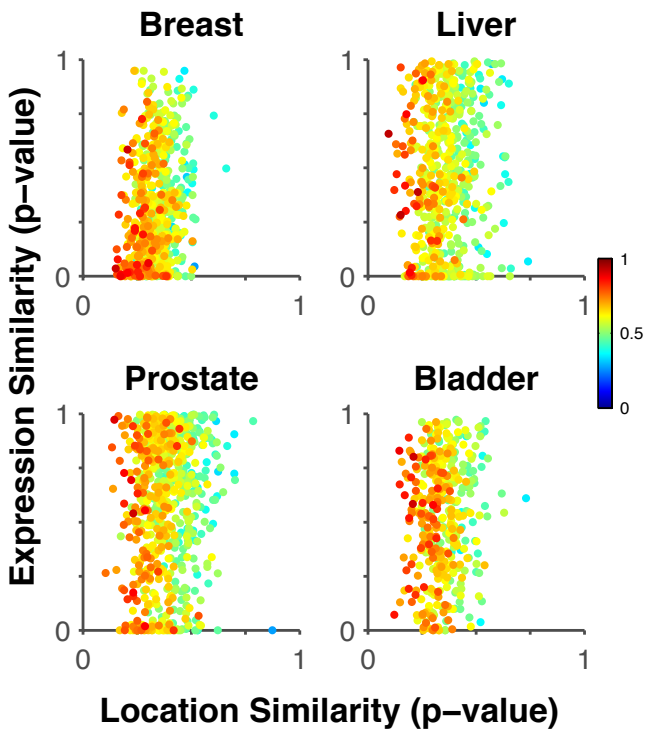


Fig. 4. Distinguishing between intensity and location changes. Each dot shows the p-values for the hypotheses that location or expression are different between normal and tumor tissue for a given protein. The correlation between location and expression p-values is weak suggesting that proteins that change location in the cancer state do not necessarily change expression, as seen in the top left corner. The color indicates the classification accuracy for separating normal and cancer images of that protein using subcellular location information alone. Proteins with high classification accuracy for distinguishing normal and cancer images are represented by a red dot. Red proteins closest to the top left corner are potential location biomarkers and their discovery would have been missed by traditional experiments that measure changes in protein expression.

Distinguishing location and expression changes

The features we used are designed to minimize the effect of differences in protein staining level. Even so, a major change in expression may cause a change in image texture that would be detected by our features even if subcellular location remains the same. This may cause proteins that do not change their location significantly but do change their expression dramatically to rank highly on our lists. We therefore used the expression p-values and location p-values together to analyze each protein's change.

Fig. 4 shows the relationship between the expression change and location change for proteins in various tissues. The first conclusion we can draw is that the two values are not correlated, suggesting that proteins that change location do not always change expression, and vice versa. Secondly, the points in the upper left corner of each scatter plot represent proteins that have significantly changed location (low p-values) but have not changed expression (high p-values). The color of each point indicates how well that protein can be used to train a classifier to distinguish images from normal and cancerous tissue (see Materials and Methods; the accuracy values are listed in Dataset S1). Thus we expect proteins whose symbols are dark red and in the upper left corner to be potential biomarkers useful in a clinical setting for recognizing cancerous tissue by measuring differences in subcellular location. These proteins would not have been identified as potential markers by measuring expression changes alone. Dataset S1 is ranked for each tissue using the Euclidian distance from the upper left corner, that is, proteins that change location and do not change expression. The five top-ranked proteins for each tissue using this criterion are shown in Table 1, and images of the top three from each tissue are shown in Fig. 5.

Of course, we expected that classic biomarkers that are known to translocate in cancer, such as E-cadherin, B-catenin and NFKB, would be ranked highly in this list. These proteins were not part of our analysis sets because the HPA did not contain a

sufficient number of images to meet the threshold of our pipeline. We therefore separately calculated location p-values for those proteins using the images that were available for breast and prostate cancers. The p-values for two E-cadherin antibodies with high reliability, CAB000087 and HPA004812 were higher than 0.20. The p-values for three B-catenin antibodies, CAB000108, HPA029159 and HPA029160 were higher than 0.32. The two antibodies against NFKB in prostate cancer are CAB004031 and HPA027305 with p-values greater than 0.22. Thus our tests indicate that none of these are strong location biomarkers in these tissues, contrary to expectation based on previous literature reports.

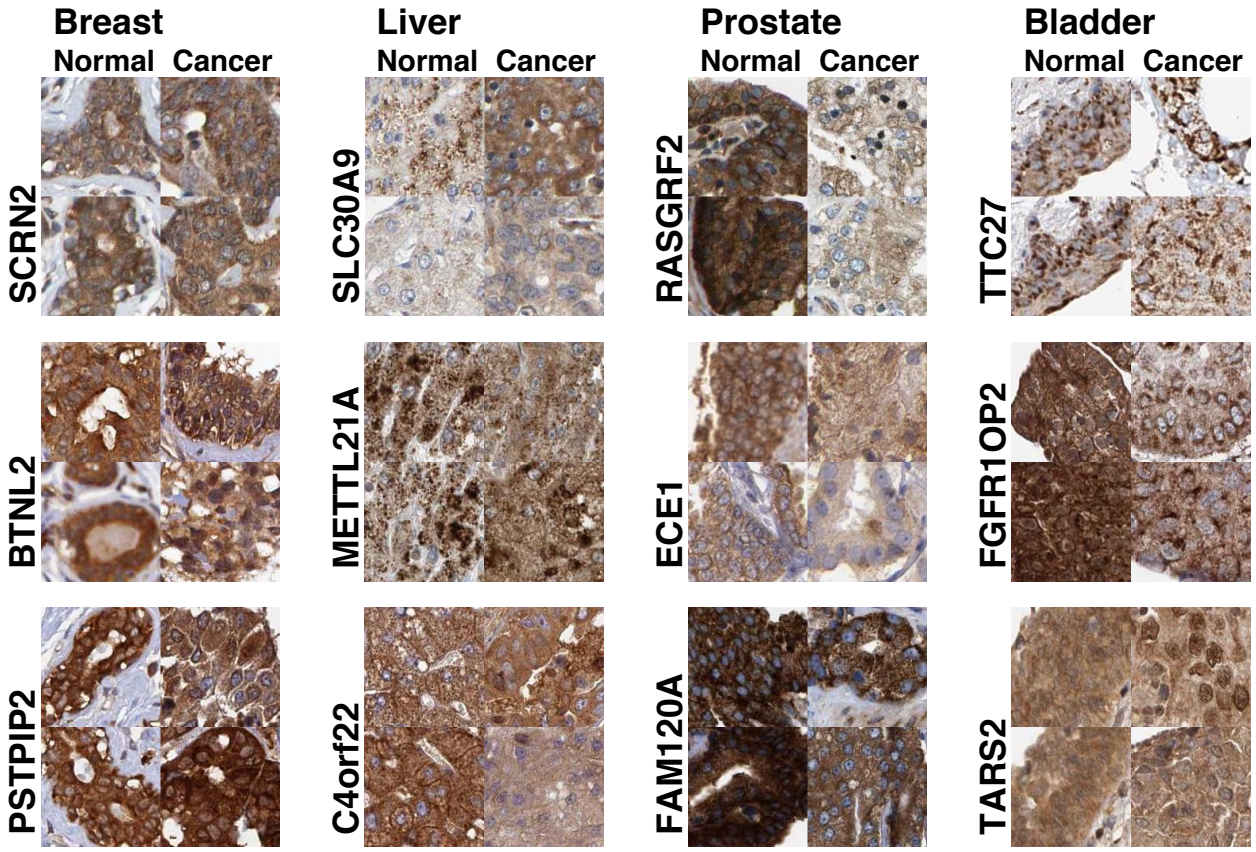


Fig. 5. Example images from top location biomarker predictions with very small mean intensity changes. For every protein the features from each disease state were clustered using k-means (k=2) and the region closest to each centroid is displayed.

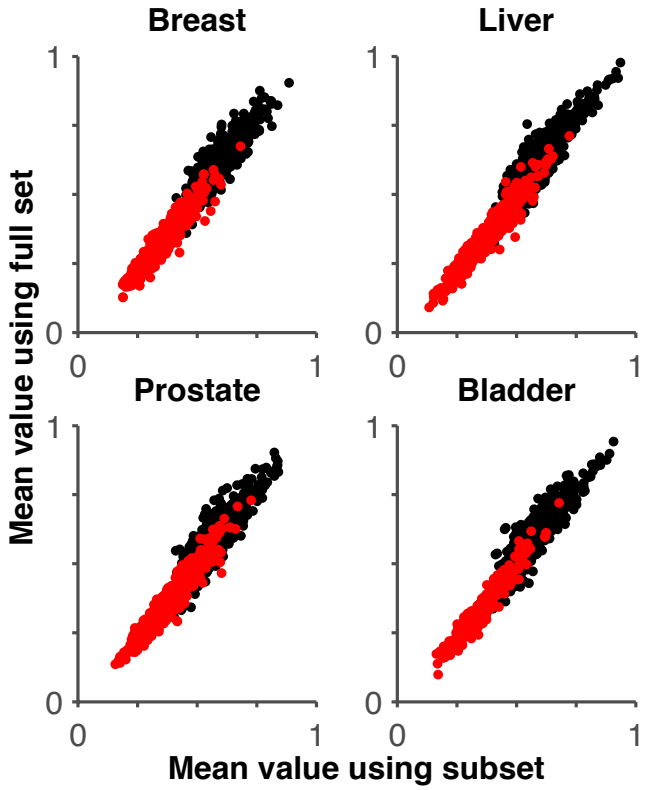


Fig. 6. Estimation of generalizability of identifying location biomarkers. For each protein, classification accuracies (black) and location biomarker rankings (red) are compared for estimates using one or two normal images. The correlation coefficients were 0.90, 0.91, 0.90 0.90 for the accuracies and 0.95, 0.96, 0.96, 0.96 for the location biomarker rankings. The high correlations for the rankings suggest that highly ranked proteins would also be highly ranked in new images.

Gene Name	HPA Ab	Loc pvalue	Exp pvalue.	Acc.
Breast				
SCRN2	HPA023434	0.24	0.95	0.69
BTNL2	HPA039844	0.24	0.95	0.59
PSTPIP2	HPA040944	0.28	0.95	0.58
USP10	HPA006749	0.27	0.90	0.58
NT5DC3	HPA041634	0.27	0.89	0.62
Liver				
SLC30A9	HPA004014	0.15	0.96	0.81
METTL21A	HPA034712	0.15	0.87	0.65
C4orf22	HPA043383	0.19	0.92	0.65
WDR24	HPA039506	0.16	0.85	0.67
PARP12	HPA003584	0.22	0.94	0.78
Prostate				
RASGRF2	HPA018679	0.14	0.97	0.90
ECE1	HPA001490	0.17	0.99	0.77
FAM120A	HPA019734	0.18	0.94	0.73
PLA2G4C	HPA043083	0.19	0.95	0.69
TMEM194A	HPA014394	0.13	0.85	0.79
Bladder				
TTC27	HPA031246	0.19	0.89	0.84
FGFR1OP2	HPA038696	0.14	0.83	0.89
TARS2	HPA028626	0.25	0.96	0.54
STAC	HPA035143	0.19	0.83	0.71
-	CAB009119	0.20	0.82	0.72

Table 1. Potential location biomarkers. The five proteins with the greatest location change and the smallest expression change are shown (the full ranked list is in Dataset S1). Classification accuracies for distinguishing normal and cancer are also shown.

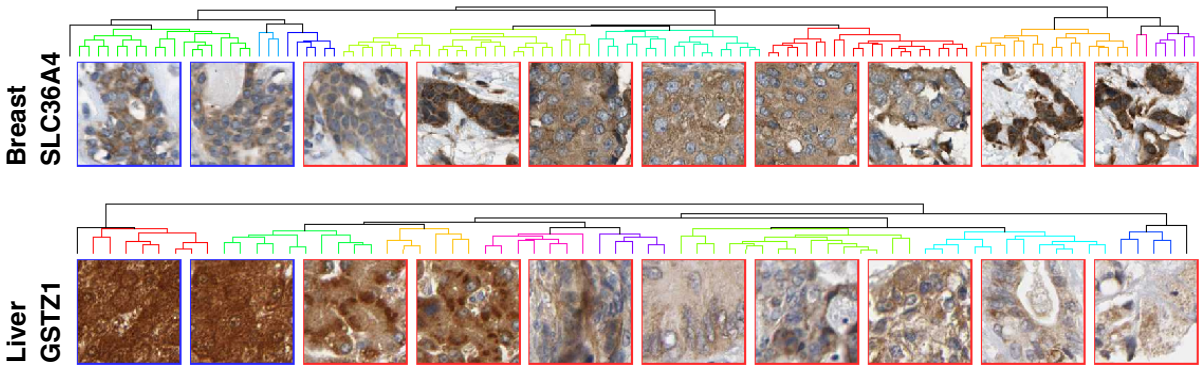


Fig. 7. Ordering regions by location change progression. We selected one top ranking protein from breast and liver: antibodies HPA017887 and HPA004701 respectively. The Euclidean distances between every pair of regions were calculated using the features and clustered into a binary hierarchical tree. The leaves were ordered to maximize the sum of similarities between adjacent leaves across the tree. The tree was 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 from left to right. Normal tiles are outlined in blue; cancer tiles are outlined in red.

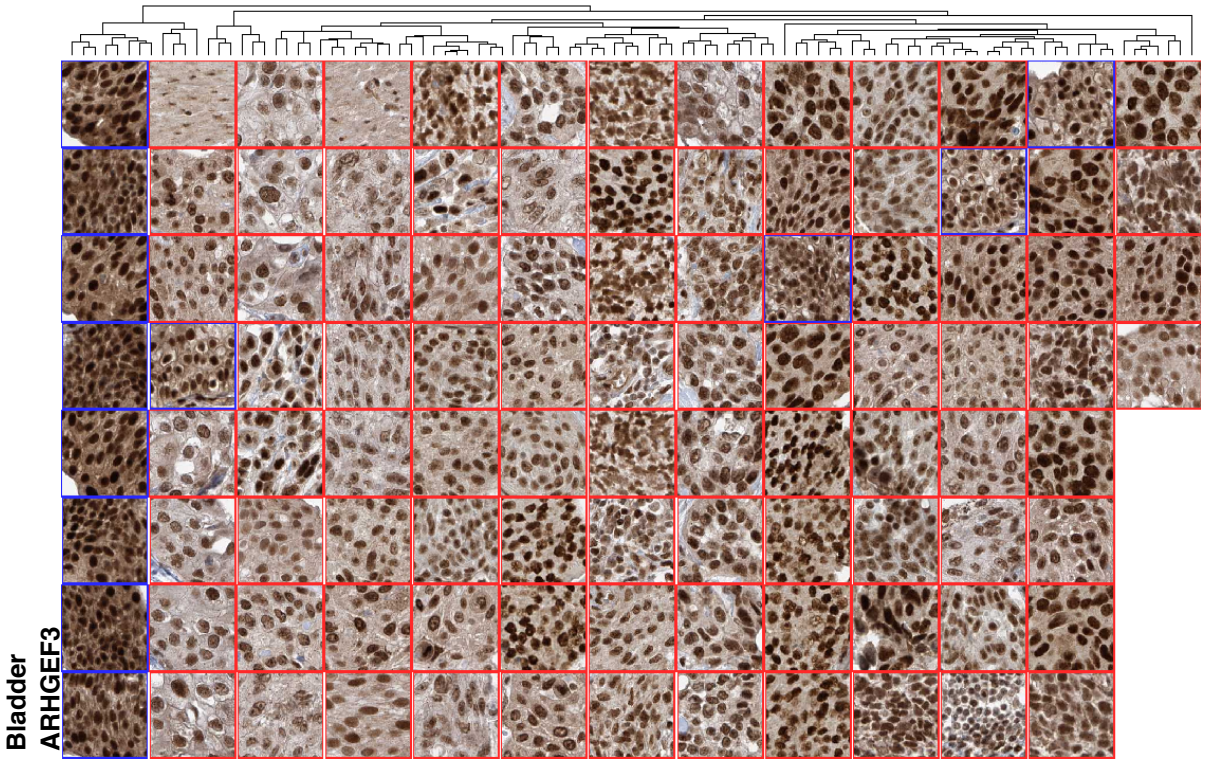


Fig. 8. Ordering regions by location change progression. We performed a hierarchical clustering of the features from the regions used in the analysis for antibody HPA034715 against ARHGEF3 in bladder tissue. We calculated the Euclidean distances between every pair of regions and then performed the binary hierarchal clustering. The leaves are ordered to maximize the sum of similarities between adjacent leaves across the tree. Regions are displayed vertically according to the ordering in the tree. Normal tiles are outlined in blue; cancer tiles are outlined in red.

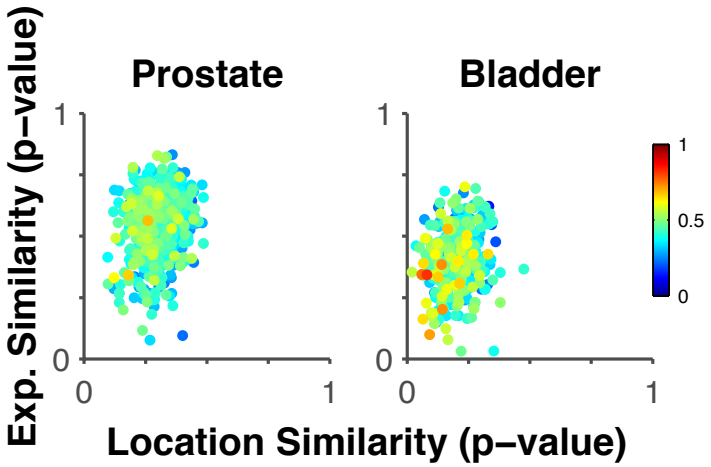


Fig. 9. Location biomarkers differ between high and low grade cancers. The prostate and bladder cancer images were partitioned into high or low grade cancer as annotated in the Protein Atlas. For each cancer, location and expression p-values were calculated between the grades. The correlation between location and expression p-values is weak suggesting that proteins with different locations between the two grades will not necessarily have different expression levels. The color of each dot indicates the accuracy of a three-class classifier trained to distinguish the normal and the two grades while using location information alone. Some proteins (marked in orange) have high classification accuracies, further they showed a significant location change and do not show a significant change in expression. These proteins are potential location biomarkers for the cancer subtypes.

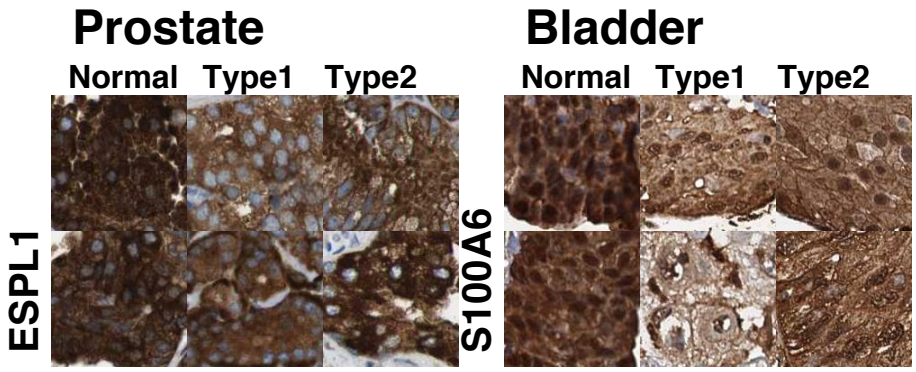


Fig. 10. Example images from top location biomarker predictions for classifying normal tissue from different tumor grades. The proteins are ranked by the 3-class classification accuracy for separating normal tissue, low grade tumors and high grade tumors.

In addition, upon visual inspection of the HPA images, we did not observe a pattern change; the pathologist annotations also did not indicate a location change. All of the antibodies for these three proteins had identical location annotations in the two disease states with the exception of HPA004812 which moved from nuclear and cytoplasmic and membranous to mostly cytoplasmic membranous in the cancer state. The basis for this difference in this dataset from previous results is unclear.

Given that we were evaluating a large number of antibodies, we were also interested in estimating the generalizability of the relative performance of the antibodies. In other words, how likely it is that proteins with low p-values or high classification accuracies would show similar values in future experiments? To do this, we calculated p-values and accuracies for a smaller number of images and compared it to the values with the larger number (see Methods). Note that this is different than the repeatability of the rankings using the same number of images. The results shown in Fig. 6 indicate a high correlation in the two estimates of the rankings of the p-values, and a slightly lower correlation of the two estimates of the rankings of the classification accuracies. This suggests that the generalizability of p-value and

classification rankings should be high (the generalizability to unseen images should be even higher for the full set than for the subsets).

Features reveal visually distinguishable changes useful for distinguishing tumor from healthy tissue

To determine whether the differences in location being identified by our pipeline were visually distinguishable changes, we performed hierarchical clustering and optimal leaf ordering to order the regions for a given antibody using our features (see Methods). Fig. 7 shows ten representative regions ordered for two example proteins from the top ranked proteins. In breast tissue stained for SLC36A4, the normal regions clustered near each other on the left. Further, the regions appear ordered by increasing nuclear localization suggesting our features can detect incremental and possibly continuous changes in this location pattern. In liver, GSTZ1 showed a decrease in nuclear localization from left to right, and also an increase in cytoplasmic graininess. The clustering grouped the normal and cancer regions separately

Location biomarkers can distinguish between cancer grades

Each cancer in the HPA has a specified grade or subtype. We partitioned the images by grade and ran the pipeline to compare the two grades to each other for the prostate and bladder cancer set (Dataset S2). We also asked how well each protein could be used as a potential biomarker in a classifier trained to distinguish three disease states: normal tissue, low grade and high grade tumors. Fig. 9 shows the location p-value and the expression p-value for each protein when comparing the two subtypes to each other. Points that fall in the upper left corner have different subcellular locations between the two grades but similar expression levels. The color of each point represents that protein's 3-class classification accuracy.

The points with warmer colors and close to the top left corner represent proteins that have different locations but similar expression levels in the two subtypes. These proteins can be used to distinguish the three disease states. Example images for the proteins with the greatest 3-class classification accuracies are shown in Fig. 10. The best classification accuracy was obtained for S100A6 in bladder: it has a classification accuracy of 83% (compared to 33% expected at random), a location p-value of 0.081 and an expression p-value of 0.34. This protein is the best example of a potential location biomarker (one that changes location but not expression) in bladder. These results provide further support for the utility of our system for identifying important location changes between disease states.

KEGG pathways and translocated proteins

Lastly, we were interested to find out whether our analysis could suggest entire pathways, or major portions thereof, that might undergo translocation together in cancer (the simplest example would be proteins that are part of a translocating complex). To answer this, for each KEGG pathway we calculated the probability that all of the proteins in it changed location or expression compared to a randomly sampled background distribution (Fig. 11). (Note that this represents an underestimate of the change in a pathway if it contains subcomponents that do not translocate.) We calculated pathway changes using either our image processing pipeline or pathologist annotations. Pathways with the largest change in either location or expression are listed in Table 2. As discussed below, some of these pathways have been previously implicated in cancer and some are novel predictions.

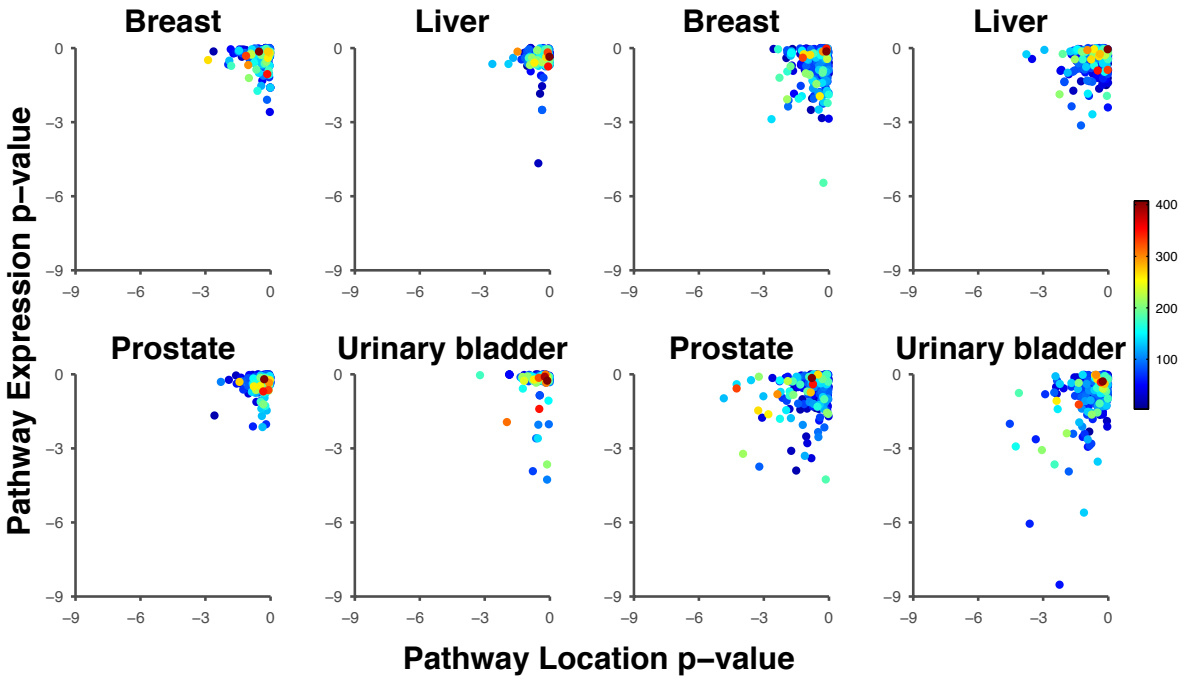


Fig. 11. Extent of expression and location change in the KEGG pathway components. The left four panels show KEGG p-values using pipeline protein values, the right four panels show pathway p-values derived from location and expression annotations. Each point represents the expression and location p-value for a single pathway. The points are colored by the number of nodes in the pathway.

Discussion

We have described a workflow to identify proteins that change their subcellular location between normal and cancerous tissue without requiring classification. We validated our ability to detect changes in location classes by using annotations provided by the HPA database.

Upon visual inspection of top hits (Fig. 5), we noted that our system was able to find texture changes between the two disease states, however these did not always represent changes between distinct subcellular location classes. In some cases our texture features were detecting changes in tissue structures and morphology.

In Fig. 5, BTNL2 in breast showed a decrease in nuclear localization. SLC30A9 in liver decreased in cytoplasmic graininess, and C4orf22 decreased in plasma membrane and vesicle localization in the cancer state. In prostate cancer ECE1 decreased in plasma membrane accumulation. In bladder cancer FGFR1OP2 changed from cytoplasmic and nuclear localization to mostly a grainy cytoplasmic localization in cancer. TARS2 in bladder increased in nuclear membrane localization in cancer. In some cases our texture features picked up changes in tissue structure but not necessarily subcellular location, as was seen in SCRN2 in breast and TTC27 in bladder.

It was of interest to consider whether our top predictions had previously been implicated as being altered in cancers, and which ones were new discoveries. In breast cancer very few studies have reported on BTNL2, however there is strong evidence that variants of this gene play a role in susceptibility to sporadic and familial prostate cancer (Fitzgerald, Kumar et al. 2013). PSTPIP2 has not been reported in breast cancer however it is implicated in the expansion of macrophage progenitors leading to autoinflammatory disease (Chitu, Ferguson et al. 2009). USP10 is translocated to the nucleus upon DNA damage and regulates p53 (Yuan, Luo et al. 2010).

In liver cancer PARP12 has been reported to play a role in genome surveillance, DNA repair pathways and it is rising as a new potential therapeutic target (Yelamos, Farres et al. 2011).

In prostate cancer, four of our top findings have been linked to prostate cancer development. Methylation of the RASGRF2 gene was found to be associated with prostate cancer (Mahapatra, Klee et al.). ECE1 has been implicated in prostate cancer cell invasion, where different isoforms of the protein were found to play different roles (Lambert, Whyteside et al. 2008). PLA2G4C is regulated by EGR, a gene that is rearranged in about 50% of prostate cancer (Massoner, Kugler et al.

2013). In bladder cancer very few of the top findings have been published in association with disease.

Pathway changes

Some of the top ranking pathways had location p-values that were about two orders of magnitude smaller than the expression p-values based on the pipeline results (Table 2 and Fig. 9). In the HTLV-1 infection pathway, the HTLV-I Tax oncoprotein initiates malignancy development in leukemia by creating an environment to facilitate DNA damage (Matsuoka and Jeang 2007). The molecular mechanism of this pathway has not been studied in the contexts of breast or urothelial cancer. Our results together with other literature reports indicate that subcellular location changes in components of HTLV-I infection pathway could play a role in driving cancer. Further, the high rank of this pathway across the four tissues indicates that these changes may be important in identifying and understanding other cancers as well.

The 'one carbon pool by folate' pathway was also found to change location in breast cancer. It is known to play an important role in DNA global hypomethylation, which can lead to DNA strand breaks (Xu and Chen 2009). As expected, when changes in expression are used to rank pathways, the ErbB pathway ranks near the top for breast cancer (Howe and Brown 2011).

In liver cancer the axon guidance pathway was found to change location more than it did expression. One of the genes contributing to the axon guidance pathway, ROBO1 was found to be overexpressed in HCC (Ito, Funahashi et al. 2006). The axon guidance pathway has not been implicated as a whole in liver cancers but it is known to be altered in pancreatic cancers (Biankin, Waddell et al. 2012). Our results with previous reports of ROBO1 suggest this pathway may play a role in liver cancer. The importance of the top pathways seen to change expression in liver is unclear.

In prostate cancer HIF-1 signaling is known to play an important role in hypoxia adaption of tumors and HIF-1alpha is known to be overexpressed in early tumors (Kimbro and Simons 2006). Our findings suggest that the proteins in this pathway undergo location changes, possibly contributing to the pathway's dysregulation in cancer. PPAR is a known prostate cancer marker [Collett, 2000 #3134], and its pathway is identified as changing expression.

Lastly, in urothelial cancer the hippo signaling pathway was the top ranking pathway to change location. Hippo signaling is responsible for tissue size and is known to lead to uncontrolled cellular proliferation and blocking of apoptosis when misregulated (Barron and Kagey 2014). When we ranked the pathways in urothelial cancer by expression changes, a number of signaling pathways known to be involved in cancers rank at the top.

We also calculated the product of the p-values for each pathway across all four tissues to find those pathways changing in all four cancers (Dataset S3). Three of the top ranking pathways for location changes were already identified in individual tissues. In addition, the p53 signaling pathway (which is known to involve location changes), was also identified. For expression changes, five pathways previously associated with cancers were highly ranked (which is encouraging with respect to the accuracy of our automated methods).

Our analyses suggest that location changes of these pathways may be important for understanding their role in disease. In addition our results link previously implicated pathways to new cancers for further investigation.

Tissue	L:P	E:P	L:A	E:A	Pathway
Breast	+	-	-	-	HTLV I infection
	+	-	-	-	One carbon pool by folate
	-	+	-	-	Proteasome
	-	+	-	-	ErbB signaling pathway
Liver	+	-	-	-	Axon guidance
	-	+	-	-	Glycosylphosphatidylinositol GPI anchor biosynthesis
	-	+	-	-	Hypertrophic cardiomyopathy HCM
	-	+	-	-	Dilated cardiomyopathy
Prostate	+	-	-	-	Fatty acid elongation
	+	-	-	-	HIF 1 signaling pathway
	-	+	-	-	Oxidative phosphorylation
	-	+	-	-	PPAR signaling pathway
	-	+	-	-	Viral myocarditis
Bladder	+	-	+	-	Hippo signaling pathway
	-	+	-	-	NF kappa B signaling pathway
	-	+	-	-	p53 signaling pathway
	-	+	-	-	Transcriptional misregulation in cancer
	-	+	-	-	Apoptosis
	-	+	-	-	Cell cycle
	-	+	-	-	mRNA surveillance pathway
	-	+	-	-	Ribosome biogenesis in eukaryotes

Table 2 – Pathways with the largest location or expression changes. Pathway p-values were calculated using individual protein location (L) or expression (E) p-values from the pipeline (L:P,E:P) and using p-values from pathologist annotations (L:A, E:A). + indicates pathway p-values less than 0.01. The values for all pathways are in Dataset S3.

Conclusion

Using staining patterns of proteins in four tissues, we have identified proteins that show altered subcellular location in cancer and/or whose patterns can be used to distinguish normal and cancerous tissue or different cancer subtypes. Many of these proteins do not have significant expression level changes and would not have been found as biomarkers if we had considered expression level changes alone. Further, some proteins have high classification accuracies but visually similar location patterns. The subtle changes that are being detected may nonetheless be useful for distinguishing disease states. Extended analysis with more images of the potential markers we have identified will be necessary to assess their utility or significance. We note that the analysis pipeline we have described is not only useful for identifying cancer biomarkers, but should also be valuable for automating the process of analyzing IHC images to assess disease state. We are currently carrying out collaborative translational studies to determine whether our technology combined with any of the potential biomarkers is useful for distinguishing lesions with various diagnoses or prognoses.

Materials and Methods

Data

We used images from the HPA (www.proteinatlas.org) that appeared online on September 24, 2013. Proteins were placed in the analysis set for each tissue if they met three criteria: (i) the staining annotation was strong or moderate, (ii) if the quantity field was annotated as greater than 75%, (iii) at least three images of that protein were available for the normal tissue. Approximately 500 proteins per tissue passed this filtering procedure (see Dataset S1 for the list of proteins in the full analysis set for each tissue).

Identification of validation sets

A validation set of proteins whose location was known to change (which we define as true positives) was created for each tissue. These were found using HPA annotations. We identified the set of true positives for a given tissue by finding those proteins for which the set of location annotations for all normal images did not intersect the set of location annotations for all cancer images. In the data set for breast, liver, prostate and bladder there were 5, 3, 7, 13 true positives, respectively.

Selecting regions

For each image, we selected regions that showed significant staining. First a low pass filter was applied to the protein mask of each image. We selected regions centered on the peaks of the filtered image. This was done under the assumption that the cellular regions of the tissue would have the highest staining levels, as opposed to the connective tissue, stroma and other non-cellular regions which would have much lower levels of staining primarily due to non-specific antibody binding.

Removing outlier images

Next we removed outlier regions and images based on DNA and protein intensity. For each tissue we calculated the mean and standard deviation of the protein and DNA stains for all images. This same process was repeated for all of the regions from each tissue. We removed images and regions from the dataset that were farther than 4 standard deviations from the mean.

Pipeline for testing changes in location or expression

Our pipeline calculates p-values for the hypotheses that the location or expression of each protein are the same between normal and cancer images. The pipeline requires inputs for the number of images to use, the number of regions to select per image, the region size, and the number of estimates to average when reporting p-values and accuracies (choice of these parameters is discussed below).

For location testing, a set of features that do not require segmentation of the image into individual cell regions was extracted from each region as described previously (Newberg and Murphy 2008), with the modification that horizontal and vertical features were combined to produce a set of 592 rotation invariant features. These include texture features at many different levels of resolution and nuclear overlap features. We used the 57 feature subset that was previously selected to be able to distinguish eight subcellular location classes with high accuracy (Newberg and Murphy 2008). The equivalence of the distributions of these features for normal and cancer regions was evaluated by the Friedman-Rafsky (FR) test. Because the test is non-parametric and does not make assumptions about the distributions from which the samples are drawn, it is suitable for small numbers of regions and large numbers of features.

Expression p-values were calculated by normalizing the mean protein intensity level across each region used in the location analysis by the respective mean nuclear intensity level. This results in a one-dimensional set of points corresponding to the regions for normal protein expression, and cancer protein expression. We calculated a p-value that the points in the two sets were drawn from the same distribution using the Wald Wolfowitz test, the one dimensional version of the FR test.

The reported p-values and accuracies for each protein were calculated by taking the average of 35 estimates which we found produced consistent ranked lists (see estimation of generalizability).

Selecting image sets, number of regions and region size

The database has up to 3 images for each normal tissue and up to 30 images for each cancer tissue. In order to have the same null distribution for the nonparametric p-values, we needed to use the same number of normal and cancer images for each antibody. To identify the optimal number of each, we randomly selected 200 antibodies for each tissue, selected 2 regions from each image, and assessed the extent to which the validation markers were ranked highly by location p-values (using the area under ROC curves, AUC). The number of normal images was varied from 1 to 3 and the number of cancer images from 3 to 24. We found that the best AUC average over all 200 antibodies resulted from using 2 normal images and 17 cancer images.

Next, the optimal number of regions was found by using the same 200 antibody training set, and 2 normal and 17 cancer images. We varied the region count from 2 to 5 from each of the images. The best performance as measured by AUC resulted from using 5, 3, 3 and 4 regions per image for breast, liver, prostate and bladder tissues respectively, and we therefore used these values for the full analysis sets. Differences in the optimal number of regions for different tissues presumably reflect tissue-specific variations across the normal and cancer states. Limiting the number of regions per image prevents the sampling of non-cellular regions in each tissue.

The optimal region size was chosen by assessing the performance of 100 randomly chosen proteins in ROC curves for each tissue. Ideally the region should be small enough so as to only capture cellular areas from a tissue image, since capturing non-cellular regions introduces new textures to the analysis that would affect the subcellular location features. The optimal radius was selected to be 75 pixels with an average AUC of 0.67 for the four tissues.

Classification

We used nearest neighbor classifiers with the 57 z-scored features and used cross-validation to estimate of the ability of a given protein to distinguish normal from cancer images, or to distinguish low-grade from high-grade cancer images. Images were assigned the majority class of their regions.

Distinguishing cancer grades

The prostate and bladders cancers are identified as high or low grade in the HPA, with approximately equal numbers of each. We partitioned the cancer images by grade and ran the pipeline to compare them. We also randomly selected 3 images from each set and calculated the 3-class classification accuracy for a nearest neighbor classifier (using leave-one-out cross-validation). This was repeated 35 times (producing a consistent ranking, as explained above) for different sets of randomly selected images and the average of the 35 accuracies is reported. Proteins that did not have at least 3 images from each disease state were excluded. The results are contained in Dataset S2.

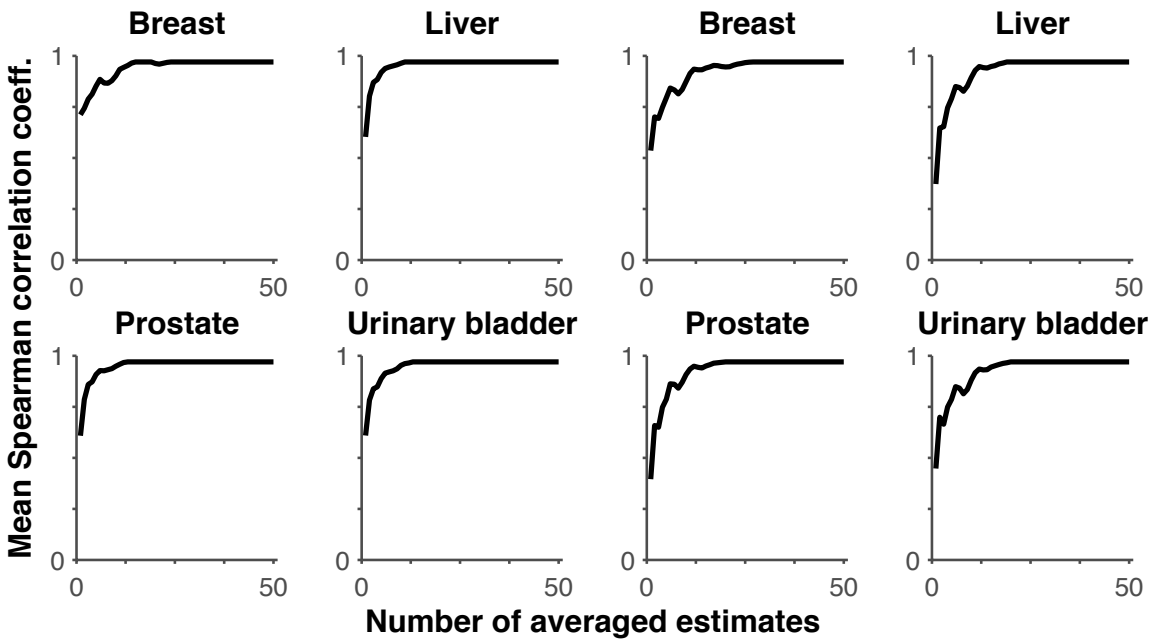


Fig. 12. Protein rank correlations using location p-values and classification accuracies when averaging different numbers of estimates. The number of estimates used to calculate the location p-value and accuracy was varied from 1 to 50. The Spearman correlation coefficient was used to measure the consistency of the ranked protein lists when different numbers of estimates were used.

Rank Consistency

Proteins are ranked by their p-values to find location biomarker candidates. Each p-value and accuracy is calculated by sampling 2 normal and 17 cancer images from the respective image sets for each protein. This list would presumably be different if we picked a different set of 2 normal and 17 cancer images. A solution would be to average the p-values from many random samplings of 2 and 17 images. Therefore, we determined how many p-value estimates we would need to average to produce a consistently ranked list.

To do this, we created ranked lists from protein p-values that had been averaged from different numbers of random samplings (from 1 to 50). We did this 10 times

for each number of samplings, and calculated the Spearman correlation between all pairwise combinations of the resulting 10 lists (the Spearman correlation coefficient is a nonparametric measure of how well two variables monotonically increase together).

The left panels of Fig. 12 show the p-value ranking consistency (measured by the average Spearman correlation coefficients) for the four tissues as a function of the number of estimates. The plots show that the rank becomes highly consistent (correlation close to 1) as the number of estimates increases. This process was repeated for classification accuracy (right panels Fig. 12) and a similar trend was observed. Therefore, we chose to use the average of 35 estimates for all of the p-values and accuracies reported in Datasets S1 and S2.

Estimation of generalizability

For each antibody, a classifier was trained using regions from one normal image and regions from one cancer image, where the number of regions was determined independently for each tissue (see methods). One held out normal image and one held out cancer image were then classified. A second classifier was trained with two normal images and two cancer images. The third normal image and a third cancer image were then classified. These steps were repeated for 35 samplings of training and testing images for each antibody and the mean accuracy for each level of cross-validation was calculated (i.e., the average accuracy when training with one image of each class and the average accuracy when training with two images of each class).

We then calculated the correlation between the two accuracies for each tissue (Fig. 6), and found them to range from 0.90 to 0.91. While this is much higher than the zero correlation that would occur at random, it indicates (as might be expected) that one image of each class is not sufficient to train a highly generalizable classifier.

We performed a similar test of the generalizability of p-value estimates from the FR test. In this case, the first estimate was made by sampling 2 normal and 17 cancer images, a second estimate was made by sampling a subset from the 2 and 17 images (1 image and 16 images respectively), the average of 35 samplings are reported for each estimate (Fig. 6). We found the correlations between the two p-values to be greater than 0.94 for all tissues, indicating that our reported p-values are likely good estimates of performance on new images.

Displaying regions

For each antibody, we performed hierarchical clustering using the features for each region and applied optimal leaf ordering to the leaves. For visualization purposes, we cut the tree to give 10 clusters. For each of these, we found the region closest to the mean feature value for the leaves in that cluster. We selected one representative antibody for breast and liver.

To illustrate how the features reflect the patterns for the full set of regions, the full hierarchical clustering tree and the ordered regions are shown in Fig. 8 for one antibody in bladder (HPA034715 against ARHGEF3). For this antibody, pathologist annotations indicated a subcellular location in every cancer sample from nuclear/cytoplasmic/membranous to nuclear (it was thus one of the true positives used in measuring performance of our system). The clustering shows a progressive change in the location pattern and most normal and cancer regions cluster with each other, as expected. Upon close inspection it can be seen that while annotations indicate that normal and cancer images have distinct non-overlapping location distributions, our method organizes the regions to show a progression of location change, highlighting the visually overlapping distributions for the two disease states.

Family wise-error calculation

We calculated expression and location p-values for each pathway, and we ranked the pathways by the extent of expression and location changes (Dataset S3). To determine whether any of the pathways had statistically significant changes we calculated a Bonferroni-Holm (BH) correction, which controls the Familywise error rate when making multiple comparisons. The correction keeps the effective Familywise error rate at alpha when there is more than one comparison. Given a set of hypotheses of size *m the* corrected significance threshold for all hypothesis (H) (pathways) is a function of its rank position (k) and the naiive significance level (α), in our case 0.05. Null hypotheses H_1 to H_k can be rejected by finding the smallest k that satisfies the inequality : $P_{(k)} > \alpha / (m+1-k)$.

KEGG Pathways and translocated proteins

Biochemical pathway networks were downloaded from the KEGG database (http://www.genome.jp/kegg/pathway.html) as KGML files. The files were parsed to directed graphs where nodes represent proteins referenced by Entrez ID numbers and edges represented interactions. The parsing into a graph structure was done in R using KEGGgraph package available from Bioconductor (http://www.bioconductor.org/). In some cases the original pathways in KEGG have nodes that represent metabolites or gene products, or for some metabolic pathways the edges represent proteins and the nodes are reactions. The default KEGGgraph package parses the graphs to a consistent format of protein at the nodes and interactions at the edges. We selected the option to list all paralogs for each protein to account for the possibility of multiple names for the same protein.

Next we mapped the Ensemble ID numbers for the proteins in the analysis set to the respective Entrez ID numbers and labeled the nodes in each graph with the respective location and expression scores from our analysis. This resulted in 268

KEGG pathways where each pathway i has n_i nodes, and m_i nodes in the network have pipeline or annotation values assigned, where $m_i <= n_i$.

To determine whether a pathway significantly changed location we calculated a network score from the location p-values of the m_i known proteins. Network scores were calculated by taking the sum of logs of the protein node p-values. We tested the hypothesis that the pathway score was drawn from a background distribution of 100 random networks scores of size m_i. For example a pathway with 30 known proteins was tested against a background distribution of 100 random networks, where each random network had 30 known proteins, while a pathway of 500 known proteins was tested against a background distribution containing random networks of 500 known proteins. Random networks were created by sampling m_i proteins from all of the known protein p-values in the 268 KEGG pathways. The score of pathway i was compared to its background distribution in a t-test to determine the probability that the pathway changed location. The significance threshold on the p-values was corrected using the Bonferroni-Holm multiple hypothesis correction to control for Familywise-error rate. We then repeated the same analysis using the expression p-values from the pipeline.

Pathway p-values were also calculated using the pathologist annotations. Under the assumption that the cancer images are independent, the annotation p-value for a given protein was calculated as I^N, where I is the empirical probability of change in that tissue, and N is the number of cancer images with a different annotation label. This was done by tissue for both location and expression. In breast cancer the empirical probability for a subcellular location change was 0.27 and for expression it was 0.50; in liver cancer it was 0.43 and 0.56; in prostate it was 0.26 and 0.49; in bladder it was 0.30 and 0.56 for location and expression, respectively. All results for pathway p-values are listed in Dataset S3, and presented in Fig. 11.

Robustness to JPEG compression

The images in the Human Protein Atlas database are approximately 3000x3000 pixels and are stored as JPEG compressed files. JPEG compression is a lossy format that aims to preserve visually distinguishable characteristics of an image while downsampling parts of the image that are not visually distinguishable. Our texture features quantify changes at varying levels of resolution. To investigate the dependence of the performance of our system upon potential JPEG compression artifacts, we compressed the original images from the Protein Atlas at varying JPEG compression levels using the imwrite function in Matlab. We then assessed how well the known location biomarkers were found by constructing ROC curves (as in Fig. 3) for varying extents of additional compression. As shown in Fig. S7, the AUCs were not extensively reduced in three tissues suggesting that the JPEG compressed images in the Atlas may not have had much effect on our detection pipeline. Further studies using uncompressed images will be needed to fully assess the impact of compression.

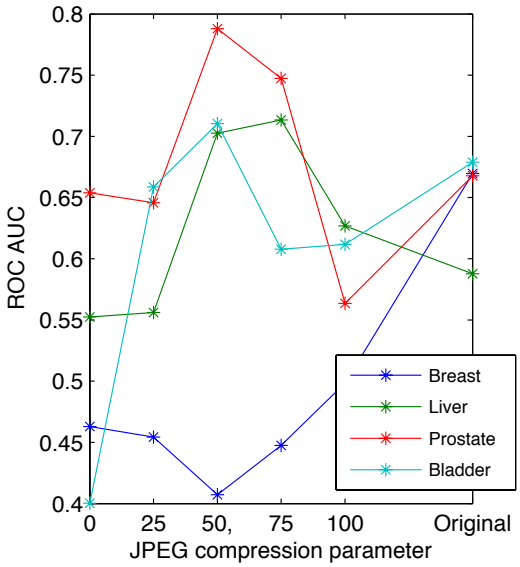


Fig. 13. The effect of JPEG compression on the performance of the pipeline. The images were JPEG compressed at different levels and processed through the pipeline. The performance for detecting known location biomarkers was measured and reported as areas under ROC curves (AUCs) as was done for Fig. 3.

Acknowledgments

We thank the Human Protein Atlas project team for making the valuable collection of IHC images publicly available, and E. Lundgren and M. Uhlén for helpful discussions. We would also like to thank members of the Murphy laboratory for valuable suggestions, and the anonymous reviewers of the manuscript for comments resulting in significant improvements in the analysis. This work was supported by NIH grant GM075205 and by grant 4100059192 from the Commonwealth of Pennsylvania CURE program. AK was supported by NIH training grant EB009403 and AR by a postdoctoral fellowship from the Lane Fellows program.

Chapter 3

Differential protein subcellular location and protein signatures for classifying adult and pediatric liver lesions

Abstract

Primary liver tumors occur in people of all ages from newborns to geriatric patients as hepatic lesions. Hepatic lesions range in severity from benign to malignant where the most malignant cases are fatal. Identifying premalignant and malignant liver lesions usually requires a biopsy however diagnosis of the biopsy can be challenging because different lesions can have similar morphological appearances (Ferrell 1995). Thus, various tests are run to help the clinician gather information about the disease including measurements of well-studied markers include differential protein expression levels, tissue architecture, and patient demographic and health information. We extended the previously developed computational IHC image analysis pipeline (Kumar, Rao et al. 2014) to investigate whether differential protein subcellular location can help to identify different liver lesions. We extended the IHC pipeline to process full slide images. Using the pipeline we explored how measuring changes in protein subcellular location can aid in identifying adult and pediatric liver lesions. Our results indicate that most of the time single protein measurements are poor markers for the lesions. Next we explored lesion-specific protein signatures for identifying diseases. Given our dataset we found a signature set of proteins that can successfully identify liver lesions in adult and pediatric populations with perfect accuracy. Finally we report two new proteins that aid in classifying the lesions when used as part of a signature protein set.

Introduction

Primary liver tumors occur in people of all ages from newborns to geriatric patients as hepatic lesions. Hepatic lesions range in severity from benign to malignant where the most malignant cases are fatal. Identifying premalignant and malignant liver lesions usually requires a biopsy however diagnosis of the biopsy can be challenging because different lesions can have similar morphological appearances (Ferrell 1995).

Pediatric lesions

Pediatric liver lesions are challenging to diagnose; the three states are: normal liver (nl), fetal-type hepatoblastoma (FHB), and well-differentiated hepatocellular carcinoma (WDHCC). The ability to differentiate a pure fetal-type hepatoblastoma from normal liver or a WDHCC in a liver resection specimen where the entire tumor is available for microscopic examination can be challenging let alone in a needle biopsy sample where visual and architectural context may be not available. Fetal hepatoblastoma (FHB) and well-differentiated hepatocellular carcinoma (WDHCC) can have similar phenotypes but the treatments of FHB and WDHCC are very different (Isaacs 2007) (Litten and Tomlinson 2008) making confident and accurate diagnoses of these lesions very important.

In the clinic, biopsies are stained with protein-specific antibodies to detect differential expression levels. Glypican-3, beta-catenin, Heppar-1 and other stains are almost always used (Li, Liu et al. 2013) (Libbrecht, Severi et al. 2006) (Kandil, Leiman et al. 2007) (Wang, Anatelli et al. 2008). Patient demographic information, specifically age, can also help identify the disease because some diseases onset at specific ages and stages of development or aging. For example, more than 90% of patients who get FHB are under the age of 5, as opposed to WDHCC. While WDHCC, has some architectural differences including three cell hepatocyte plate thickness and greater nuclear atypia and mitoses than FHB (LV 2004), both can appear very

similar. Some reports have shown that there are significant changes in transcription and in the genome of patients with HCC and FHB, and FHB and normal liver (Yamada, Ohira et al. 2004) (Luo, Ren et al. 2006).

Adult lesions

Adult liver lesion diagnosis is very challenging due to the number of lesions with similar morphological appearances (Ferrell 1995): WDHCC, hepatocellular adenoma (HCA), nodular regenerative hyperplasia (NRH), focal nodular hyperplasia (FNH), macroregenerative nodules (MRN), and borderline dysplastic nodules (BN). Age can be used to help identify FNH and HCA, as they occur as a single mass in a younger age group than NRH and HCC which occurs in older patients and have multiple masses. NRH is associated with other diseases and toxin and drug exposure, vascular disorders and connective tissue disorders. HCC is associated with cirrhotic liver disease. MRN, BH and HCC occur in damaged and cirrhotic livers. MRN and BN can be found in the same nodule in the liver.

Light microscopy is used to assess the morphological characteristics of the lesions. Distinctive qualities include single or multiple nodules, presence of central scar, thickness of hepatocellular trabeculae, pattern of vascularization, presence of Mallory's hyaline, and small vascular thrombi. Structure and characteristics of the nucleus have been shown to be unreliable (LV 2004).

Improving diagnostic methods

As clinicians collect information about patients from various clinical tests their confidence in the initial disease diagnosis may change. That is, the clinician's confidence in the first diagnosis was low and new information allowed for a more informed later diagnosis. Finding strong markers that can identify diseases will significantly improve the diagnostic process and lead to improved patient care.

Various tests are run to help the clinician gather information about the disease. Measurements of well-studied markers include differential protein expression levels, tissue architecture, and patient demographic and health information. The need to improve liver lesion characterization is immediate and this area has received a lot of attention from the medical community recently (Sukru Emre) (Dolores López-Terrada 2013) (Esmeralda Celia Marginean 2013).

Some subclasses of tumors can be identified by the expression levels of protein (Sotiriou and Piccart 2007) (Reis-Filho, Weigelt et al. 2010). In some cases protein subcellular location can predict patient outcome. For example in pediatric hepatoblastomas nuclear localization of B-catenin is an indicator of patient survival (Sang Park, Ra Oh et al. 2001). Subcellular location of proteins can predict patient survival in other cancers as well (Hung and Link 2011). Here, we investigate whether protein subcellular location measured from protein-specific IHC images can aid and improve the identification of liver lesions in pediatric and adult populations.

Computational quantitative pathology

Automatic protein quantitative systems can help to identify protein biomarkers and signatures and in this case to improve the accuracy of liver lesion diagnosis. In diagnostic research, computational pathology methods have identified predictive features that were previously unknown (Hoque, Lippman et al. 2001), and in some cases these methods were able to show the importance of features that had been regarded as unimportant (Beck, Sangoi et al. 2011) (Guillaud, Adler-Storthz et al. 2005). Identifying discriminating features is valuable not only while making the diagnosis but also they may provide insight into the development of the disease.

We extended the computational IHC image analysis pipeline from Chapter 2 to investigate whether differential protein subcellular location can help to identify different liver lesions. The image processing section was revised to handle full slide

images. Our pipeline uses protein expression and subcellular location together to make the classification. Next we construct a series of classifiers for the liver lesions in pediatric populations, and in adult populations. Finally we report the best classifier for the lesions and we make recommendations about the best markers and measurements

Data

The dataset consists of full slide IHC images of liver lesions stained with 13 protein-specific antibodies, H&E and Feulgen. The images were collected from Omnyx V4 scanner at the UPMC Shadyside Hospital in Pittsburgh, PA. The tissue sections came from a collaboration with Dr. John Ozolek at the UPMC Children's hospital.

Paraffin embedded tissue sections were prepared from the patient samples and each section was stained with Hematoylin making the nuclei of basophilic cells appear blue. Next, a protein-specific primary antibody was used to bind the protein of interest, followed by a secondary horseradish peroxidase (HPR) -conjugated antibody was used to bind the primary antibody. Finally DAB was added to the slide which forms a brown precipitate when it reacts with HRP. RGB images of each section are captured using the scanner.

The full tissue sections were scanned on the Omnyx scanner by collecting small partially overlapping tiles. The tiles were stitched together and stored in JPEG 2000 format. The images are collected at 60x magnification, 0.1385 um/pixel. The size of each image is approximately 1.4e5 x 3.0e5 pixels.

The dataset consists of 86 patients, where tissue sections from each patient are individually stained with 13 protein-specific antibodies, H&E and Feulgen. The distribution of patients referenced by ID numbers across diseases and age groups is as follows:

Pediatric cases:

- Fetal hepatoblastoma (FHB 6) : FHB1, FHB2, FHB3, FHB4, FHB5, FHB6.
- Hepatocellular adenoma (HCA 4) : HCA1, HCA2, HCA3, HCA4.
- Hepatocellular carcinoma (HCC 1) : HCC1.
- Macroregenerative nodules (MRN 1) : MRN1.
- Focal nodular hyperplasia (FNH 3) : FNH1, FNH2, FNH3.

Adult cases:

- Dysplastic nodules (DN 13) : DN1, DN2, DN3, DN4, DN5, DN6, DN7, DN8, DN9, DN10, DN11, DN12, DN13.
- Focal nodular hyperplasia (FNH 23) : FNH4, FNH5, FNH6, FNH7, FNH8, FNH9, FNH10, FNH11, FNH12, FNH13, FNH14, FNH15, FNH16, FNH17, FNH18, FNH19, FNH20, FNH21, FNH22, FNH23, FNH24, FNH25, FNH26.
- Hepatocellular adenoma (HCA 5) : HCA5, HCA6, HCA7, HCA8, HCA9.
- Hepatocellular carcinoma (HCC 16) : HCC2, HCC3, HCC4, HCC6, HCC7, HCC8, HCC9, HCC10, HCC11, HCC12, HCC13, HCC14, HCC15, HCC16, HCC18, HCC19.
- Macroregenerative nodules (MRN 14) : MRN2, MRN3, MRN4, MRN5, MRN6, MRN7, MRN8, MRN9, MRN10, MRN11, MRN12, MRN13, MRN14, MRN15.

Stains:

H&E – Hematoxylin binds DNA/RNA. Eosin binds proteins.

Feulgen – binds chromosomal material or DNA in cells.

13 protein-specific antibodies:

	Antibody name	**Protein name**	**Gene ID**	**Entrez Gene**
1	B-catenin	Catenin (Cadherin-Associated Protein), Beta 1	CTNNB1	1499
2	CRP	C-Reactive Protein, Pentraxin-Related	CRP	1401
3	Glutamine Synthase	Glutamate-Ammonia Ligase	GLUL	2752
4	Glypican-3	Glypican-3	GPC3	2719
5	HSP70	HSPA (Heat Shock 70kDa) Binding Protein, Cytoplasmic Cochaperone 1	HSPBP1	23640
6	L-FABP	fatty acid binding protein 1, liver	FABP1	1268
7	DEK	DEK Oncogene	DEK	7913
8	DKC1	dyskeratosis congenita 1, dyskerin	DKC1	1736
9	IRX6	Iroquois homeobox 6	IRX6	79190
10	KI67	Marker of proliferation Ki-67	MKI67	4288
11	NDUFAF1	NADH dehydrogenase (ubiquinone) complex I, assembly factor 1	NDUFAF1	51103
12	NPM1	nucleophosmin (nucleolar phosphoprotein B23, numatrin)	NPM1	4869
13	TIP1	Tax1 (Human T-Cell Leukemia Virus Type I) Binding Protein 3	TAX1BP3	30851

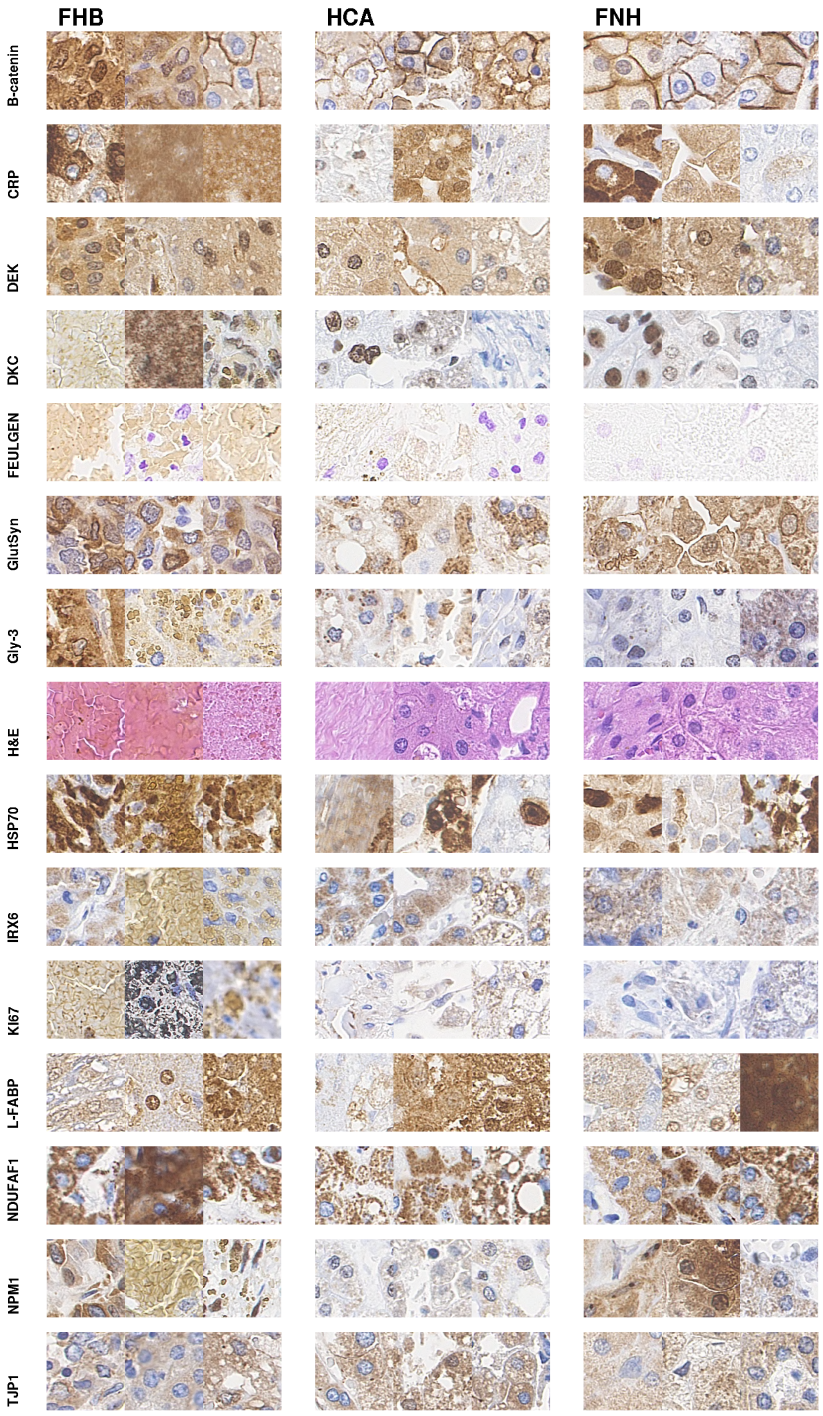


Fig. 1: Pediatric liver lesions. Sections from each patient were stained with 14 antibodies and H&E stains. There were three pediatric liver lesion groups used in the analysis: FHB, HCA, FNH. Each disease-antibody panel shows regions from three different patients.

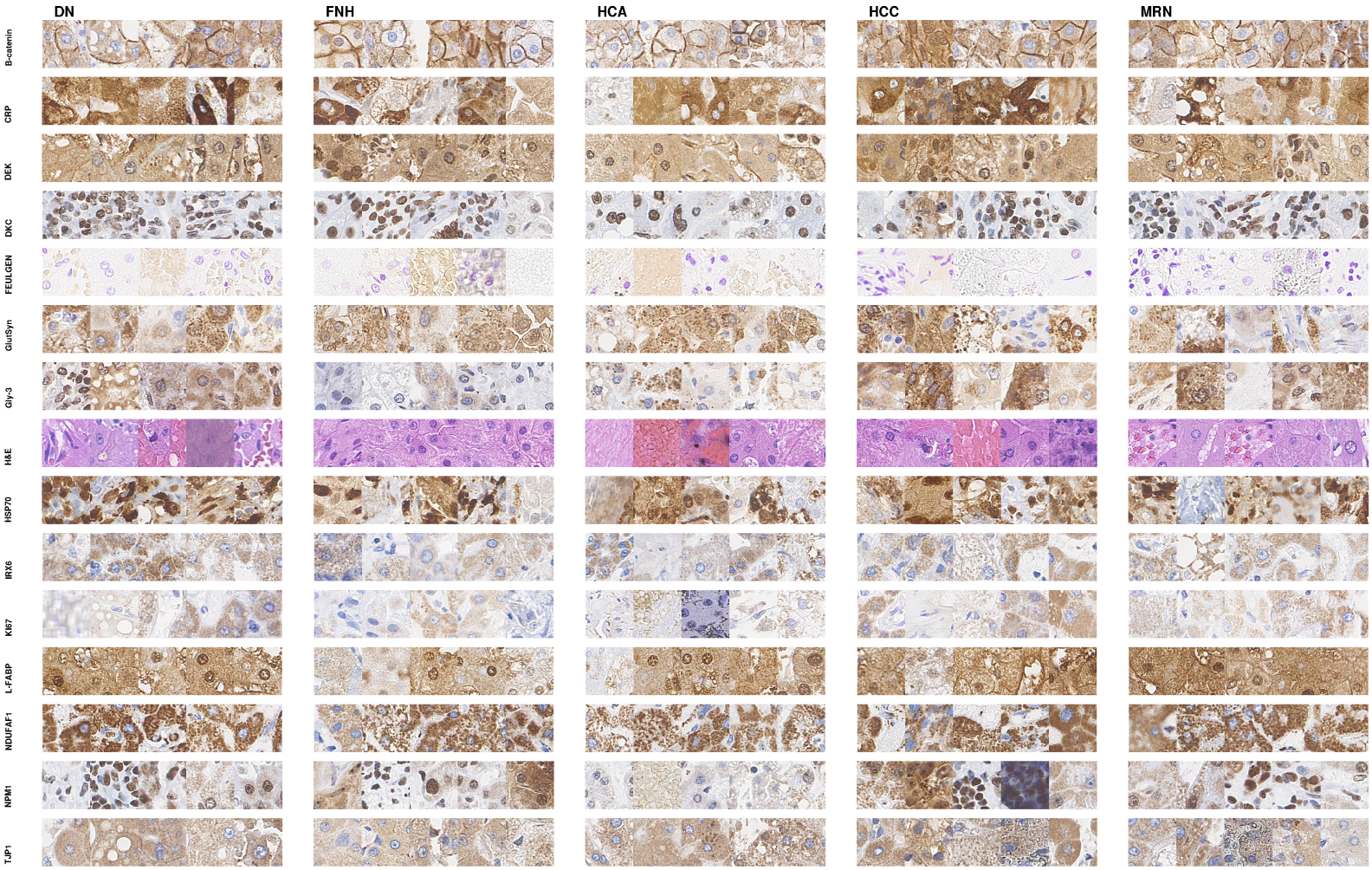


Fig. 2: Adult liver lesions. Sections from each patient were stained with 14 antibodies and H&E stains. Five adult liver lesion groups were used in the analysis: DN, FNH, HCA, HCC, MRN. Each disease-antibody panel shows regions from five different patients.

Results

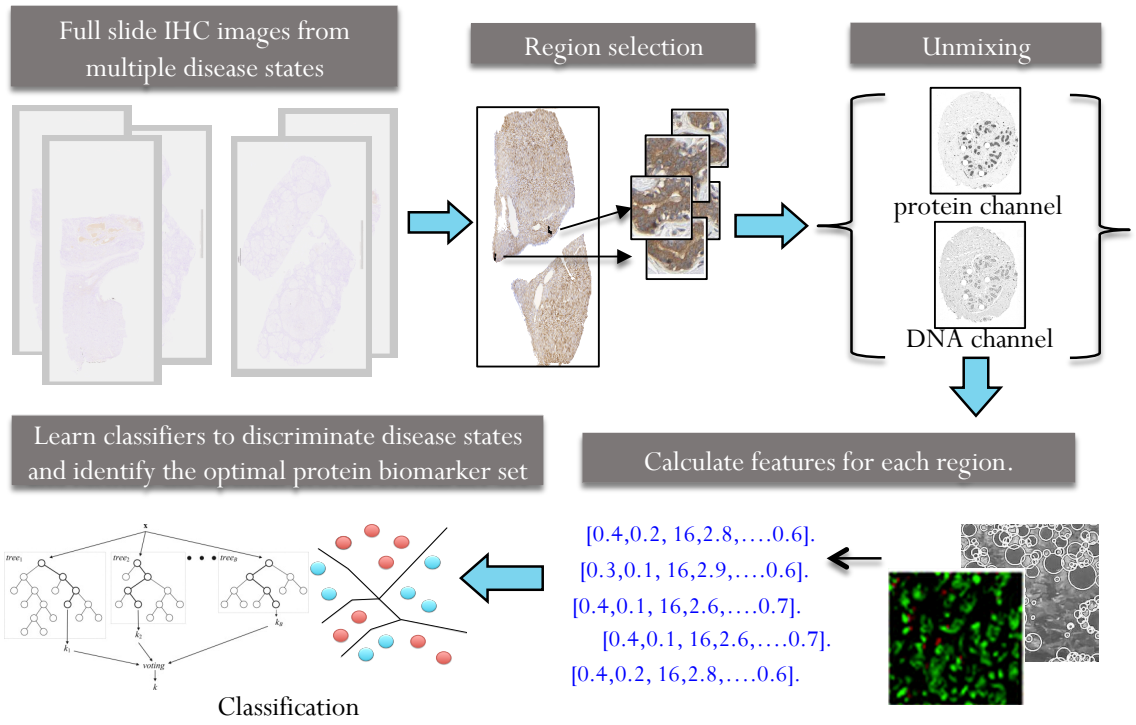


Fig 3. The automatic IHC image analysis pipeline was extended to process full slide images of multiple proteins and to perform multiclass classification using the sets of input proteins. The tissue object was segmented and regions were selected with high Brenner scores to select in-focus tiles. Regions were unmixed and features were calculated including original features, and new local features. Finally three types of classifiers were trained using the features, where each classifier was designed to model a diagnostic scenario using single proteins as markers, or a signature of proteins. Single class and multiclass classifiers were trained. The new pipeline outputs the best classifier and the optimal protein marker signatures for distinguishing the input disease states.

The IHC image analysis pipeline was extended to include new image processing steps, classification and feature selection. The pipeline is as follows:

Image processing:

1. Tissue object detection:

The tissue object was segmented from the full slide image so that background was not processed during region selection. The tissue object corresponded to the regions of the image that had the highest chromaticity and greatest frequency together.

2. Region selection

Regions were selected from the downsampled tissue object based on the maximum pixel intensity that corresponded to the highest stained regions. Region coordinates were mapped to the original image and selected. The top 200 regions with the highest protein stain were selected with varying diameters from 62, 125, 312, 625, 1250, 2500 pixels.

3. Unmixing and Features

Next the average color basis matrix for the top 200 regions from each image was calculated and the regions were linearly unmixed. For each region we calculated 1) the Brenner score, 2) Murphy lab features. The top 150 regions with the highest Brenner score were selected for the analysis.

Classification:

4. Classifiers were trained to identify disease states using the subsets of input proteins. The results from each classifier show the sets of proteins that may be potential biomarkers.

Classifiers were trained through several rounds of cross validation on the training set to find optimal parameters and finally the classifier was evaluated with the held out test set.

In some cases a disease state was represented by as few as three patients across the full dataset. To check whether the parameter estimation determined from single patients was generalizable to the rest of the data set we trained one-vs-rest classifiers with one data point, and two data points and cross-validated with a single held out point. This was done for all disease states and all antibodies in the analysis set. The results shown in Fig. 4 indicate a relatively good correlation between both subsets of the data, suggesting our results are generalizable during the different rounds of cross validation in the three classification scenarios we present below. The classification accuracies were also calculated using one patient per disease to train, and then two patients per disease. The generalizability results for accuracy are on the right of Fig. 4.

4.1 Single lesion classification: Pairwise

We determined how accurately a patient's disease can be identified when considering one of two possible lesions. That is, we determined how well each protein can discriminate pairwise lesions. We selected the disease states in question and then divided the set of images into training and testing sets as described above. The optimal model parameters, region count and region radius were learned for each classifier. The classification results are shown in Fig. 5 and Table 1. The left plot shows the pediatric pairwise classification accuracies for the three lesion types, the right plot shows the accuracies for the adult lesions. The

results are plotted in a square heatmap where the diagonal and the lower half of the plots are omitted due to redundancy.

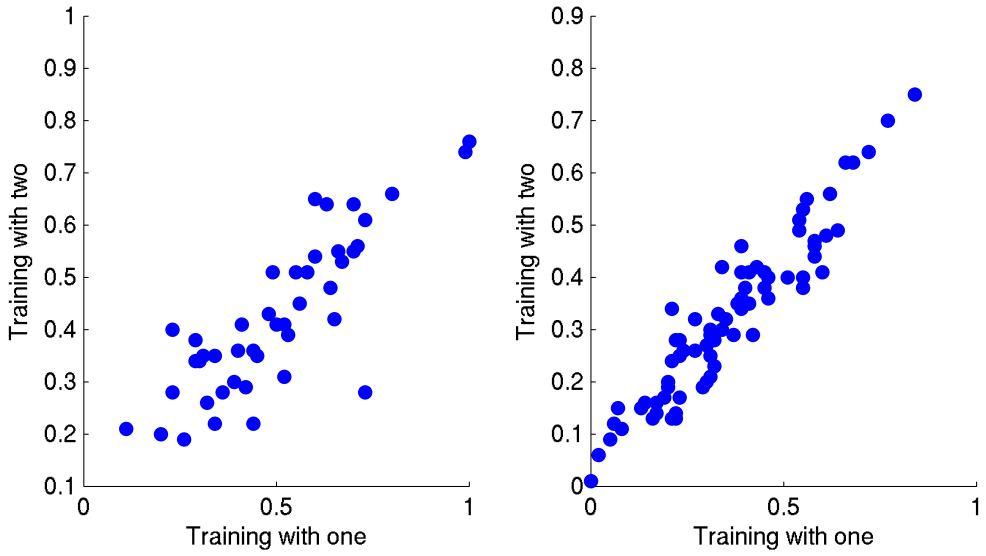
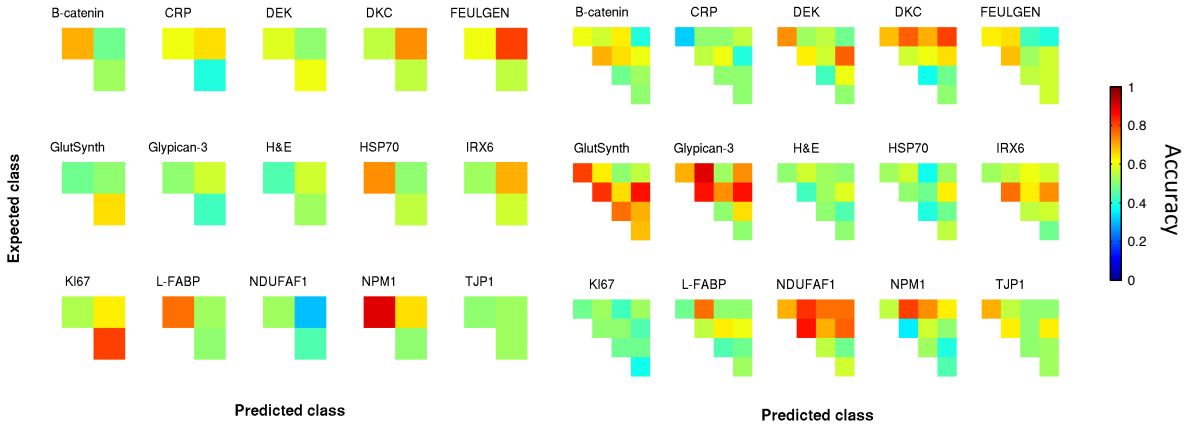


Fig. 4. The generalizability of the parameters (left) and classification accuracies (right) calculated from subsets of data. One-vs-rest classifiers were trained for each disease state and antibody combination. The x-axis shows the first round of estimation from training with one patient from each disease group. The y-axis shows the second round of estimation where the training group had two patients



from each disease group.

Fig. 5: Pairwise classification accuracy for the pediatric (left) and adult (right) lesions. The lesions were classified pairwise using a random forest classifier where the number of regions, size the region and model parameters were learned jointly. The order of the pediatric lesions are FHB, HCA, FNH from top the bottom, left to right. The adult lesions are DN, HCA, HCC, MRN, FNH from top to bottom, left to right. Accuracies are listed in Table 1.

In pediatric lesions NPM1 (90%) had the highest classification accuracy for separating FHB and HCA, while all other proteins had a lower accuracy. L-FABP (75%), HPS70 (73%) and B-catenin (70%) did not perform as well in FHB vs HCA. The best separation for FHB and FNH was achieved by Feulgen (80%), DKC (73%) and IRX6 (70%). The highest classification accuracy for HCA and FNH was achieved by KI67 (80%). Some markers that are currently useful in the clinic for identifying lesions were not useful in this system when we measured subcellular location differences, such as glutamine synthase, glypican-3 and CRP which all performed poorly as shown in Table 1.

In adult lesions glypican-3 had the highest pairwise classification accuracy (90%) for separating DN and HCC, while NDUFAF1 (83%) and NPM1 (80%) had lower performance. Glypican-3 had moderate performance for separating HCA and HCC where the highest accuracy was (85%) and glutamine synthase had comparable performance (83%). Glutamine synthase (80%) moderately separated DN and HCA. MRN was poorly separated from the 4 other adult lesions where NDUFAF1 (75%) performed the best against DN and lower against the other lesions. In FNH, DKC (80%) separated DN with moderate power, while glutamine synthase (85%) and glypican-3 (85%) performed slightly better against HCA. FNH and HCC were poorly separated, where the best marker was glutamine synthase (70%).

Bcat.	HCA	FNH
FHB	0.7	0.47
HCA		0.52

Gl.Syn.	HCA	FNH
FHB	0.48	0.5
HCA		0.65

KI67	HCA	FNH
FHB	0.54	0.63
HCA		0.8

CRP	HCA	FNH
FHB	0.6	0.65
HCA		0.4

Gly-3	HCA	FNH
FHB	0.5	0.58
HCA		0.43

LFABP	HCA	FNH
FHB	0.75	0.53
HCA		0.5

DEK	HCA	FNH
FHB	0.58	0.5
HCA		0.6

H&E	HCA	FNH
FHB	0.45	0.57
HCA		0.53

NDUF	HCA	FNH
FHB	0.53	0.3
HCA		0.45

DKC	HCA	FNH
FHB	0.55	0.73
HCA		0.55

HSP70	HCA	FNH
FHB	0.73	0.5
HCA		0.55

NPM1	HCA	FNH
FHB	0.9	0.65
HCA		0.5

FLEU	HCA	FNH
FHB	0.6	0.8
HCA		0.55

IRX6	HCA	FNH
FHB	0.53	0.7
HCA		0.58

TJP1	HCA	FNH
FHB	0.5	0.53
HCA		0.53

Table 1a. Pairwise classification accuracy for the pediatric liver lesions by protein.

B-cat.	HCA	HCC	MRN	FNH
DN	0.6	0.58	0.63	0.4
HCA		0.68	0.48	0.53
HCC			0.48	0.5
MRN				0.5

H&E	HCA	HCC	MRN	FNH
DN	0.5	0.58	0.53	0.5
HCA		0.43	0.53	0.58
HCC			0.5	0.45
MRN				0.5

CRP	HCA	HCC	MRN	FNH
DN	0.33	0.5	0.5	0.55
HCA		0.55	0.6	0.4
HCC			0.5	0.5
MRN				0.5

HSP70	HCA	HCC	MRN	FNH
DN	0.53	0.58	0.38	0.53
HCA		0.5	0.48	0.63
HCC			0.4	0.48
MRN				0.55

DEK	HCA	HCC	MRN	FNH
DN	0.73	0.53	0.55	0.48
HCA		0.63	0.58	0.78
HCC			0.43	0.6
MRN				0.5

IRX6	HCA	HCC	MRN	FNH
DN	0.53	0.55	0.6	0.58
HCA		0.75	0.63	0.73
HCC			0.55	0.58
MRN				0.48

DKC	HCA	HCC	MRN	FNH
DN	0.68	0.78	0.7	0.8
HCA		0.58	0.6	0.65
HCC			0.37	0.48
MRN				0.5

KI67	HCA	HCC	MRN	FNH
DN	0.48	0.53	0.43	0.53
HCA		0.5	0.5	0.45
HCC			0.48	0.48
MRN				0.38

FEUL	HCA	HCC	MRN	FNH
DN	0.63	0.65	0.43	0.4
HCA		0.68	0.58	0.58
HCC			0.55	0.58
MRN				0.58

LFABP	HCA	HCC	MRN	FNH
DN	0.48	0.75	0.5	0.5
HCA		0.55	0.63	0.6
HCC			0.45	0.48
MRN				0.53

Gl.Syn.	HCA	HCC	MRN	FNH
DN	0.8	0.63	0.5	0.55
HCA		0.83	0.65	0.85
HCC			0.75	0.7
MRN				0.68

NDUF	HCA	HCC	MRN	FNH
DN	0.7	0.83	0.75	0.75
HCA		0.85	0.7	0.78
HCC			0.55	0.48
MRN				0.58

Gly-3	HCA	HCC	MRN	FNH
DN	0.7	0.9	0.53	0.73
HCA		0.85	0.73	0.85
HCC			0.5	0.65
MRN				0.5

NPM1	HCA	HCC	MRN	FNH
DN	0.55	0.8	0.73	0.63
HCA		0.35	0.58	0.5
HCC			0.53	0.4
MRN				0.45

TJP1	HCA	HCC	MRN	FNH
DN	0.7	0.55	0.5	0.5
HCA		0.63	0.5	0.63
HCC			0.5	0.53
MRN				0.53

Table 1b. Pairwise classification accuracy for the adult liver lesions by protein.

4.2 Multi-class single protein classifier

Next, we determined which proteins can be used as markers to identify a patient's disease given that the patient could have any disease present in our dataset their respective age group. Our approach was to find proteins that can act as markers to identify multiple disease states simultaneously. In each age group we constructed a multiclass random forest classifier for each protein. The number of trees, number of regions and radius of the regions was optimized jointly and then a held out test set was used to calculate the cross-validation accuracy. In almost all pediatric cases the radius of the region was 1250 pixels, and 25 regions were used. In the adult cases the radius size and count varied across the full range.

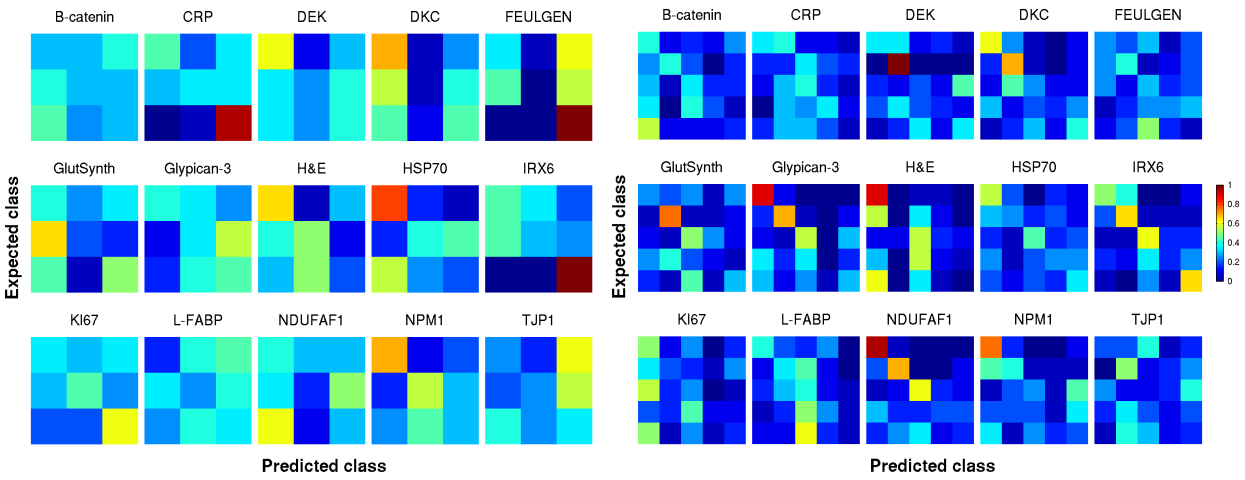


Fig. 6. The multi-class confusion matrices for pediatric (left) and adult (right) liver lesions. The three pediatric liver lesions in order on the confusion matrix are: FHB, HCA, FNH. In order, the adult lesions are DN, HCA, HCC, MRN, FNH. For each age group the lesions were classified using each protein independently in a random forest classifier where the number of trees, the number of regions and the size of the regions was learned at each round of testing. Each heatmap represents the cross-validation accuracies for each protein.

Our results suggest that some proteins are potential markers for specific disease states; they can identify specific lesions with high sensitivity in a multiclass system. For example, CRP (95%), Feulgen (100%), and IRX6 (100%) were able to correctly classify FHN however FHB and HCA were confused for FNH and the specificity was 65%, 42%, 78% in for CRP, Feulgen and IRX6, respectively. HSP70 (80%) could identify FHB, and on average HCA and FNH were confused with a specificity of 65%. Pediatric HCA was not easily identifiable with a single protein in a multiclass system, however multiple proteins together allowed for improved accuracy as discussed in the next section.

Bcat.	FHB	HCA	FNH
FHB	0.3	0.3	0.4
HCA	0.4	0.3	0.3
FNH	0.45	0.25	0.3

Gl.Syn.	FHB	HCA	FNH
FHB	0.4	0.25	0.35
HCA	0.65	0.2	0.15
FNH	0.45	0.05	0.5

KI67	FHB	HCA	FNH
FHB	0.35	0.3	0.35
HCA	0.3	0.45	0.25
FNH	0.2	0.2	0.6

CRP	FHB	HCA	FNH
FHB	0.45	0.2	0.35
HCA	0.3	0.35	0.35
FNH	0	0.05	0.95

Gly-3	FHB	HCA	FNH
FHB	0.4	0.35	0.25
HCA	0.1	0.35	0.55
FNH	0.15	0.4	0.45

LFABP	FHB	HCA	FNH
FHB	0.15	0.4	0.45
HCA	0.35	0.25	0.4
FNH	0.25	0.4	0.35

DEK	FHB	HCA	FNH
FHB	0.6	0.1	0.3
HCA	0.35	0.25	0.4
FNH	0.35	0.25	0.4

H&E	FHB	HCA	FNH
FHB	0.65	0.05	0.3
HCA	0.4	0.5	0.1
FNH	0.3	0.5	0.2

NDUF	FHB	HCA	FNH
FHB	0.4	0.3	0.3
HCA	35	0.15	0.5
FNH	0.6	0.1	0.3

DKC	FHB	HCA	FNH
FHB	0.7	0.05	0.25
HCA	0.55	0.05	0.4
FNH	0.45	0.1	0.45

HSP70	FHB	HCA	FNH
FHB	0.8	0.15	0.05
HCA	0.15	0.4	0.45
FNH	0.55	0.25	0.2

NPM1	FHB	HCA	FNH
FHB	0.7	0.1	0.2
HCA	0.15	0.55	0.3
FNH	0.25	0.45	0.3

FLEU	FHB	HCA	FNH
FHB	0.35	0.05	0.6
HCA	0.45	0	0.55
FNH	0	0	1

IRX6	FHB	HCA	FNH
FHB	0.45	0.35	0.2
HCA	0.45	0.3	0.25
FNH	0	0	1

TJP1	FHB	HCA	FNH
FHB	0.25	0.15	0.6
HCA	0.2	0.25	0.55
FNH	0.4	0.25	0.35

Table 2a. The multi-class confusion matrices for pediatric liver lesions.

B-cat.	DN	HCA	HCC	MRN	FNH
DN	0.4	0.1	0.15	0.1	0.25
HCA	0.25	0.4	0.2	0	0.15
HCC	0.3	0.05	0.35	0.15	0.15
MRN	0.35	0	0.4	0.2	0.05
FNH	0.55	0.1	0.1	0.1	0.15

H&E	DN	HCA	HCC	MRN	FNH
DN	0.9	0	0.05	0.05	0
HCA	0.55	0	0.35	0.1	0
HCC	0.1	0.1	0.55	0.15	0.1
MRN	0.3	0	0.55	0.1	0.05
FNH	0.6	0	0.35	0.05	0

CRP	DN	HCA	HCC	MRN	FNH
DN	0.35	0.4	0.1	0.1	0.05
HCA	0.15	0.15	0.35	0.2	0.15
HCC	0.1	0.3	0.4	0.15	0.05
MRN	0	0.3	0.25	0.35	0.1
FNH	0.15	0.3	0.3	0.2	0.05

HSP70	DN	HCA	HCC	MRN	FNH
DN	0.55	0.2	0	0.14	0.1
HCA	0.3	0.25	0.15	0.2	0.1
HCC	0.15	0.05	0.45	0.15	0.2
MRN	0.25	0.05	0.2	0.25	0.25
FNH	0.2	0.1	0.2	0.15	0.35

DEK	DN	HCA	HCC	MRN	FNH
DN	0.35	0.35	0.1	0.15	0.05
HCA	0	1	0	0	0
HCC	0.15	0.1	0.2	0.1	0.45
MRN	0.2	0.35	0.2	0.15	0.1
FNH	0.05	0.15	0.35	0.1	0.35

IRX6	DN	HCA	HCC	MRN	FNH
DN	0.5	0.4	0	0	0.1
HCA	0.2	0.65	0.05	0.05	0.05
HCC	0.05	0.05	0.6	0.15	0.15
MRN	0.15	0.1	0.35	0.15	0.25
FNH	0.05	0	0.25	0.05	0.65

DKC	DN	HCA	HCC	MRN	FNH
DN	0.6	0.25	0.05	0	0.1
HCA	0.15	0.7	0.05	0	0.1
HCC	0.1	0.45	0.25	0.1	0.1
MRN	0.3	0.1	0.2	0.15	0.25
FNH	0.05	0.15	0.03	0.1	0.4

KI67	DN	HCA	HCC	MRN	FNH
DN	0.5	0.1	0.25	0	0.15
HCA	0.35	0.2	0.25	0.05	0.25
HCC	0.55	0.15	0.25	0	0.05
MRN	0.2	0.15	0.45	0.1	0.1
FNH	0.5	0.05	0.25	0.05	0.15

FEUL	DN	HCA	HCC	MRN	FNH
DN	0.25	0.2	0.3	0.05	0.2
HCA	0.25	0.4	0.03	0.1	0.2
HCC	0.25	0.1	0.3	0.15	0.2
MRN	0.05	0.15	0.25	0.25	0.3
FNH	0.1	0.2	0.5	0.15	0.05

LFABP	DN	HCA	HCC	MRN	FNH
DN	0.4	0.2	0.15	0.25	0
HCA	0.1	0.35	0.45	0.1	0
HCC	0.15	0.3	0.4	0.1	0.05
MRN	0.05	0.15	0.5	0.25	0.05
FNH	0.1	0.1	0.6	0.15	0.05

Gl.Syn.	DN	HCA	HCC	MRN	FNH
DN	0.25	0.2	0.25	0.05	0.25
HCA	0.05	0.75	0.05	0.05	0.1
HCC	0.2	0.05	0.5	0.25	0.1
MRN	0.25	0.4	0.2	0.1	0.05
FNH	0.15	0.05	0.45	0.05	0.3

NDUF	DN	HCA	HCC	MRN	FNH
DN	0.95	0.05	0	0	0
HCA	0.2	0.7	0	0	0.1
HCC	0.05	0.1	0.6	0.15	0.1
MRN	0.3	0.15	0.15	0.2	0.2
FNH	0.15	0.25	0.35	0.1	0.15

Gly-3	DN	HCA	HCC	MRN	FNH
DN	0.9	0.1	0	0	0
HCA	0.15	0.7	0.05	0	0.1
HCC	0.1	0.05	0.55	0	0.3
MRN	0.35	0.15	0.35	0	0.15
FNH	0.3	0.1	0.25	0.05	0.3

NPM1	DN	HCA	HCC	MRN	FNH
DN	0.75	0.15	0	0	0.1
HCA	0.45	0.4	0.05	0.05	0.05
HCC	0.05	0.2	0.25	0.05	0.45
MRN	0.2	0.2	0.2	0.05	0.35
FNH	0.35	0.05	0.25	0.15	0.2

TJP1	DN	HCA	HCC	MRN	FNH
DN	0.2	0.2	0.4	0.05	0.15
HCA	0	0.5	0.1	0.15	0.25
HCC	0.25	0.1	0.1	0.15	0.4
MRN	0.15	0.35	0.2	0.1	0.2
FNH	0.05	0.4	0.3	0.1	0.15

Table 2b. The multi-class confusion matrices for adult liver lesions.

In the adult population many lesions could not be identified with high accuracy in a multiclass system. One exception was DN which had high specificity when stained with glypican-3 (90%), NDUFAF1 (95%) and H&E (90%), but lower sensitivity (87%, 82.5, 76%, respectively). It is interesting to note that while DEK performed moderately in a pairwise system for all adult lesions its specificity for HCA was 100% and the sensitivity was 77% in a multiclass system. These results suggest that glypican-3, NDUFAF1, DEK, H&E staining and other proteins may be useful for discriminating lesions with similar morphological appearances and they may provide additional information for clinicians when diagnosing a patient.

4.3 Multi-protein voting classifier

We next determined whether a set of proteins could be used together to identify a patient's disease given any lesion in their age group. Using a subset of patients from the training set multi-class single protein classifiers were trained as described in 4.2. For each population the held out set contained all proteins for one patient from each disease group. Each patient was either entirely in the training set or the held out set. The results from each classifier were used in a plurality voting scheme to identify the disease of the patient, as described below.

For each disease, signature sets of proteins were found by classifying the held out training set and for each disease selecting proteins that had a true positive rate greater than the false positive rate. Next, the classifiers were re-trained using the full training set. Finally, the held out test patients were classified by the plurality vote of the signature proteins for any disease. The signature determines if the patient had the disease, or not. The process was repeated 100 times by randomly selecting different training and held out sets. This resulted in 100 protein signatures for each disease and respective classification accuracies.

Fig. 7 shows that different signatures were found during different independent rounds of the learning process. Some signatures were found with greater frequency (as indicated by the colorbar) and consistently resulted in high classification accuracy of the held out patients set. Each bar represents a protein signature and is centered on the mean accuracy. The vertical length of the bar indicates two standard deviations from the mean. The sets of proteins at the top right of the plots are potential protein marker sets for identifying specific liver lesions.

We defined the optimal protein signature as the most frequent signature having a mean classification accuracy = 1, and standard deviation = 0. That is, these signatures were found repeatedly and always resulted in accurate classification. The signatures for each disease are clustered in the clustergram in the lower panel

of Fig. 7. The results show that the markers and disease fall roughly into dark clusters in the top right and bottom left corners of the clustergram.

MRN and FNH in adult and pediatric cases were distinguished by proteins sets that both included glypican-3, Feulgen and NDUFAF1. The remaining lesions were mostly classified by H&E and the other proteins: B-catenin, DEK, TJP1, DKC, CRP and HSP70. Interestingly adult and pediatric HCA were marked by different proteins, as were adult and pediatric FNH. B-catenin was only used to mark HCC in adults, and TJP1 and IRX6 were used to distinguish HCA in adult and pediatric groups. The optimal protein signatures are presented in Table 3.

Within each age group the proteins in each signature are mutually exclusive with the exception of glypican-3 and H&E in pediatric lesions, and FNH and MRN proteins in adult lesions. Given the accuracy of the protein signatures a multiclass classifier can be assembled where a patient is first classified to determine if they have the one of the diseases. If they do not then the patient can be classified in the second classifier, and if necessary finally in the third classifier to make the diagnosis. Given our results and dataset we expect such a scheme using the signatures presented in Table 3 for the adult and pediatric groups would provide perfect classification.

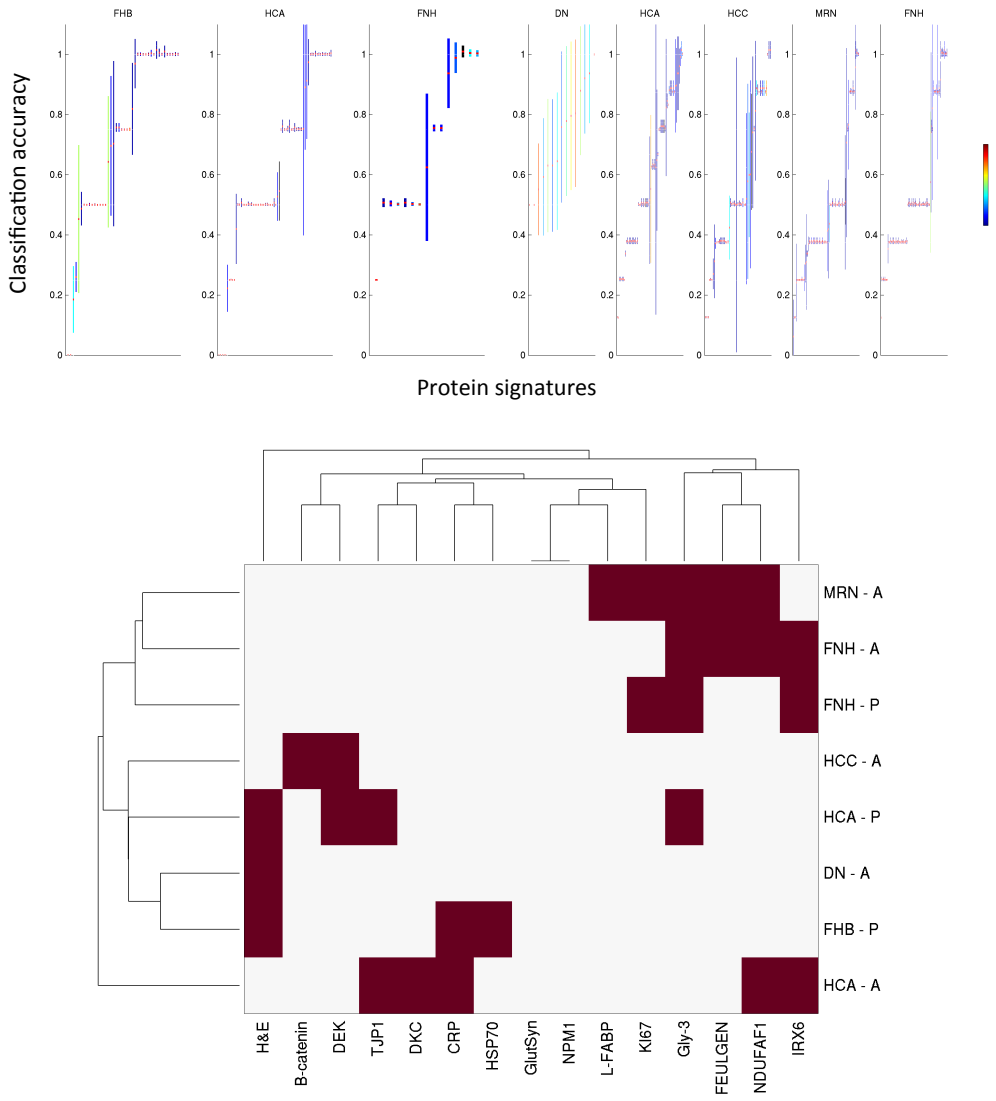


Fig. 8. Protein marker signatures for distinguishing liver lesions. The single protein multiclass class classifiers were selected to distinguish each lesion by the plurality vote. The signatures were found by cross-validation 10 times on the training set and then evaluating the performance of the signature on the held out test set. The process was repeated 100 times to result in 100 protein signatures. In many cases the same signature was found across the 100 runs. The mean accuracy and deviation in test accuracy is shown in the top panels above. The sets of proteins that resulted in perfect performance represent the optimal protein signatures. The clustergram is a binary representation of the proteins that are in the best protein signatures for each lesion. The dark tiles indicate when a protein was selected to identify a disease state. The x and y axes were clustered using Euclidean distance and optimal leaf ordering.

FHB - ped

	multiclass	Protein voting
Accuracy	**0.80**	**1.00**
Protein	**HSP70**	**CRP,H&E,HSP70**
# feats	**592**	**1776**

HCA - adult

	multiclass	Protein voting
Accuracy	**1.00**	**1.00**
Protein	**DEK**	**CRP,DKC,IRX6,NDFAF1, TJP1**
# feats	**592**	**2960**

HCA - ped

	multiclass	Protein voting
Accuracy	**0.55**	**1.00**
Protein	**NPM1**	**DEK,Gly-3,H&E,TJP1**
# feats	**592**	**2960**

HCC - adult

	multiclass	Protein voting
Accuracy	**0.60**	**1.00**
Protein	**IRX6**	**B-cat, DEK**
# feats	**592**	**1184**

FNH - ped

	multiclass	Protein voting
Accuracy	**1.00**	**1.00**
Protein	**FEULGEN**	**Gly-3,IRX6,KI67**
# feats	**592**	**1776**

MRN - adult

	multiclass	Protein voting
Accuracy	**0.35**	**1.00**
Protein	**CRP**	**FEULGEN, Gly-3, KI67, L_FABP, NDUFAF1, NPM1**
# feats	**592**	**2960**

DN - adult

	multiclass	Protein voting
Accuracy	**0.95**	**1.00**
Protein	**NDUFAF1**	**H&E**
# feats	**592**	**592**

FNH - adult

	multiclass	Protein voting
Accuracy	**0.65**	**1.00**
Protein	**IRX6**	**FEULGEN, Gly-3, IRX6, NDUFAF1**
# feats	**592**	**2368**

Table 3. Comparison of the best performing proteins and protein signatures found from the different classification schemes.

Discussion

We have extended our image analysis IHC pipeline to process full slide images and to classify multiple disease states using multiple potential protein markers. The image processing portion of the pipeline was modified to segment the tissue object and to select regions prior to unmixing. This was done to limit the search space for regions and to make the unmixing problem computationally tractable. A range of region sizes and counts were tested to make sure sufficient portions of the image were sampled and used in the classification. Features included the set of 592 from the Chapter 2 pipeline.

We constructed three different classifiers to investigate the diagnostic potential of the set of tagged proteins in our dataset. Each classifier was designed to represent a specific clinical scenario where a patient presents with a disease. The first two classifiers involved using a single protein as a marker to identify the lesions (classification using a single protein), the third classifier involved using a set of signature proteins to identify the lesions (a subset of proteins that are potentially more predictive than the proteins independently). The three classifiers we constructed were built from the random forest classifier.

In the first two classifiers we assessed how well a single protein could perform at classifying a disease. In the majority of cases a single protein gave performance above chance, but not at the accuracy necessary to translate into a clinical setting. Next we investigated how well combination of the proteins could perform. In classifier 3 we tested different combinations of multiclass classifiers and we found optimal protein signatures that consistently resulted in 100% classification accuracy with different test points.

Thus, with our system we explored a set of current protein biomarkers, and a set of new potential protein biomarkers. Our system was able to classify some of the lesions with high accuracy using protein markers that are currently used in the

clinic for lesion identification. For example: glypican.3 and NDUFAF1 were able to identify adult DN in a multiclass system (90%, 95%); DEK was able to identify adult HCA in a multiclass system (100%). When measuring protein subcellular location we found that B.catenin was not useful for distinguishing any lesions in the pediatric population with high accuracy, in contrast to the utility of B.catenin in the clinic (Yamaoka, Ohtsu et al. 2006).

Next, subsets of protein were selected as signatures to identify the lesions in a one. vs.rest classifier. For each disease we were able to identify a signature that consistently resulted in perfect classification. We found two new proteins that were previously not used in liver lesion diagnosis, IRX6 and TJP1, were part of the protein signatures that separate lesions in both age groups. TJP1 was an important protein in the signatures found for plurality voting for HCA in adult and pediatric diseases. IRX6 was part of the signatures for adult HCA and adult and pediatric FNH.

The one-vs-rest protein signatures were combined to form single multi-class voting classification systems for liver lesions for each age group. We reported these signatures and the classification structure as a potential approach for diagnosing liver lesions in the clinic.

While these results are promising, the number of patients in our dataset is small and our data comes from a single source. To show the study the utility and generalizability of our results in a clinical setting further work on a larger dataset with more patients and from different sources is necessary.

Cross-referencing results

Glypican-3

We identified Glypican-3 as a top predictor in our system for adult liver lesions. This result is in agreement with current clinical practices : Glypican-3, beta-catenin, Heppar-1 and other stains are used to differentiate disease states based on expression levels (Li, Liu et al. 2013) (Libbrecht, Severi et al. 2006) (Kandil, Leiman et al. 2007) (Wang, Anatelli et al. 2008). More specifically Glypican-3 has been reported as a useful biomarker in hepatocellular carcinoma (Shirakawa , Kuronuma et al. 2009).

Comparing results from chapters 2 and 3

NDUFAF1 was predictive of cancer in two independent liver cancer datasets. It was the 33rd out of 609 proteins in Chapter 2, and this protein was able to separate liver cancer types in Chapter 3. We have not found any reports of this protein as a marker for liver cancer. We suggest follow up studies to understand the utility of NDUFAF1 as a novel protein biomarker.

NPM1 had relatively high classification accuracy in children's cancer (90%) and adult cancer (80%). In Chapter 2 this protein ranked at the 11th percentile : 67th out of 609 proteins. We have not found reports identifying NPM1 as a marker in the context of liver cancer. Based on the agreements between our two studies we encourage further research on NPM1 in liver cancer.

DEK has high specificity for hepatocellular adenoma (HCA). In Chapter 2 DEK ranked in the lower 50th percentile as the 349th protein out of 609. We have not found literature reports suggesting the utility of this protein in HCA liver cancer. These confounding results suggest this protein may be a marker for a specific subtype of cancer and further investigation may yield valuable insights.

Glypican-3 (GPC3) was not part of our HPA analysis set.

Comparing chapter 2 staining patterns with HPA and Uniprot

We compared the subcellular location of the top 4 proteins from the analysis dataset displayed in Figure 2 with two well known bioinformatics databases : Uniprot and Human Protein Atlas (HPA). Uniprot subcellular locations are from unspecified tissues. HPA subcellular annotations are from normal liver tissue. We compared the database subcellular location information with the protein location from the dataset in Chapter 3, Table 4.

The subcellular location of NPM1 in Chapter 3 dataset include nucleus, nuclear membrane, and cytoplasmic organelles while NPM1 in HPA normal liver tissue is listed as nucleus. This difference highlights the importance of considering subcellular locations of proteins to understand, diagnose and develop cancer therapeutics. The location of NDUFAF1 and DEK agree between Chapter 3 dataset and HPA normal tissue. Glypican-3 is not available in HPA.

We will not emphasize these differences between the Chapter 3 dataset and Uniprot since the locations in Uniprot are from unspecified tissues. The differences may not be due to tissue-dependent variations and not the diseases we are studying.

	Dataset		
Protein	**Chapter 3 dataset**	**Uniprot (unspecified normal tissue)**	**HPA (normal liver)**
NDUFAF1	Cytoplasmic organelle	Organelle, membraneous, cytoplasm	Cytoplasmic, membraneous
NPM1	Nucleus, nuclear membrane, cytoplasmic organelle	Nucleus, organelle	Nuclear
DEK	Cytoplasmic, nuclear membrane	Nucleus	Cytoplasmic, membraneous, nuclear
Glypican-3	Organelle, nuclear membrane	Cell membrane, membrane.	Not available.

Table 4 : Proteins that discriminated liver cancer with subcellular location. Locations are collected from Figure 2 and two bioinformatics databases : Uniprot[1] and Human Protein Atlas[2].

Conclusion

We have extended our image analysis pipeline to process full slide images. We applied our pipeline to a dataset of pediatric and adult liver lesions to determine whether differential protein subcellular location allows us to distinguish and classify liver lesions. Further, we found a subset of proteins that can identify liver lesions perfectly given our dataset. The generalizability of this subset of proteins needs to be investigated on a larger dataset, and from images acquired from other imaging systems and institutions.

We found that some markers currently used in the clinic were highly predictive in our system, such as glypican-3 and NDUFAF1 to identify adult DN in a multiclass system; DEK

was able to identify adult HCA in a multiclass system. NPM1 was a good discriminator in pediatric cancers. In addition we found two proteins that are currently not used in the clinic to identify liver lesions: IRX6 and TJP1. Both proteins were important in the protein signatures used to identify liver lesions.

Methods

Image processing

1.) Tissue object detection

The full slide image was loaded into memory after downsampling by a factor of 2^5. To segment the tissue object from the image the areas with the maximum overlapping chromaticity and frequency were used to find the tissue regions in the image. A 300 pixel border was removed from the segmented object to account for edge effects.

2.) Region selection

Regions were selected from the downsampled tissue object based on the maximum pixel intensity that corresponds to the highest stained regions. A circular region of 10 pixels in diameter was used to scan the image as a sliding window in 5 pixel increments. The top 200 regions with the highest protein stain were selected from each tissue object. Next, regions from the original image with centers matching the 200 top downsampled regions were selected. The diameter of the regions from the original image varied at 62, 125, 312, 625, 1250, 2500 pixels. Regions were ranked in ascending order by the Brenner score (Brenner 1976). The Brenner score is a gradient.based measure of focus where larger values correspond to higher frequencies in the image and indicate greater focus. The top 150 regions were selected for the analysis.

3.) Unmixing and Features

For each region, the color basis matrix (W) was calculated and the region was linearly unmixed into the protein and DNA masks, as previously described (Newberg and Murphy 2008). 592 Murphy lab features were calculated as first described by Boland et al. (Boland, Markey et al. 1998).

Classification

4.) Training classifiers : region count, region size, model parameters, classification accuracy.

For each round of cross validation one patient from each disease group was held out, that is all of the regions from the patient belonged to the same held out group. Each region was classified independently and the patient label was assigned the plurality vote of the regions.

For each classifier the optimal model parameters, region count and region size were learned together by maximizing the training accuracy of the classifier. For example, every combination of model parameter, region size and region count (count = 50, 100, 150) was tested. The combination that gave the greatest cross.validation accuracy on the training set was selected.

In each classifier the training points were assigned weights corresponding to the inverse frequency of each class. The held out training and testing sets were always selected to have one patient from each disease group.

Random forest classifiers (Breiman 2001) were trained using Matlab's TreeBagger function. The model parameter, the number of trees in the ensemble, was learned through cross.validation.

4.1.) Single lesion classification: pairwise

A binary random forest classifier was trained to distinguish every pairwise combination of disease states within each age group. The reported accuracy is the mean of 10 rounds of cross-validation.

4.2.) Multi class single protein classifier

A single protein multi.class random forest classifier was trained to identify different disease states within each age group. The reported accuracy is the mean of 10 rounds of cross-validation.

4.3*.) Multi-protein voting classifier*

For each patient, the disease was assigned the plurality vote from the multi.class single protein classifiers trained on a portion of the training set as described in 4.2. Proteins were selected to be in the signature set if the true positive rate was greater than the false positive rate. The protein set was selected and the classifiers were retrained using the full training set. Finally, the test set was classified using the plurality vote of the signature proteins. The protein signatures and accuracies are reported from 100 rounds of cross-validation.

Chapter 4

Conclusions and future work

Summary of the chapters and significance

This thesis focused on developing an automated system to process IHC images to quantitate protein subcellular location. Protein location labels were used to discriminate cancers and to identify systems changes in the disease.

In Chapter 2, we described a robust pipeline for identifying proteins whose subcellular location undergoes statistically significant changes. We quantified changes in location for hundreds of proteins in four cancers and we presented a list of proteins ranked by their extent of location change between the normal and cancer states. Using those results we identified biochemical pathways that are enriched for proteins that translocate. Future investigation of these proteins and pathways may provide new insight into oncogenesis. Further, the analysis pipeline is expected to be useful for assessing disease type and severity in a clinical setting.

In Chapter 3 we extended the pipeline from Chapter 2. The image processing section of the pipeline was modified to process and extract features from full slide images. We also added a series of classifiers to recreate situations for identifying a patient's disease, where any of the given cancer types could be stained with up to 15 different protein specific antibodies and stains. Three of the classifiers were designed to find the optimal protein signatures for the liver lesions. We reported the optimal protein signature and the classification accuracy of the protein set. With further development, these protein signatures are expected to be useful in a clinical setting for discriminating difficult to identify liver lesions in pediatric and adult populations.

Thesis contributions

Chapter 2

1. Previous work classified the subcellular location of proteins into one of 8 classes from IHC images (Newberg and Murphy 2008) however the system did not estimate more diverse location patterns or mixtures of locations. We extended the previous pipeline to measure the extent of protein subcellular location change between two sets of images with more diverse locations, without explicit classification. Given two input sets of images the pipeline outputs 3 measurements of change: 1) a Freidman-Rafsky p-value on the null hypothesis that the feature distributions in the two sets of images are the same, 2) a Wald Wolfowitz p-value on the null that the expression distributions between the two sets of images are the same, and 3) a classification accuracy on how well the two sets of images can be discriminated based on subcellular location features.

2. We improved the performance of the pipeline by selecting cellular regions of interest from each IHC image for analysis. By selecting regions we were able to isolate cellular regions of the tissue and omit stroma, connective tissue and other non-cellular components that have minimal cross-reactivity with the antibody. Thus features were calculated on cellular parts of the image with moderate or strong protein levels.

3. We reported a ranked list of potential location biomarkers using the protein location and expression results from the pipeline. We showed the generalizability of the FR p-values and accuracies to unseen data. The list ranks proteins by the largest changes in subcellular location and the smallest changes in expression. These potential biomarkers would have been missed by traditional experiments that measure expression alone.

4. We reported biochemical signaling pathways that we predict are altered in the cancer state based on subcellular location changes of the individual

proteins. Most pathways implicated in cancer are found by changes in protein expression or mutations. We were able to identify new pathways that are implicated in cancers based on changes in protein location.

Chapter 3

5. We reported optimal signatures of proteins for liver lesions for pediatric and adult populations. The lesions within each population can be challenging to identify in a clinical setting given the current markers for the diseases. We explored different approaches for finding protein signatures and reported the optimal proteins for each disease and the respective accuracies.

6. We showed that two new proteins IRX6 and TJP1 are important for identifying liver lesions that are difficult to distinguish in the clinic. The analysis set contained sections of each disease stained with antibodies against current biomarkers, and a set of new potential biomarkers. We found that in addition to the current biomarkers IRX6 and TJP1 provided valuable information for separating the lesions.

Future work

Below we discuss a set of extensions to the projects in this thesis. The second part of this section described new projects related to these findings.

Improved region selection to study of low staining proteins

Currently we are selecting regions of interest from IHC images by finding areas with the highest levels of protein stain, under the assumption that cellular regions expressing the protein will stain more strongly than other parts of the tissue. However low staining levels may happen due to poor protein specificity, or low abundance of the protein. Changes in subcellular location of proteins with low staining levels are an unexplored part of this project.

To select regions from these types of images a more robust region selection approach is necessary. One approach is to scan the DNA mask of the unmixed image and find regions that have nuclei, as opposed to other types of tissue. Methods to identify cellular regions of based on the presence of nuclei through the classification of superpixels have been previously described (Schüffler, Fuchs et al. 2010) (Schuffler, Fuchs et al. 2013) (Beck, Sangoi et al. 2011) (Kong, Gurcan et al. 2011). Such a change can improve the performance of the pipeline on the current dataset, and also allow for the analysis of a much larger set of proteins across more diverse tissues.

Extended unmixing for lighter and darker stains

In this work we unmix the IHC images by calculating the color basis matrix for each antibody tissue combination. The matrix is found under the assumption that the peaks of the DNA and protein stains will appear at comparable levels along the hue

(H) dimension in HSV space. Further we assume that one peak will be greater than 0.3 and one will be less than 0.3 on the H axis when oriented from 0 to 1. While this assumption holds for the images in our datasets, the assumption will break down when we extend the pipeline to process lightly stained images or heavy stained images.

Some reports have tried to improve upon these methods, particularly when the staining is very weak or very strong (Tadrous 2010) (Onder, Zengin et al. 2014). It would be interesting to test these methods on more diversely stained images.

Some additional improvements may result from learning a new smoothing parameter for the hue component prior to identifying peaks, or setting a prior on the peaks in HSV space. Further, constraining the sum of the unmixed stains of each pixel to the original value can significantly improve unmixing of lightly stained images that are currently unmixing roughly equal levels of both stains.

Improved features set that acts at resolution of cells

We are using image-level features during our analysis. While this feature set has worked well in the past for distinguishing protein location (Glory, Newberg et al. 2008) (Newberg and Murphy 2008) in some cases the results from identifying changes in protein location between normal and cancer in Chapter 2 showed changes in tissue structure. It would be interesting and valuable to test the effect of local features and image level features together. Local features (Bay 2008) (Lowe 1999) (Lowe 2004) have been used in the past to create robust systems in biological and non-biological frameworks. In addition Coelho et al. (Coelho, Kangas et al. 2013) have used local features on cellular images and they have reported increased performance compared to using image-level features alone. Testing local features may improve our results and remove changes in tissue structure from ranking at the top of our potential biomarker list.

Separating cell types

In Chapter 2 we analyzed breast cancer that is a combination of ductal carcinomas and lobular carcinomas; we also analyzed liver cancers that are a combination of cholangiocarcinoma and hepatocellular carcinoma. We analyzed breast cancer and liver cancer as single diseases arising from a single cell type, however both breast cancers and liver cancers arise from different cells in their respective tissues. For instance cholangiocarcinoma arises from bile duct cells and hepatocellular carcinoma arises from hepatocytes. Currently our pipeline does not segment and classify cell types. Segmentation and classification by cell types would allow us to compare cancers directly to their respective cell types and would yield more accurate results.

Graph based approaches to identify translocated protein complexes

We have reported KEGG pathways where a significant number of the members have changed subcellular location in the cancer state. The results are an underestimate of the changes taking place in each pathway since the pathway is analyzed as a whole. It would be interesting to analyze how groups of proteins change together. These groups may represent protein complexes that play a role in the diseased state.

Larger datasets

The analysis we have presented in Chapters 2 and 3 is based on images in the HPA, and from our collaboration with UPMC. In both cases the datasets are limited and the number of patients stained with each antibody in the HPA and the number of patients representing each lesion-type in the UPMC data are limited. Further a second independent dataset collected from a different source to test our methods is not available at this time.

More data should be collected from new patients to improve the generalizability of the potential location biomarkers, and of the protein signatures for identifying liver lesions. For example, in Chapter 2 we discussed a previous study on B-catenin that reports nuclear translocation in the prostate cancer state. In our dataset none of the prostate cancer images stained with B.catenin showed translocation into the nucleus. This suggests that a subpopulation of prostate cancer patients undergo translocation, and this subpopulation is missing from our data.

In Chapter 3 some diseases were represented with as few as 1 patient, other diseases had 3 patients, and some had up to 26 patients. In the cases with few patients the disease was either removed from our analysis if the count was smaller than 3, or a very small training set was used if the number of patients was greater than or equal to 3. To report a more robust discriminating protein signature the analysis should be rerun on a large set of patients images.

Applications of deep learning

The Human Protein Atlas is a large database with patient information such as age and cancer type, and their respective IHC images. The images have human annotations for expression and subcellular location. Below are a set of experiments that apply deep learning on this dataset to improve our understanding of cancer.

1. Convolutional neural networks have properties to capture structure within images (LeCun, Kavukcuoglu et al. 2010), and this has made them effective and popular in computer vision. In this experiment a convolutional model can be applied to the dataset. The images can be grouped by organ, cancer-type or pooled as one large set to predict cancer and normal states. We expect this model to outperform the classifiers described in this thesis.

2. Deep learning multi-task multi-label (MTML) models are effective when there are two types of labels on a dataset (Huang, Wang et al 2013). In NLP MTMLs have been used to predict topic and sentiment from a single dataset simultaneously (Huang, Peng, 2013). Since each label is predicted from a common trunk, the labels promote and reinforce each other in the model.

   In the proposed experiment an MTML can predict protein expression and protein subcellular location as two distinct types of labels from IHC images. The embedding layers can be used to explore relationships between the tissues, cancer-types and patients. Information about patients can be used as features in the model, or as labels to analyze the proximity of similar aged patients in the embedding space. It is expected that this model will outperform the classifiers described in these chapters.

Biology experiments

Cancer protein studies have typically focused on understanding changes in expression. In this thesis we showed that changes in protein subcellular location are important factors in understanding disease onset and development.

Cell based experiments can lead to insights about the applicability of these results. A suitable experiment design will consist of a cell line that matches the cancer type of the protein target we are testing. To understand the effect of the translocated protein in cancer, an antibody can be used to bind to the protein and block its function. Measurements of cell growth and similar proxies can be an early indication of the validity of these protein targets.

References

Amin, W., M. Srinivasan, S. Y. Song, A. V. Parwani and M. J. Becich (2014). "Use of automated image analysis in evaluation of Mesothelioma Tissue Microarray (TMA) from National Mesothelioma Virtual Bank." Pathol Res Pract **210**(2): 79.82.

Andreadi, C., C. Noble, B. Patel, H. Jin, M. M. Aguilar Hernandez, K. Balmanno, S. J. Cook and C. Pritchard (2012). "Regulation of MEK/ERK pathway output by subcellular localization of B.Raf." Biochem Soc Trans **40**(1): 67.72.

Arihiro, K., S. Umemura, M. Kurosumi, T. Moriya, T. Oyama, H. Yamashita, Y. Umekita, Y. Komoike, C. Shimizu, H. Fukushima, H. Kajiwara and F. Akiyama (2007). "Comparison of evaluations for hormone receptors in breast carcinoma using two manual and three automated immunohistochemical assays." Am J Clin Pathol **127**(3): 356.365.

Barron, D. A. and J. D. Kagey (2014). The role of the Hippo pathway in human disease and tumorigenesis. Clin Transl Med. **3:** 25.

Bauer, T. W., L. Schoenfield, R. J. Slaw, L. Yerian, Z. Sun and W. H. Henricks (2013). "Validation of whole slide imaging for primary diagnosis in surgical pathology." Arch Pathol Lab Med **137**(4): 518.524.

Bay, H., Ess, A., Tuytelaars, T., Gool, L.V. (2008). "Speeded.Up Robust Features (SURF)." Computer Vision and Image Understanding **110**(3): 346.359.

Beck, A. H., A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn and D. Koller (2011). "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival." Sci Transl Med **3**(108): 108ra113.

Biankin, A. V., N. Waddell, K. S. Kassahn, M. C. Gingras, L. B. Muthuswamy, A. L. Johns, D. K. Miller, P. J. Wilson, A. M. Patch, J. Wu, D. K. Chang, M. J. Cowley, B. B. Gardiner, S. Song, I.

Harliwong, S. Idrisoglu, C. Nourse, E. Nourbakhsh, S. Manning, S. Wani, M. Gongora, M. Pajic, C. J. Scarlett, A. J. Gill, A. V. Pinho, I. Rooman, M. Anderson, O. Holmes, C. Leonard, D. Taylor, S. Wood, Q. Xu, K. Nones, J. L. Fink, A. Christ, T. Bruxner, N. Cloonan, G. Kolle, F. Newell, M. Pinese, R. S. Mead, J. L. Humphris, W. Kaplan, M. D. Jones, E. K. Colvin, A. M. Nagrial, E. S. Humphrey, A. Chou, V. T. Chin, L. A. Chantrill, A. Mawson, J. S. Samra, J. G. Kench, J. A. Lovell, R. J. Daly, N. D. Merrett, C. Toon, K. Epari, N. Q. Nguyen, A. Barbour, N. Zeps, N. Kakkar, F. Zhao, Y. Q. Wu, M. Wang, D. M. Muzny, W. E. Fisher, F. C. Brunicardi, S. E. Hodges, J. G. Reid, J. Drummond, K. Chang, Y. Han, L. R. Lewis, H. Dinh, C. J. Buhay, T. Beck, L. Timms, M. Sam, K. Begley, A. Brown, D. Pai, A. Panchal, N. Buchner, R. De Borja, R. E. Denroche, C. K. Yung, S. Serra, N. Onetto, D. Mukhopadhyay, M. S. Tsao, P. A. Shaw, G. M. Petersen, S. Gallinger, R. H. Hruban, A. Maitra, C. A. Iacobuzio.Donahue, R. D. Schulick, C. L. Wolfgang, R. A. Morgan, R. T. Lawlor, P. Capelli, V. Corbo, M. Scardoni, G. Tortora, M. A. Tempero, K. M. Mann, N. A. Jenkins, P. A. Perez.Mancera, D. J. Adams, D. A. Largaespada, L. F. A. Wessels, A. G. Rust, L. D. Stein, D. A. Tuveson, N. G. Copeland, E. A. Musgrove, A. Scarpa, J. R. Eshleman, T. J. Hudson, R. L. Sutherland, D. A. Wheeler, J. V. Pearson, J. D. McPherson, R. A. Gibbs and S. M. Grimmond (2012). "Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes." Nature **491**(7424): 399.405.

Boland, M. V., M. K. Markey and R. F. Murphy (1998). "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images." Cytometry **33**(3): 366.375.

Breiman, L. (2001). "Random Forests." *Machine)Learning* **45** (1): 5–32.

Brenner, J. F., Dew, B.S., Horton, J.B., King, T., Neurath, P.W., Selles, W.D. (1976). "An automated microscope for cytologic research a preliminary evaluation." Journal of Histochemistry and Cytochemistry **24**(1): 100.111.

Camp, R. L., G. G. Chung and D. L. Rimm (2002). "Automated subcellular localization and quantification of protein expression in tissue microarrays." Nat Med **8**(11): 1323.1327.

Chitu, V., P. J. Ferguson, R. de Bruijn, A. J. Schlueter, L. A. Ochoa, T. J. Waldschmidt, Y. G. Yeung and E. R. Stanley (2009). "Primed innate immunity leads to autoinflammatory disease in PSTPIP2.deficient cmo mice." Blood **114**(12): 2497. 2505.

Chung, G. G., E. Provost, E. P. Kielhorn, L. A. Charette, B. L. Smith and D. L. Rimm (2001). "Tissue microarray analysis of beta.catenin in colorectal cancer shows nuclear phospho.beta.catenin is associated with a better prognosis." Clin Cancer Res **7**(12): 4013.4020.

Coelho, L. P., J. D. Kangas, A. W. Naik, E. Osuna.Highley, E. Glory.Afshar, M. Fuhrman, R. Simha, P. B. Berget, J. W. Jarvik and R. F. Murphy (2013). "Determining the subcellular location of new proteins from microscope images using local features." Bioinformatics **29**(18): 2343.2349.

Coons AH Creech HJ, J. R. (1941). "Immunological properties of an antibody containing a fluorescent group." Proc Soc Exp Biol Med **47**: 200.202.

Dai, Y., Z. Wei, C. F. Sephton, D. Zhang, D. H. Anderson and D. D. Mousseau (2007). "Haloperidol induces the nuclear translocation of phosphatidylinositol 3'.kinase to disrupt Akt phosphorylation in PC12 cells." J Psychiatry Neurosci **32**(5): 323.330.

Dobson, L., C. Conway, A. Hanley, A. Johnson, S. Costello, A. O'Grady, Y. Connolly, H. Magee, D. O'Shea, M. Jeffers and E. Kay (2010). "Image analysis as an adjunct to manual HER.2 immunohistochemical review: a diagnostic tool to standardize interpretation." Histopathology **57**(1): 27.38.

Dolores López.Terrada, R. A., Maria T de Dávila, Piotr Czauderna, Eiso Hiyama, Howard Katzenstein, Ivo Leuschner, Marcio Malogolowkin, Rebecka Meyers, Sarangarajan Ranganathan, Yukichi Tanaka, Gail Tomlinson, Monique Fabrè, Arthur Zimmermann and Milton J Finegold (2013). "Towards an international pediatric liver tumor consensus

classification: proceedings of the Los Angeles COG liver tumors symposium." Modern Pathology.

Edgington, N. P. and B. Futcher (2001). "Relationship between the function and the location of G1 cyclins in S. cerevisiae." J Cell Sci **114**(Pt 24): 4599.4611.

Esgiar, A. N., R. N. Naguib, B. S. Sharif, M. K. Bennett and A. Murray (2002). "Fractal analysis in the detection of colonic cancer images." IEEE Trans Inf Technol Biomed **6**(1): 54.58.

Esmeralda Celia Marginean, A. M. G., Dhanpat Jain (2013). "Diagnostic Approach to Hepatic Mass Lesions and Role of Immunohistochemistry." **6**(2): 333–365.

Ferrell, L. (1995). "Malignant liver tumors that mimic benign lesions: analysis of five distinct lesions." Semin Diagn Pathol **12**(1): 64.76.

Fitzgerald, L. M., A. Kumar, E. A. Boyle, Y. Zhang, L. M. McIntosh, S. Kolb, M. Stott. Miller, T. Smith, D. M. Karyadi, E. A. Ostrander, L. Hsu, J. Shendure and J. L. Stanford (2013). "Germline missense variants in the BTNL2 gene are associated with prostate cancer susceptibility." Cancer Epidemiol Biomarkers Prev **22**(9): 1520.1528.

Fowler, C. B., Y. G. Man, S. Zhang, T. J. O'Leary, J. T. Mason and R. E. Cunningham (2011). "Tissue microarrays: construction and uses." Methods Mol Biol **724**: 23.35.

Geda, P., S. Patury, J. Ma, N. Bharucha, C. J. Dobry, S. K. Lawson, J. E. Gestwicki and A. Kumar (2008). "A small molecule.directed approach to control protein localization and function." Yeast **25**(8): 577.594.

Ghaznavi, F., A. Evans, A. Madabhushi and M. Feldman (2013). "Digital imaging in pathology: whole.slide imaging and beyond." Annu Rev Pathol **8**: 331.359.

Glory, E., J. Newberg and R. F. Murphy (2008). "Automated comparison of protein subcellular location patterns between images of normal and cancerous tissues." Proc IEEE Int Symp Biomed Imaging **ISBI 2008**: 304.307.

Guillaud, M., K. Adler.Storthz, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekirk, D. Cox, N. Poulin, M. Follen and C. Macaulay (2005). "Subvisual chromatin changes in cervical epithelium measured by texture image analysis and correlated with HPV." Gynecol Oncol **99**(3 Suppl 1): S16.23.

Gurcan, M. N., L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener (2009). "Histopathological image analysis: a review." IEEE Rev Biomed Eng **2**: 147. 171.

Hara, M. R., N. Agrawal, S. F. Kim, M. B. Cascio, M. Fujimuro, Y. Ozeki, M. Takahashi, J. H. Cheah, S. K. Tankou, L. D. Hester, C. D. Ferris, S. D. Hayward, S. H. Snyder and A. Sawa (2005). "S.nitrosylated GAPDH initiates apoptotic cell death by nuclear translocation following Siah1 binding." Nat Cell Biol **7**(7): 665.674.

Hoque, A., S. M. Lippman, I. V. Boiko, E. N. Atkinson, N. Sneige, A. Sahin, D. M. Weber, S. Risin, M. D. Lagios, R. Schwarting, W. J. Colburn, K. Dhingra, M. Follen, G. J. Kelloff, C. W. Boone and W. N. Hittelman (2001). "Quantitative nuclear morphometry by image analysis for prediction of recurrence of ductal carcinoma in situ of the breast." Cancer Epidemiol Biomarkers Prev **10**(3): 249.259.

Howe, L. R. and P. H. Brown (2011). "Targeting the HER/EGFR/ErbB family to prevent breast cancer." Cancer Prev Res (Phila) **4**(8): 1149.1157.

Htun, H., J. Barsony, I. Renyi, D. L. Gould and G. L. Hager (1996). "Visualization of glucocorticoid receptor translocation and intranuclear organization in living cells with a green fluorescent protein chimera." Proc Natl Acad Sci U S A **93**(10): 4845. 4850.

Huang, S., Peng, W., Li, J., & Lee, D. (2013). Sentiment and topic analysis on social media: a multi-task multi-label classification approach. *WebSci*.

Huang, Y., Wang, W., Wang, L., & Tan, T. (2013). Multi-task deep neural network for multi-label learning. *2013 IEEE International Conference on Image Processing*, 2897-2900.

Hung, M. C. and W. Link (2011). "Protein localization in disease and therapy." J Cell Sci **124**(Pt 20): 3381.3392.

Isaacs, H., Jr. (2007). "Fetal and neonatal hepatic tumors." J Pediatr Surg **42**(11): 1797.1803.

Ismail, H. A., L. Lessard, A. M. Mes.Masson and F. Saad (2004). "Expression of NF. kappaB in prostate cancer lymph node metastases." Prostate **58**(3): 308.313.

Ito, H., S. Funahashi, N. Yamauchi, J. Shibahara, Y. Midorikawa, S. Kawai, Y. Kinoshita, A. Watanabe, Y. Hippo, T. Ohtomo, H. Iwanari, A. Nakajima, M. Makuuchi, M. Fukayama, Y. Hirata, T. Hamakubo, T. Kodama, M. Tsuchiya and H. Aburatani (2006). "Identification of ROBO1 as a novel hepatocellular carcinoma antigen and a potential therapeutic and diagnostic target." Clin Cancer Res **12**(11 Pt 1): 3257. 3264.

Kallioniemi, O. P., U. Wagner, J. Kononen and G. Sauter (2001). "Tissue microarray technology for high.throughput molecular profiling of cancer." Hum Mol Genet **10**(7): 657.662.

Kandil, D., G. Leiman, M. Allegretta, W. Trotman, L. Pantanowitz, R. Goulart and M. Evans (2007). "Glypican.3 immunocytochemistry in liver fine.needle aspirates : a novel stain to assist in the differentiation of benign and malignant liver lesions." Cancer **111**(5): 316.322.

Khan, J., L. H. Saal, M. L. Bittner, Y. Chen, J. M. Trent and P. S. Meltzer (1999). "Expression profiling in cancer using cDNA microarrays." Electrophoresis **20**(2): 223.229.

Kimbro, K. S. and J. W. Simons (2006). "Hypoxia.inducible factor.1 in human breast and prostate cancer." Endocr Relat Cancer **13**(3): 739.749.

Klapczynski, M., G. D. Gagne, S. J. Morgan, K. J. Larson, B. E. LeRoy, E. A. Blomme, B. F. Cox and E. W. Shek (2012). Computer.assisted imaging algorithms facilitate histomorphometric quantification of kidney damage in rodent renal failure models. J Pathol Inform. **3**.

Kleinfeld, D., A. Bharioke, P. Blinder, D. D. Bock, K. L. Briggman, D. B. Chklovskii, W. Denk, M. Helmstaedter, J. P. Kaufhold, W. C. A. Lee, H. S. Meyer, K. D. Micheva, M. Oberlaender, S. Prohaska, R. C. Reid, S. J. Smith, S. Takemura, P. S. Tsai and B. Sakmann (2011). "Large.Scale Automated Histology in the Pursuit of Connectomes." J Neurosci **31**(45): 16125.16138.

Kong, H., M. Gurcan and K. Belkacem.Boussaid (2011). "Partitioning histopathological images: an integrated framework for supervised color.texture segmentation and cell splitting." IEEE Trans Med Imaging **30**(9): 1661.1677.

Kononen, J., L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter and O. P. Kallioniemi (1998). "Tissue microarrays for high.throughput molecular profiling of tumor specimens." Nat Med **4**(7): 844. 847.

Kozlowski, C., S. Jeet, J. Beyer, S. Guerrero, J. Lesch, X. Wang, J. Devoss and L. Diehl (2013). "An entirely automated method to score DSS.induced colitis in mice by digital image analysis of pathology slides." Dis Model Mech **6**(3): 855.865.

Kumar, A., A. Rao, S. Bhavani, J. Y. Newberg and R. F. Murphy (2014). "Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers." Proc Natl Acad Sci U S A **111**(51): 18249.18254.

Lambert, L. A., A. R. Whyteside, A. J. Turner and B. A. Usmani (2008). "Isoforms of endothelin.converting enzyme.1 (ECE.1) have opposing effects on prostate cancer cell invasion." Br J Cancer **99**(7): 1114.1120.

Lau, J. F., J. P. Parisien and C. M. Horvath (2000). "Interferon regulatory factor subcellular localization is determined by a bipartite nuclear localization signal in the DNA.binding domain and interaction with cytoplasmic retention factors." Proc Natl Acad Sci U S A **97**(13): 7278.7283.

LeCun, Yann ; Kavukcuoglu, Koray ; Farabet, Clément (2010). "Convolutional networks and applications in vision." ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems. pp. 253-256

Lejeune, M., J. Jaen, L. Pons, C. Lopez, M. T. Salvado, R. Bosch, M. Garcia, P. Escriva, J. Baucells, X. Cugat and T. Alvaro (2008). "Quantification of diverse subcellular immunohistochemical markers with clinicobiological relevancies: validation of a new computer.assisted image analysis procedure." J Anat **212**(6): 868.878.

Lessard, L., P. I. Karakiewicz, P. Bellon.Gagnon, M. Alam.Fahmy, H. A. Ismail, A. M. Mes.Masson and F. Saad (2006). "Nuclear localization of nuclear factor.kappaB p65 in primary prostate tumors is highly predictive of pelvic lymph node metastases." Clin Cancer Res **12**(19): 5741.5745.

Leung, F., E. P. Diamandis and V. Kulasingam (2012). "From bench to bedside: discovery of ovarian cancer biomarkers using high.throughput technologies in the past decade." Biomark Med **6**(5): 613.625.

Li, B., H. Liu, H. Shang, P. Li, N. Li and H. Ding (2013). "Diagnostic value of glypican.3 in alpha fetoprotein negative hepatocellular carcinoma patients." Afr Health Sci **13**(3): 703.709.

Libbrecht, L., T. Severi, D. Cassiman, S. Vander Borght, J. Pirenne, F. Nevens, C. Verslype, J. van Pelt and T. Roskams (2006). "Glypican.3 expression distinguishes small hepatocellular carcinomas from cirrhosis, dysplastic nodules, and focal nodular hyperplasia.like nodules." Am J Surg Pathol **30**(11): 1405.1411.

Litten, J. B. and G. E. Tomlinson (2008). "Liver tumors in children." Oncologist **13**(7): 812.820.

Lowe, D. G. (1999). "Object recognition from local scale.invariant features." The Proceedings of the Seventh IEEE International Conference on Computer Vision **2**: 1150 . 1157

Lowe, D. G. (2004). "Distinctive image features from scale.invariant keypoints." *International)Journal)of)Computer)Vision* **60**(2): 91.110.

Luo, J. H., B. Ren, S. Keryanov, G. C. Tseng, U. N. Rao, S. P. Monga, S. Strom, A. J. Demetris, M. Nalesnik, Y. P. Yu, S. Ranganathan and G. K. Michalopoulos (2006). "Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas." Hepatology **44**(4): 1012.1024.

LV, R. J. a. A. (2004). *Surgical)pathology* Mosby, Edinburgh. Lyon, R. C., S. Lange and F. Sheikh (2013). "Breaking down protein degradation mechanisms in cardiac muscle." Trends Mol Med **19**(4): 239.249.

Mahapatra, S., E. W. Klee, C. Y. Young, Z. Sun, R. E. Jimenez, G. G. Klee, D. J. Tindall and K. V. Donkena (2012). "Global methylation profiling for risk prediction of prostate cancer." Clin Cancer Res **18**(10): 2882.2895.

Mardis, E. R. and R. K. Wilson (2009). "Cancer genome sequencing: a review." Hum Mol Genet **18**(R2): R163.168.

Massoner, P., K. G. Kugler, K. Unterberger, R. Kuner, L. A. Mueller, M. Falth, G. Schafer, C. Seifarth, S. Ecker, I. Verdorfer, A. Graber, H. Sultmann and H. Klocker (2013). "Characterization of transcriptional changes in ERG rearrangement.positive prostate cancer identifies the regulation of metabolic sensors such as neuropeptide Y." PLoS One **8**(2): e55207.

Matos, L. L., D. C. Trufelli, M. G. de Matos and M. A. da Silva Pinhal (2010). "Immunohistochemistry as an important tool in biomarkers detection and clinical practice." Biomark Insights **5**: 9.20.

Matsuoka, M. and K. T. Jeang (2007). "Human T.cell leukaemia virus type 1 (HTLV. 1) infectivity and cellular transformation." Nat Rev Cancer **7**(4): 270.280.

Maulik, N., R. M. Engelman, J. A. Rousou, J. E. Flack, 3rd, D. Deaton and D. K. Das (1999). "Ischemic preconditioning reduces apoptosis by upregulating anti.death gene Bcl.2." Circulation **100**(19 Suppl): II369.375.

Mawdesley.Thomas, L. E. and P. Healey (1969). "Automated analysis of cellular change in histological sections." Science **163**(3872): 1200.

Muller, S., L. Ronfani and M. E. Bianchi (2004). "Regulated expression and subcellular localization of HMGB1, a chromatin protein with a cytokine function." J Intern Med **255**(3): 332.343.

Mulrane, L., E. Rexhepaj, S. Penney, J. J. Callanan and W. M. Gallagher (2008). "Automated image analysis in histopathology: a valuable tool in medical diagnostics." Expert Rev Mol Diagn **8**(6): 707.725.

Nakopoulou, L., E. Mylona, I. Papadaki, N. Kavantzas, I. Giannopoulou, S. Markaki and A. Keramopoulos (2006). "Study of phospho.beta.catenin subcellular distribution in invasive breast carcinomas in relation to their phenotype and the clinical outcome." Mod Pathol

19(4): 556.563.

Nativ, N. I., A. I. Chen, G. Yarmush, S. D. Henry, J. H. Lefkowitch, K. M. Klein, T. J. Maguire, R. Schloss, J. V. Guarrera, F. Berthiaume and M. L. Yarmush (2014). "Automated image analysis method for detecting and quantifying macrovesicular steatosis in hematoxylin and eosin.stained histology images of human livers." Liver Transpl **20**(2): 228.236.

Newberg, J. and R. F. Murphy (2008). "A framework for the automated analysis of subcellular patterns in human protein atlas images." J Proteome Res **7**(6): 2300. 2308.

O'Neill, E. M., A. Kaffman, E. R. Jolly and E. K. O'Shea (1996). "Regulation of PHO4 nuclear localization by the PHO80.PHO85 cyclin.CDK complex." Science **271**(5246): 209.212.

Onder, D., S. Zengin and S. Sarioglu (2014). "A review on color normalization and color deconvolution methods in histopathology." Appl Immunohistochem Mol Morphol **22**(10): 713.719.

Pratt, W. B. (1992). "Control of steroid receptor function and cytoplasmic.nuclear transport by heat shock proteins." Bioessays **14**(12): 841.848.

Ramos.Vara, J. A. and M. A. Miller (2014). "When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry..the red, brown, and blue technique." Vet Pathol **51**(1): 42.87.

Reis-Filho, J. S., B. Weigelt, D. Fumagalli and C. Sotiriou (2010). "Molecular profiling: moving away from tumor philately." Sci Transl Med **2**(47): 47ps43.

Sang Park, W., R. Ra Oh, J. Young Park, P. Joon Kim, M. Sun Shin, J. Heun Lee, H. Sug Kim, S. Hyung Lee, S. Young Kim, Y. Gyu Park, W. Gun An, H. Seung Kim, J. June Jang, N. Jin Yoo and J. Young Lee (2001). "Nuclear localization of Beta catenin is an important prognostic factor in

hepatoblastoma." Journal of Pathology **Volume 193**(4): 483.490.

Schuffler, P. J., T. J. Fuchs, C. S. Ong, P. J. Wild, N. J. Rupp and J. M. Buhmann (2013). "TMARKER: A free software toolkit for histopathological cell counting and staining estimation." J Pathol Inform **4**(Suppl): S2.

Schuffler, P. J., T. J. Fuchs, C. S. Ong, V. Roth and J. M. Buhmann (2010). Computational TMA analysis and cell nucleus classification of renal cell carcinoma. Proceedings of the 32nd DAGM conference on Pattern recognition, Springer-Verlag.

Scolyer, R. A., H. M. Shaw, J. F. Thompson, L. X. Li, M. H. Colman, S. K. Lo, S. W. McCarthy, A. A. Palmer, K. D. Nicoll, B. Dutta, E. Slobedman, G. F. Watson and J. R. Stretch (2003). "Interobserver reproducibility of histopathologic prognostic variables in primary cutaneous melanomas." Am J Surg Pathol **27**(12): 1571.1576.

Shariff, A., J. Kangas, L. P. Coelho, S. Quinn and R. F. Murphy (2010). "Automated image analysis for high.content screening and analysis." J Biomol Screen **15**(7): 726. 734.

Shelton, M. D., P. B. Chock and J. J. Mieyal (2005). "Glutaredoxin: role in reversible protein s.glutathionylation and regulation of redox signal transduction and protein translocation." Antioxid Redox Signal **7**(3.4): 348.366.

Song, J. J. and Y. J. Lee (2003). "Differential role of glutaredoxin and thioredoxin in metabolic oxidative stress.induced activation of apoptosis signal.regulating kinase 1." Biochem J **373**(Pt 3): 845.853.

Sotiriou, C. and M. J. Piccart (2007). "Taking gene.expression profiling to the clinic: when will molecular signatures become relevant to patient care?" Nat Rev Cancer **7**(7): 545.553.

Stromberg, S., M. G. Bjorklund, C. Asplund, A. Skollermo, A. Persson, K. Wester, C. Kampf, P. Nilsson, A. C. Andersson, M. Uhlen, J. Kononen, F. Ponten and A. Asplund (2007). "A

high.throughput strategy for protein profiling in cell microarrays using automated image analysis." Proteomics **7**(13): 2142.2150.

Sukru Emre, V. U., Manuel Rodriguez.Davalos "Current concepts in pediatric liver tumors." Pediatric Transplantation **16**(6): 549.563.

Tadrous, P. J. (2010). "Digital stain separation for histological images." J Microsc **240**(2): 164.172.

Uhlen, M., E. Bjorling, C. Agaton, C. A. Szigyarto, B. Amini, E. Andersen, A. C. Andersson, P. Angelidou, A. Asplund, C. Asplund, L. Berglund, K. Bergstrom, H. Brumer, D. Cerjan, M. Ekstrom, A. Elobeid, C. Eriksson, L. Fagerberg, R. Falk, J. Fall, M. Forsberg, M. G. Bjorklund, K. Gumbel, A. Halimi, I. Hallin, C. Hamsten, M. Hansson, M. Hedhammar, G. Hercules, C. Kampf, K. Larsson, M. Lindskog, W. Lodewyckx, J. Lund, J. Lundeberg, K. Magnusson, E. Malm, P. Nilsson, J. Odling, P. Oksvold, I. Olsson, E. Oster, J. Ottosson, L. Paavilainen, A. Persson, R. Rimini, J. Rockberg, M. Runeson, A. Sivertsson, A. Skollermo, J. Steen, M. Stenvall, F. Sterky, S. Stromberg, M. Sundberg, H. Tegel, S. Tourle, E. Wahlund, A. Walden, J. Wan, H. Wernerus, J. Westberg, K. Wester, U. Wrethagen, L. L. Xu, S. Hober and F. Ponten (2005). "A human protein atlas for normal and cancer tissues based on antibody proteomics." Mol Cell Proteomics **4**(12): 1920.1932.

Wang, H. L., F. Anatelli, Q. J. Zhai, B. Adley, S. T. Chuang and X. J. Yang (2008). "Glypican.3 as a useful diagnostic marker that distinguishes hepatocellular carcinoma from benign hepatocellular mass lesions." Arch Pathol Lab Med **132**(11): 1723.1728.

Webster, J. D., E. R. Simpson, A. M. Michalowski, S. B. Hoover and R. M. Simpson (2011). Quantifying Histological Features of Cancer Biospecimens for Biobanking Quality Assurance Using Automated Morphometric Pattern Recognition Image Analysis Algorithms. J Biomol Tech. **22:** 108.118.

Xu, X. and J. Chen (2009). "One.carbon metabolism and breast cancer: an epidemiological

perspective." J Genet Genomics **36**(4): 203.214.

Yamada, S., M. Ohira, H. Horie, K. Ando, H. Takayasu, Y. Suzuki, S. Sugano, T. Hirata, T. Goto, T. Matsunaga, E. Hiyama, Y. Hayashi, H. Ando, S. Suita, M. Kaneko, F. Sasaki, K. Hashizume, N. Ohnuma and A. Nakagawara (2004). "Expression profiling and differential screening between hepatoblastomas and the corresponding normal livers: identification of high expression of the PLK1 oncogene as a poor.prognostic indicator of hepatoblastomas." Oncogene **23**(35): 5901.5911.

Yamaoka, H., K. Ohtsu, T. Sueda, T. Yokoyama and E. Hiyama (2006). "Diagnostic and prognostic impact of beta.catenin alterations in pediatric liver tumors." Oncol Rep **15**(3): 551.556.

Yelamos, J., J. Farres, L. Llacuna, C. Ampurdanes and J. Martin.Caballero (2011). "PARP.1 and PARP.2: New players in tumour development." Am J Cancer Res **1**(3): 328.346.

Yuan, J., K. Luo, L. Zhang, J. Cheville and Z. Lou (2010). "USP10 Regulates p53 Localization and Stability by Deubiquitinating p53." Cell **140**(3): 384.