Supporting Information

Kumar et al. 10.1073/pnas.1415120112

Estimation of Generalizability

For each antibody, a classifier was trained by using regions from one normal image and regions from one cancer image, with the number of regions determined independently for each tissue (*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. S3), and found them to range from 0.90 to 0.91, indicating that our estimates of classification accuracies are likely good estimates of future performance.

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), and the average of 35 samplings are reported for each estimate (Fig. S3). 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.

Robustness to JPEG Compression

The images in the HPA database are ~3,000 × 3,000 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 HPA 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. S2) 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.

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. S8 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, although annotations indicate that normal and cancer images have distinct nonoverlapping location distributions, our method organizes the regions to show a progression of location change, highlighting the visually overlapping distributions for the two disease states.

Familywise 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 correction, which controls the familywise error rate when making multiple comparisons. The correction keeps the effective familywise error rate at α when there is more than one comparison. Given a set of hypotheses of size m, the corrected significance threshold for all hypothesis (H; i.e., pathways) is a function of its rank position (k) and the naive 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 $P_{(k)} > \alpha/(m + 1 - k)$.

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. S9 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 (i.e., correlation close to 1) as the number of estimates increases. This process was repeated for classification accuracy (Fig. S9, *Right*) 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.

KEGG Pathways and Translocated Proteins

Biochemical pathway networks were downloaded from the KEGG database (www.genome.jp/kegg/pathway.html) as KGML files. The files were parsed to undirected graphs in which nodes represent proteins referenced by Entrez ID numbers and edges represent interactions. The parsing into a graph structure was done in R with the use of the KEGGgraph package available from Bioconductor (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 Ensembl IDs for the proteins in the analysis set to the respective Entrez IDs and labeled the nodes in each graph with the respective location and expression scores from our analysis. This resulted in 268 KEGG pathways, for which each pathway i has n_i nodes, and m_i nodes in the network have pipeline or annotation values assigned, where $m_i \leq n_i$.

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 the logarithms 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, in which each random network had 30 known proteins, whereas 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 known protein P values in the 268 KEGG pathways. The score of pathway i was compared with 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 by using Bonferroni–Holm multiple hypothesis correction to control for familywise error rate. We then repeated the same analysis by using the expression P values from the pipeline.

Pathway P values were also calculated by 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 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, the respective probabilities were 0.43 and 0.56; in prostate, 0.26 and 0.49; and in bladder, 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. S6.

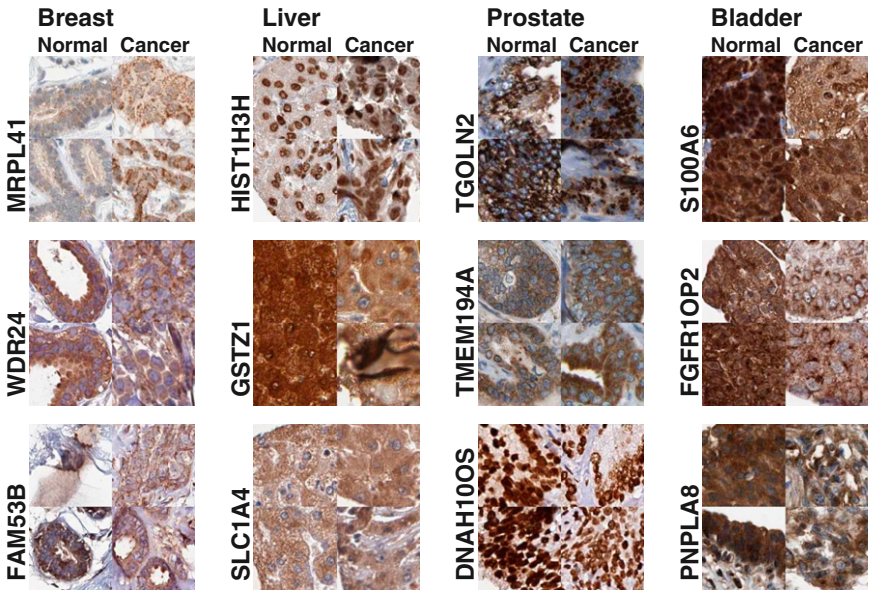

Fig. S1. Example images from top-ranked potential location biomarkers. The three proteins with the lowest location P values are shown for each tissue (without considering expression level). The two regions closest to the two centroids found from k-means clustering ($k = 2$) for the normal and cancer feature distributions are displayed for each of the top hits. Note that some of the top hits may have been detected as a result of expression changes.

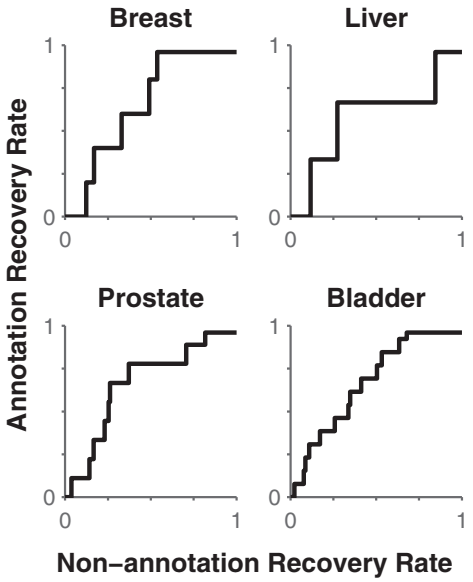

Fig. S2. Ability of the system to detect known location biomarkers. ROC curves were constructed for each tissue by determining how many true positives and false positives were found as a threshold on the *P* value was varied. The validation set for a given tissue consisted of those proteins from the 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.

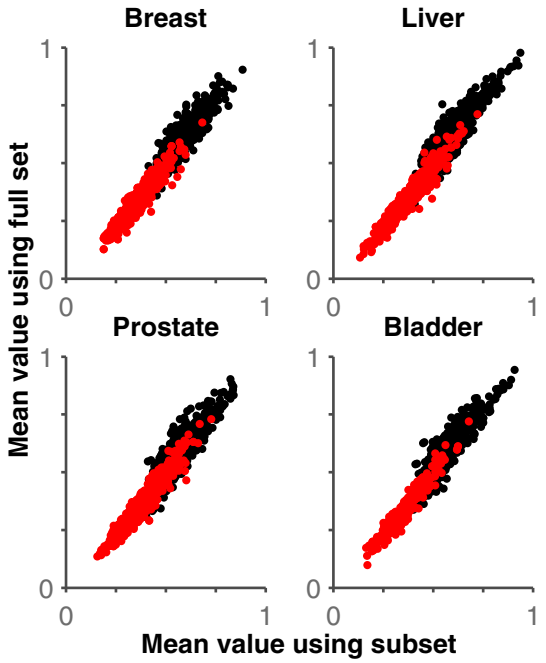

Fig. S3. Estimation of generalizability of identifying location biomarkers. For each protein, classification accuracies (black) and location biomarker rankings (red) are compared for estimates by using one or two normal images. The correlation coefficients were 0.90, 0.91, 0.90, and 0.90 for the accuracies and 0.95, 0.96, 0.96, and 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.

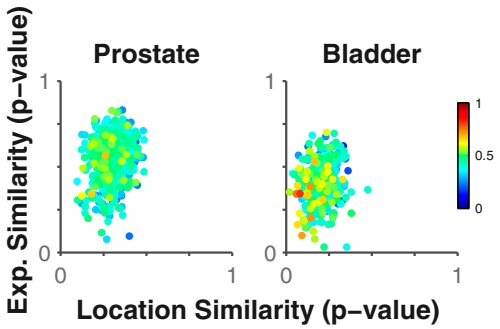

Fig. S4. Location biomarkers differ between high- and low-grade cancers. The prostate and bladder cancer images were partitioned into high- or low-grade cancer as annotated in the HPA. For each cancer, location and expression *P* values were calculated between the grades. The correlation between location 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.

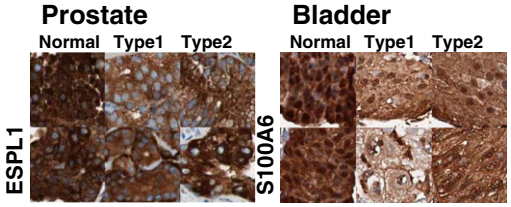

Fig. S5. Example images from top location biomarker predictions for classifying normal tissue from different tumor grades. The proteins are ranked by the three-class classification accuracy for separating normal tissue, low-grade tumors, and high-grade tumors.

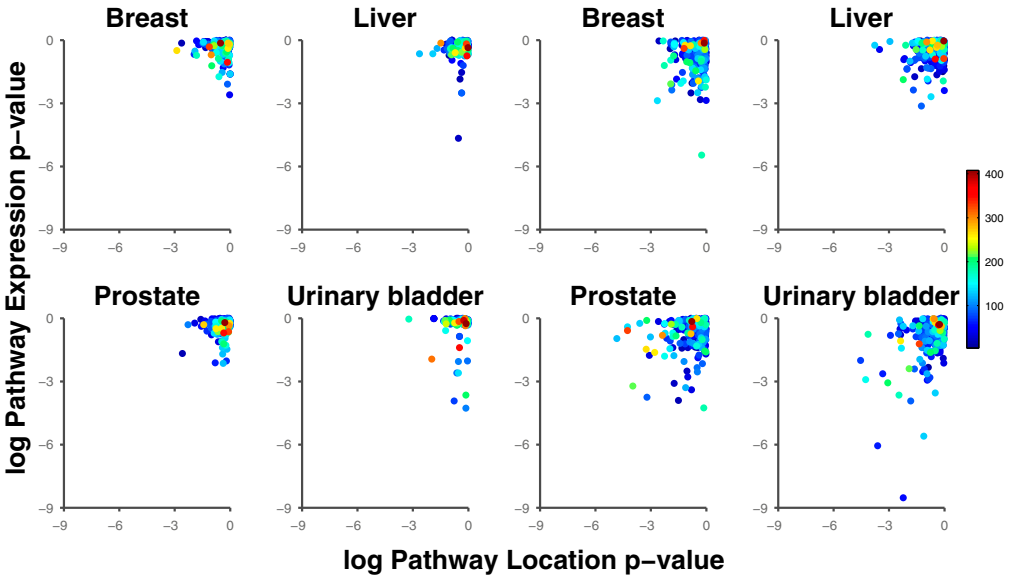

Fig. S6. Extent of expression and location change in the KEGG pathway components. The left four panels show KEGG *P* values by using pipeline protein values, and the right four panels show pathway *P* values derived from location and expression annotations. Each point represents the expression and location *P* value for a single pathway. The points are colored by the number of nodes in the pathway.

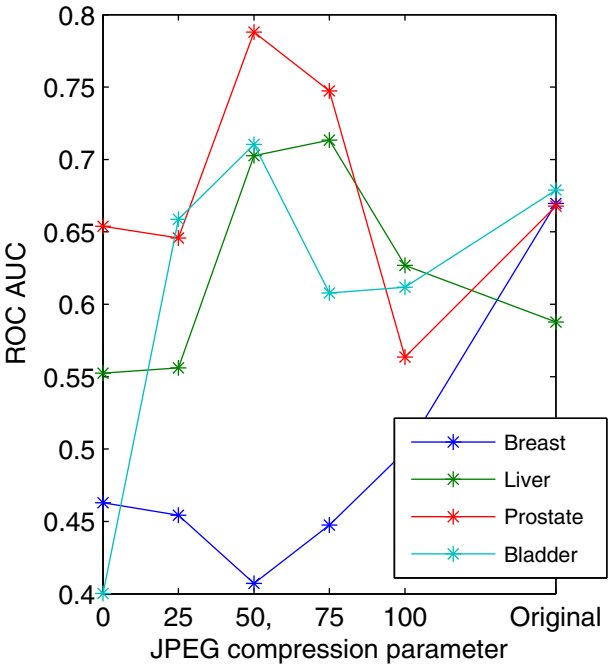

Fig. S7. The effects of JPEG compression on the performance of the pipeline. The images were JPEG compressed at different levels and processed through the pipeline. The performance for detecting known location biomarkers was measured and reported as AUC values, as was done for Fig. S2.

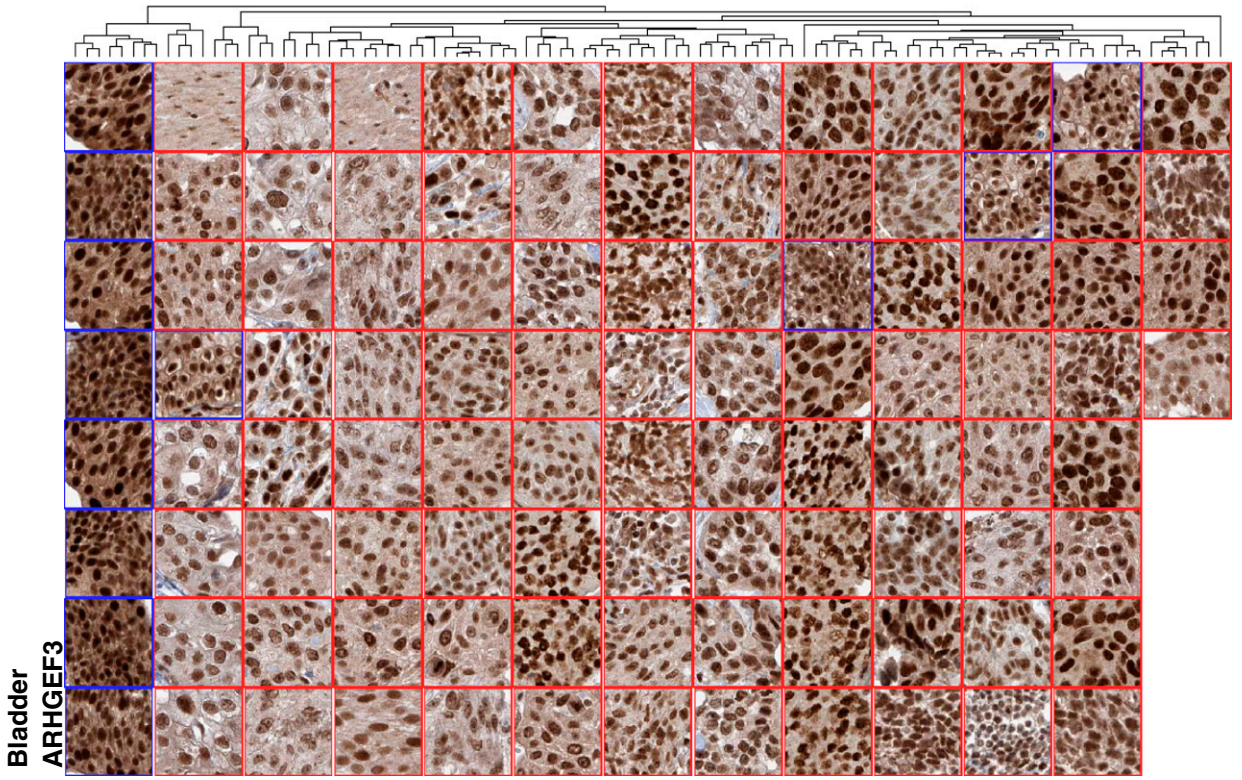

Fig. S8. Ordering regions by location change progression. We performed a hierarchical clustering of the features from the regions used in the analysis for antibody HPA034715 against ARHGEF3 in bladder tissue. We calculated the Euclidean distances between every pair of regions and then performed hierarchical clustering. The leaves are ordered to maximize the sum of similarities between adjacent leaves across the tree. Regions are displayed from top to bottom and then left to right according to the ordering in the tree. Normal tiles are outlined in blue; cancer tiles are outlined in red.

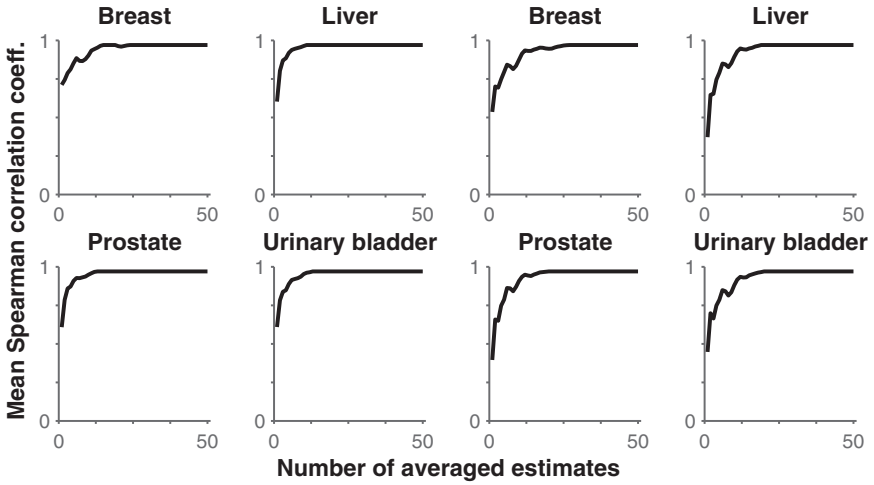

Fig. S9. Protein rank correlations using location *P* values and classification accuracies when averaging different numbers of estimates. The number of estimates used to calculate the location *P* value and accuracy was varied from 1 to 50. The Spearman correlation coefficient was used to measure the consistency of the ranked protein lists when different numbers of estimates were used.

Dataset S1. The sets of analyzed proteins for each tissue, ranked by the combination of location change and minimal expression change

Dataset S1

Validation marker proteins (*Methods*) are indicated with a 1 in the Validation Marker column.

Dataset S2. Estimation of expression and location differences between prostate and bladder tumor grades

Dataset S2

The proteins are ranked by the three-class classification accuracy for separating normal tissue, low-grade tumors, and high-grade tumors.

Dataset S3. Estimates of expression and location changes for pathways

Dataset S3

P values that the KEGG pathways changed (against a background distribution) were calculated by using the image analysis pipeline or the image annotations, both for expression and location (*Methods*). Blank entries indicate pathways in which there was an insufficient number of images for the analysis.