# Discriminative bag-of-cells for imaging-genomics

Benjamin Chidester
*Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA*
*E-mail: bchidest@cs.cmu.edu*

Minh N. Do
*Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA*
*E-mail: minhdo@illinois.edu*

Jian Ma
*Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA*
*Email: jianma@cs.cmu.edu*

Connecting genotypes to image phenotypes is crucial for a comprehensive understanding of cancer. To learn such connections, new machine learning approaches must be developed for the better integration of imaging and genomic data. Here we propose a novel approach called Discriminative Bag-of-Cells (DBC) for predicting genomic markers using imaging features, which addresses the challenge of summarizing histopathological images by representing cells with learned discriminative types, or codewords. We also developed a reliable and efficient patch-based nuclear segmentation scheme using convolutional neural networks from which nuclear and cellular features are extracted. Applying DBC on TCGA breast cancer samples to predict basal subtype status yielded a class-balanced accuracy of 70% on a separate test partition of 213 patients. As data sets of imaging and genomic data become increasingly available, we believe DBC will be a useful approach for screening histopathological images for genomic markers. Source code of nuclear segmentation and DBC are available at: https://github.com/bchidest/DBC.

*Keywords*: Imaging-genomics; Histopathological image analysis; Computational pathology

## 1. Introduction

Cancer is a genetic disease that develops from accumulated genomic alterations that disrupt normal cellular processes and give rise to phenotypic changes, such as cell size, shape, and structural relationship within a tumor.[1] High-throughput genomic approaches have revealed new understandings of the complexity of cancer. We now know that even patient samples from the same type of cancer may exhibit a high level of inter-tumor heterogeneity.[2] For example, breast cancer patients can be largely categorized into four molecular subtypes (luminal A and

B, HER2-enriched, and basal) with distinct prognosis.[3,4] Additionally, 'imaging-genomics' has been coined to refer to recent developments in leveraging new insights gained from genomics with traditional imaging of radiology or histopathology.[5–8] For histopathological images, features of cells and nuclei, which are used by pathologists to diagnose cancer, provide the most direct connection to the genomic signatures of a patient's tumor. Yet whole slide images (WSIs) contain tens of thousands of cells with diverse characteristics, which makes associating phenotype with genomics challenging. Previous works in imaging-genomics have sought to draw connections by first clustering nuclei and cells into types in an unsupervised fashion and then associating with genomic markers, such as gene expression.[5,7,8] Others have looked for connections with specific cell types, leveraging biological understanding, such as the affect of the cellularity of lymphocytes on copy number variation in tumors.[6] There remains a need for machine learning tools that can effectively capture the diversity of cellular phenotypes within and across tumors for high-throughput investigation of general image-genomic associations.

To address this need, we propose a novel, general framework for predicting arbitrary genomic markers from imaging features called Discriminative Bag-of-Cells (DBC). In this framework, cells within a histopathological image are grouped into types and the image is summarized succinctly by a histogram of cell types, and a classifier is learned from these histograms of types to predict genomic markers (e.g., mutation, gene expression, or molecular subtype). Our framework is inspired by the bag-of-words (BoW) approach, which has been successfully applied to document classification and image classification.[9] In addition to learning the BoW classifier, our method also learns the cell-type, or codeword, assignments in a discriminative fashion to find types that are more informative of the specific genomic marker. This avoids the trouble of unsupervised clustering of cellular features,[5,7,8] which does not optimize cell-type assignments for genomic markers of interest. Some works have also proposed a similar training of discriminative codebooks,[10,11] which inspired this particular enhancement for our BoW framework for histopathological images.

DBC has several advantages over other methods of high-throughput histopathological image analysis. A primary advantage is the interpretability of the learned model. From the learned cell types, we can distinguish what types of nuclei are important for classification. This allows us to trace back to the original nuclei in the sample to visualize how they are categorized and to learn which types of nuclei in the images are discriminative for specific genomic markers. Another advantage of DBC is that it can learn what heterogeneity within a tumor is informative of a particular genomic marker and what is not. For example, it is known that a tumor can be comprised of multiple molecular subtypes,[2] but there is still a need to understand what cellular phenotypes are unique to which subtype. DBC seeks to account for such diversity by learning how the mixing of cell types is informative of an assigned subtype.

Our framework relies upon extraction of nuclear features from histopathological images, from which the cell types are learned. For hematoxylin and eosin (H&E) histopathological imaging, due to the high degree of variation in sample acquisition, such as stain intensity and slice thickness, and the care required to produce quality samples, segmenting nuclei and extracting image features is often unreliable. Most studies have relied upon popular microscopy image analysis software, such as CellProfiler,[12] to perform image feature extraction,[13] but the

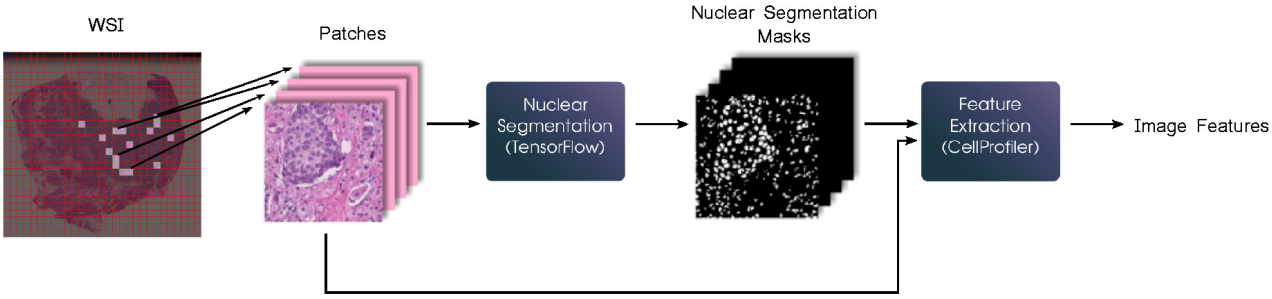


Fig. 1. Overview of our method for combined nuclear segmentation and feature extraction from WSI tiles.

commonly used available algorithms fail to generalize adequately to the significant variation of most large-scale histopathological image data sets. Recently, several methods for nuclear detection and segmentation have been developed using deep learning and convolutional neural networks (CNNs),[14–17] which have shown improvements over traditional methods. For our framework, we developed our own patch-based CNN for nuclear segmentation, which we optimized for computational efficiency and trained using additional image data from TCGA. Additionally, unlike CellProfiler, our patch-based CNN method requires no parameter tuning, which allows for easy analysis by non-specialists, which would help facilitate high-throughput imaging studies. The contribution of this work is to apply deep-learning-based nuclear segmentation in conjunction with nuclear feature extraction to show its utility for high-throughput histopathological image analysis, specifically for predicting genomic markers.

As a proof of principle, we applied our method to breast cancer patient samples from TCGA[4] to detect the basal subtype. For this task, the learned model achieved a class-balanced accuracy of 70% (i.e., sensitivity and specificity scores were both 0.7) on a separate test partition of 213 patients, of which 39 patients were of the basal subtype, which is a significant improvement over the standard method of summarizing images by the statistics of the distribution of cellular features. The algorithm also learned eight cell types relevant to the basal subtype. To our knowledge, this is the first work that attempts to predict molecular subtype solely from histopathological images. With improved nuclear segmentation and increased access to imaging-genomic data sets, DBC has the potential to be a useful tool for cancer screening of genomic markers.

## 2. Methods

### 2.1. *Overview*

To represent an H&E image as a histogram of nuclear words, we first extract nuclear and cellular features. These features are derived from the segmentation result of a patch-based CNN scheme. From the segmentation, CellProfiler computes cellular features of shape, texture, and color. The overall method for nuclear segmentation and feature extraction is shown in Fig. 1. With each cell in the image represented by its extracted feature vector, DBC jointly learns a cell-type assignment and a BoW classifier to predict a genomic marker of interest.

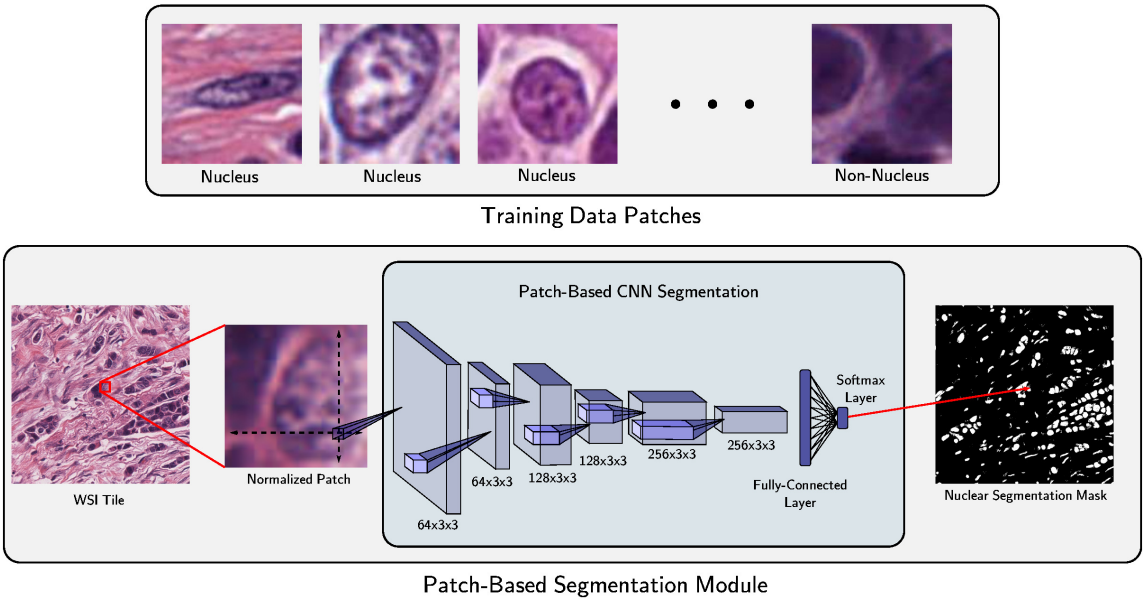


Fig. 2. Example nucleus and non-nucleus patches (top) from the training data for our CNN. Once trained, our CNN (bottom) classifies each local patch of a WSI tile individually, yielding a full nuclear segmentation mask for the tile.

### 2.2. *WSI Tiling*

Though DBC can be applied to any H&E image, whether it be a biopsy or a WSI, to process WSIs specifically, we divide them into non-overlapping tiles to be processed individually. From the WSIs, representative tiles of cancerous tissue regions that are free from artifacts, such as out-of-focus regions and tissue folds, are selected. Although this is performed mostly manually now, deep learning algorithms for segmenting cancerous regions of WSIs could be potentially incorporated.[18] For processing TCGA WSIs, we selected up to 15 representative tiles for each slide, with a tile size of 1000×1000 pixels.

### 2.3. *Nuclear Segmentation Module Based on CNN*

To achieve computationally efficient and reliable nuclear segmentation, we developed a patch-based CNN method that learns to detect nuclear pixels from the statistics of local patches. Similar to other recent patch-based CNN approaches,[17] the segmentation module produces a segmentation mask for an H&E image by generating, for each patch within the image, a binary label, indicating whether the center pixel of each patch belongs to a nucleus or not. Since nuclei only occupy a patch of a few pixels, the CNN needs to operate only on a small patch surrounding each pixel to produce its label. The training data for our network was imaged at 40× magnification, for which we chose the local patch size to be 51 × 51 pixels. Fig. 2 shows a diagram of the segmentation module operating on a WSI tile to produce a full nuclear segmentation mask.

Our network consists of six convolutional layers of $\{64, 64, 128, 128, 256, 256\}$ $3 \times 3 \times N$ filters, where $N$ is the number of filters of the previous layer, and was implemented in TensorFlow.[19] At every other layer, starting at the second layer, the convolution operator is applied

at a stride of two to gradually taper the dimension of each subsequent layer towards the fully-connected layers near the output. The final two layers are a fully-connected layer of 50 nodes and a softmax output layer of two nodes, respectively. Before being fed through the network, each tile is unmixed into its hematoxylin and eosin stains and normalized[20] to mitigate stain variation and to reduce the last dimension $N$ of the input layer filters from three to two.

For training our CNN, we used an available data set of segmented ER-positive epithelial nuclei,[17] supplemented with our own data set of nuclei patches from 68 TCGA-BRCA patients that included epithelial nuclei, stromal nuclei, and lymphocytes. For each TCGA-BRCA patient, we extracted several hundred sample patches of nuclei and non-nuclei, comprising 32,174 patches in total. The patients used in the data set represented a variety of tissue source sites to encourage the network to learn robustness to variation in sample acquisition. The pre-processing of each patch consisted of normalizing the stain,[20] separating the normalized hematoxylin and eosin images, and generating rotations at 90 degree increments, as well as horizontal and vertical flips of the images, to promote invariance to such manipulations, which naturally arise in H&E images. Several example patches collected for our data set are shown in Fig. 2.

Once the initial binary segmentation mask for each WSI tile is generated by the CNN, the mask and the corresponding H&E image are passed to CellProfiler to be further enhanced by smoothing the boundaries and separating clumped nuclei. Compared to relying solely upon CellProfiler for segmentation, the CNN-based method produces more reliable performance, as shown in Section 3.

### 2.4. *Nuclear and Cellular Feature Extraction*

After nuclear segmentation, we use CellProfiler to determine the boundaries of the cell corresponding to each nucleus. We consider only the pixels close to the nucleus boundary, but not overlapping with neighboring cells, as belonging to the cell. In our analysis, the distance we chose was 15 pixels. Features describing the shape, texture, and color of each nucleus and cell for each patient are then extracted through CellProfiler from the segmented masks and corresponding H&E images. From the nuclear and cellular segmentation boundaries computed with our CNN and CellProfiler's refinement steps, we extract a total of 219 image features for each cell-nucleus pair.

### 2.5. *Discriminative Bag-of-Cells*

Each image is then summarized by a histogram of the various cell types it contains, where the cell types are learned *discriminatively* for the specific genomic marker. We denote the extracted feature vector for a cell by $\mathbf{x} \in \mathbb{R}^d$, where $d$ is the number of extracted features. For each sample, or patient, $s$, with $N_x(s)$ segmented cells, we extract a set of $N_x(s)$ cellular feature vectors $X_s = \{\mathbf{x}_i\}_{i=1}^{N_x(s)}$. We denote the genomic marker of interest by $y \in [0, 1, \ldots, K-1]$, where $K$ is the number of possible states the marker can assume. Each patient then consists of a pair of cellular feature vectors and a marker: $(X_s, y_s)$.

Our DBC framework consists of two learners: the cell-type assignment $f_x(\cdot)$ and the BoW classifier $f_b(\cdot)$. The cell-type assignment produces the cell-type representation $\mathbf{c}_i = f_x(\mathbf{x}_i) \in$

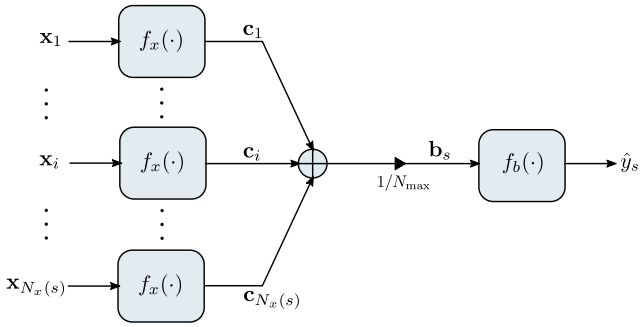


Fig. 3. Flow chart of DBC prediction on sample $s$.

$[0,1]^C$ (satisfying $\sum_j^C c_{i,j} = 1$), which acts like a soft one-hot-encoded vector assignment to one of $C$ cell types, for each cell $i$. The cell-type representations are summed over all cells, producing the BoW representation $\mathbf{b}_s$, which is normalized by $N_{\max}$, the maximal number of cells associated with any sample. In practice, this was found to help the network to learn. The final output of the DBC is the predicted marker $\hat{y}_s = f_b(\mathbf{b}_s)$. The flowchart describing DBC is shown in Fig. 3.

In the standard BoW approach, the type, or codeword, assignment $f_x(\cdot)$ is learned independent of the subsequent BoW classifier, usually by a clustering algorithm such as K-means. The drawback of the standard approach is that the assignment to codewords may not be discriminative of the classification task. However, in DBC, by constructing both learners with neural networks, we can train the combined network end-to-end, allowing $f_x(\cdot)$ to be optimized for the discriminative task at hand. To enforce that the cell-type assignment $f_x(\cdot)$ act like a soft one-hot encoding, its output layer is a softmax layer of $C$ nodes. The output layer of the BoW classifier $f_b(\cdot)$ is also a softmax layer, and the final predicted marker $\hat{y}$ is the index of the maximal value of the output softmax layer. The entire DBC framework was implemented in TensorFlow.[19]

## 3. Results

### 3.1. *Segmentation Evaluation*

We evaluated our patch-based CNN segmentation module on the UCSB Bio-Segmentation Benchmark.[21] The segmentations for this data set are pixel-wise binary masks of nuclei pixels. Boundaries between touching nuclei were not delineated on the masks, so in to order ensure objectivity and reproducibility, we inferred these boundaries automatically using CellProfiler's nuclei separation tool. Since the images were captured at a lower magnification, and not 40× magnification like the images for our training data, we resized them by a factor of 2 to approximate 40× magnification. We refer to the resulting separated nuclei masks as the gold standard. For our evaluation, a nucleus is considered a true positive if its center is within a distance of 12 pixels of the center of a nucleus in the gold standard mask. When a gold standard nucleus matches multiple predicted nuclei, its closest match is chosen.

The scores for our network, our CellProfiler pipeline, and the work of Janowczyk and Madabhushi[17] are shown in Table 1. The network of Ref. 17 produces a probability, which

Table 1. Comparative detection accuracy of our CNN, our CellProfiler pipeline, and the network Ref. 17 on UCSB breast cancer H&E images.

| Algorithm | Precision | Recall | F1-Score |
|---|---|---|---|
| Our CNN | 0.830 | 0.875 | 0.852 |
| Janowczyk and Madabhushi[17] | 0.850 | 0.850 | 0.850 |
| CellProfiler | 0.905 | 0.712 | 0.800 |

Table 2. Comparative segmentation accuracy of our CNN, our CellProfiler pipeline, and the network of Ref. 17 UCSB breast cancer H&E images.

| Algorithm | DSC Mean | DSC Median | DSC STD | HD Mean | HD Median | HD STD | MAD Mean | MAD Median | MAD STD |
|---|---|---|---|---|---|---|---|---|---|
| Our CNN | 0.72 | 0.80 | 0.20 | 4.24 | 3.00 | 3.83 | 1.82 | 1.26 | 1.43 |
| Janowczyk and Madabhushi[17] | 0.75 | 0.81 | 0.16 | 4.22 | 2.83 | 4.03 | 1.75 | 1.29 | 1.20 |
| CellProfiler | 0.76 | 0.82 | 0.18 | 4.62 | 2.83 | 4.70 | 1.84 | 1.24 | 1.64 |

allows for tuning by the user by varying the threshold to be applied to make a binary decision, so we evaluated thresholds ranging from 0 to 0.92 by increments of 0.04 and reported the the threshold with the best F1-score. It is expected that, assuming the training procedure can determine a good set of parameters, our network should perform comparably, if perhaps slightly better, than the network of Ref. 17, since we have an expanded data set with our own training patches from TCGA, and indeed, our algorithm performs comparably. What is of more importance is that compared to the results of CellProfiler, our approach shows a marked improvement.

To evaluate the quality of the resulting nuclear boundaries, we used the the Dice similarity coefficient (DSC), the Hausdorff distance (HD), and the mean absolute distance (MAD), which were adopted from other works[16] for consistency and comparison. Since these metrics can only be applied to true positive nuclei, they do not capture the effects of false negatives and false positives, so a desired balance must be considered when comparing algorithms. A comparison of the results for each of these metrics is shown in Table 2. Again, our CNN and the network of Ref. 17 perform similarly. CellProfiler performs slightly better in terms of DSC, but worse in both HD and MAD, despite it having a significantly higher precision score, which indicates that it is more conservative in what it detects as nuclei. We observed on this data set that our patch-based CNN algorithm is able to detect and segment both stromal and epithelial nuclei, but that it struggles with nuclei with fainter hematoxylin stain, which could be a consequence of the difference in resolution between the 20×UCSB images and the 40×TCGA images on which it was trained.

Data sets such as TCGA pose a much greater challenge for segmentation since the acquisition procedures across the various tissue source cites are less controlled and prone to introduce significant variation. A prominent advantage of our network is that it has been trained on WSIs from TCGA-BRCA patients, increasing its robust to these variations. We

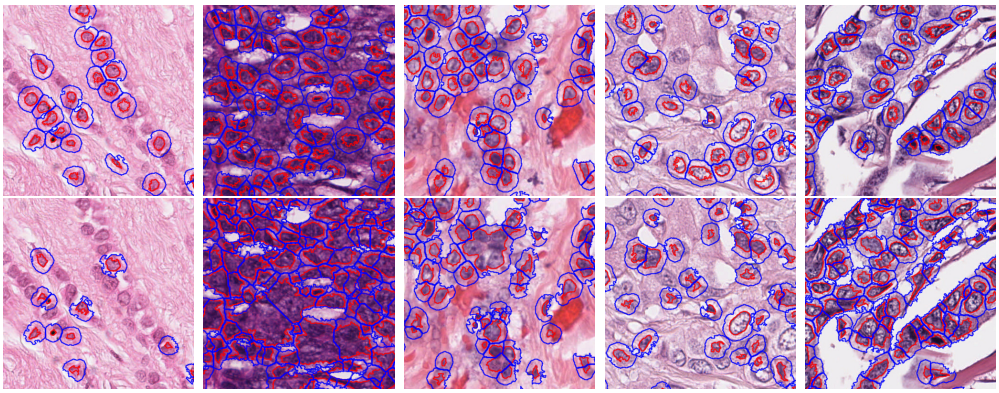

Fig. 4. Cell and nucleus segmentation results of our patch-based CNN approach (top row) and CellProfiler (bottom row) on several regions of diagnostic WSIs of the TCGA-BRCA data set. Nuclear boundaries are drawn in red and cellular boundaries in blue.

chose a small subset of 1000 × 1000 tiles from TCGA-BRCA WSIs with varying slide quality and hue on which to qualitatively compare our trained patch-based CNN approach and CellProfiler. The CellProfiler pipeline we tested consisted of adaptive thresholding to remove white background pixels, adaptive three class thresholding of the unmixed hematoxylin channel to segment nuclei, declumping of nuclei, and removal of overly large or small nuclei. We tuned the parameters to generalize the segmentation as best as possible across the test images, limiting the white pixel background threshold on the hematoxylin channel to lie between 0 to 0.3 and the adaptive three class threshold for nuclei segmentation to lie between 0.4 to 1. The yielded nuclear and cellular boundaries of both methods for regions of five images from the test set are shown in Fig. 4. Increasing the later threshold improved performance on darker images, but at the cost of missing most nuclei in lighter images. In contrast, our approach was able to detect and segment nuclei well despite the stark differences in intensity and hue. In particular, it was better able to avoid clumping large nuclei together, as seen in the two darker images. We also note that, unlike CellProfiler, our method required no parameter tuning for this test set, which is crucial for high-throughput analysis of large data sets of WSIs.

Computational cost is also an important consideration for high-throughput analysis. Segmenting each 1000×1000 tile using our approach required only an additional 5.7 seconds on average for processing with our patch-based CNN on a single Intel Xeon E5-1603 v3 (2.8GHz) CPU with 16GB of RAM and a single NVIDIA GeForce GTX 1080Ti GPU, which is a minimal overhead for the increased accuracy gained.

### 3.2. *Detecting BRCA Basal Subtype with Discriminative Bag-of-Cells*

We applied DBC on the extracted cellular features of 607 TCGA-BRCA patients to predict the status of the basal molecular subtype. The set of patients was split into 50% training, 15% validation, and 35% testing, which yielded a test set of 213 patients, of which 39 were of the basal subtype. To prevent overfitting to the training partition, during training, the accuracy of the model on the validation set was monitored and training was ended when this accuracy began to steadily decrease. Additionally, we removed all nuclear features with variance below

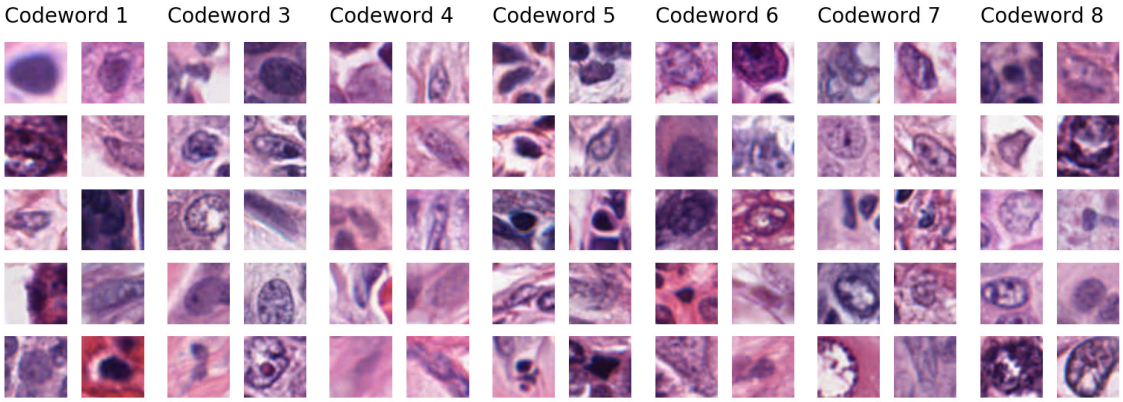


Fig. 5. Example nuclei from TCGA-BRCA patient WSIs for each of the eight learned codewords. Codeword two had no dominant example nuclei.

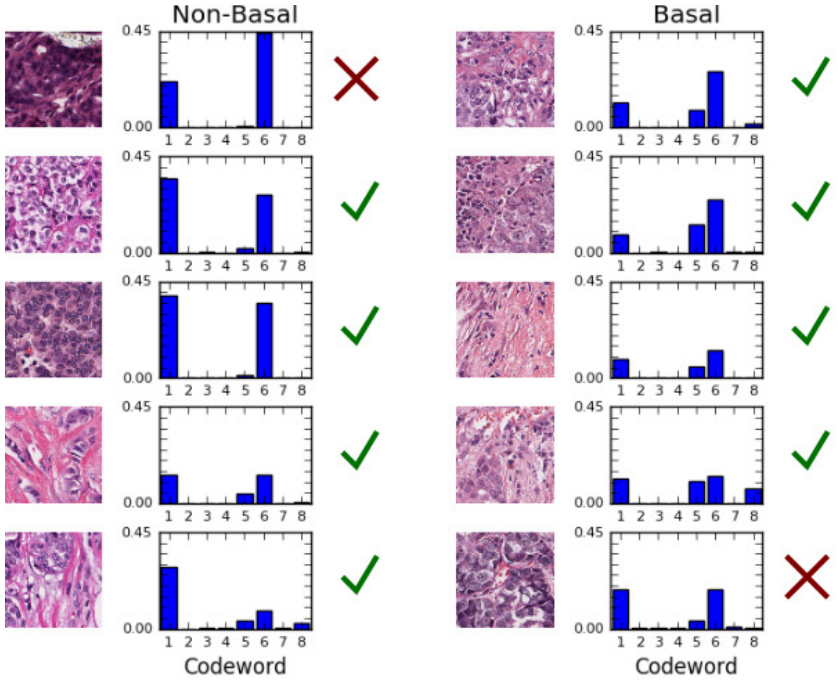


Fig. 6. BoW histograms for several example TCGA-BRCA patients. (left) Non-basal patients. (right) basal patients. Sample patches of the WSI for each patient are shown on the left of each of the two columns. The markings to the right of the BoW histograms indicate if the predicted subtype was correct.

0.001 across the training set to keep the model from fitting noise. For this particular prediction task, we trained a codeword assignment with two hidden layers of $\{25, 15\}$ nodes and a BoW classifier with just one hidden layer of six nodes. We found that a codebook of size $C = 8$ allowed for an accurate BoW classifier while minimizing overfitting. To help the model train, the nuclear features were normalized to a mean of 0.5 and a variance of 0.5.

Fig. 5 shows randomly selected example nuclei for the eight learned cell types by the codeword assignment of DBC. These are nuclei for which the maximal codeword index is the

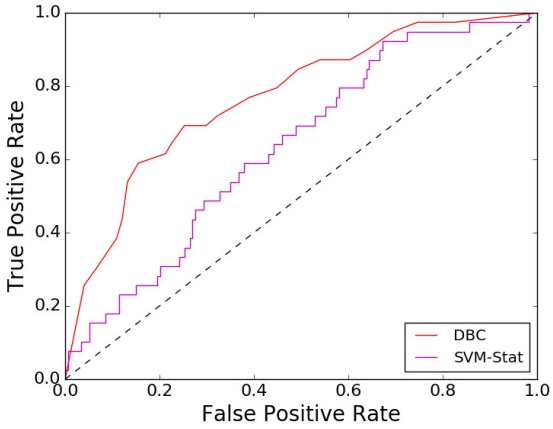


Fig. 7. ROC curves for predicting basal subtype (Positive = basal, Negative = other) on the test partition of TCGA-BRCA patients. SVM-Stat represents a standard approach to histopathological image summarization and classification, in which an image is summarized by statistics of the distribution of cellular features.

column in which they are shown, though since the codeword assignments are soft, some nuclei may be encoded as a mixture of several codewords. Patterns can be seen in the assignments, such as small, dark, lymphocytic cells and thin stromal cells, as well as the grouping of mostly hollow, highly textured nuclei. We found that codeword two never received the maximal mixture coefficient for any nuclei, which is why no examples are shown. With the knowledge of what types of cells are being mapped to which codewords, we can then investigate the BoW representations for WSIs to provide further insight. Fig. 6 shows BoW representations for the WSIs of several patients of basal and non-basal subtypes using the learned codeword assignment. The H&E images on the left of the two columns are randomly selected sample 500 $\times$ 500 pixel patches from the tiles of the WSIs of these patients which were used for feature extraction. The markings on the right indicate if the subtype of the patient was predicted correctly. A general trend for the basal subtype of higher counts of codeword five and lower counts of codewords one and six can be seen, though the true relationship lies in the weights of the learned neural network of the BoW classifier. We observe that most non-basal patients had a higher count of codeword one than codeword six. The upper left non-basal patient conversely had a higher count of codeword six and was predicted by the BoW classifier to be of the basal type. Similarly, the lower right basal patient had roughly equal counts of codewords one and six and a low count of codeword five and was predicted not to be of the basal type.

The output of our BoW classifier is a scalar value between $[0, 1]$, which can be interpreted as the probability of belonging to the basal subtype and which can be thresholded at varying levels to trade-off between the false and true positive rates. Varying this threshold, we generated the ROC curve of DBC on the separate partition of test patients shown in Fig. 7. From the ROC curve, we observe that our trained DBC is well-balanced between classifying basal and non-basal subtypes, achieving the reported class-balanced accuracy of 70% at a false positive rate of 0.3 and a true positive rate of 0.7. We compared the performance of DBC to a standard approach of summarizing histopathological images,[13] in which images are summa-

rized by statistics of the distribution of cellular features. For this comparison, we used the mean and variance, which became the feature vector for each patient to be fed into a classifier. We trained a SVM classifier with a RBF kernel using Scikit-learn[22] on this feature vector and called the overall method SVM-Stat. The ROC of this approach using the same training and testing sets is also shown in Fig. 7. DBC shows a significant improvement over this standard approach, which we attribute to the significant loss of information in summarizing all cellular features together.

## 4. Discussion

We have shown the potential for DBC as a screening tool of histopathological images for genomic markers. As a proof of principle, we showed the application of our approach for identifying the basal molecular subtype based on imaging features, though our method could easily be trained for other genomic markers, thereby learning new cell types relevant to that marker. The effectiveness of DBC stems from its flexibility in learning what types of cells are informative of the specific genomic marker and how best to jointly leverage the extracted cellular features to define these cell types. Our overall method requires only a few minutes per H&E image of 1000 $\times$1000 pixels to process, which is sufficiently fast for high-throughput image-genomic analysis if WSIs can first be efficiently pruned for representative tiles, which could be helped by the use of parallel GPU computing.

Furthermore, we believe several areas can be improved for our method. As a tumor progresses, the spatial relationship of cells changes, becoming increasingly disordered,[1] and so image features should consist not just of those of individual cells but also those describing their spatial relationship. Graphical features of cells have been shown to increase accuracy of automated non-small cell lung cancer subtype identification,[23] and these features could easily be appended to our BoW image representation to be used in the subsequent classification step. The DBC framework could also easily be extended hierarchically to learn words for local regions of nuclei in a spatial pyramid scheme[9] to capture cellular heterogeneity at various scales. Additionally, as more segmentation and imaging-genomic training data become available, the performance of DBC is expected to also increase. Nevertheless, we believe our approach has the potential to become a highly useful tool to connect imaging features with genomic signals to unravel the phenotypic impact of genomic alterations in cancer.

## 5. Acknowledgments

The authors would like to thank Jack P. Hou for helpful discussions about cancer genomics, Chang Hu for help in collecting nuclei training samples, and Sandhya Sarwate, M.D. for offering her professional pathologist expertise and consultation. We thank the TCGA Research Network for making the data publicly available.

## References

1. D. Hanahan and R. A. Weinberg, *Cell* **144**, 646 (2011).
2. K. Polyak, *The Journal of Clinical Investigation* **121**, 3786 (2011).

3. C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. a. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. a. Akslen, O. Fluge, a. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, a. L. Børresen-Dale, P. O. Brown and D. Botstein, *Nature* **406**, 747 (2000).
4. TCGA Network, *Nature* **490**, 61 (2012).
5. H. Chang, G. V. Fontenay, J. Han, G. Cong, F. L. Baehner, J. W. Gray, P. T. Spellman and B. Parvin, *BMC Bioinformatics* **12** (2011).
6. Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-f. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, N. Johnson, S. Doyle, G. Turashvili, E. Provenzano, S. Aparicio, C. Caldas and F. Markowetz, *Science Translational Medicine* **4**, 157ra143 (2012).
7. C. Wang, T. Pécot, D. L. Zynger, R. Machiraju, C. L. Shapiro and K. Huang, *Journal of the American Medical Informatics Association* **20**, 680 (2013).
8. L. A. D. Cooper, J. Kong, D. A. Gutman, W. D. Dunn, M. Nalisnik and D. J. Brat, *Laboratory Investigation* **95**, 366 (2015).
9. S. Lazebnik, C. Schmid and J. Ponce, Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
10. F. Moosmann, B. Triggs and F. Jurie, *Advances in Neural Information Processing Systems* , 985 (2007).
11. S. Lazebnik and M. Raginsky, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**, 1294 (2009).
12. A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland and D. M. Sabatini, *Genome Biology* **7**, p. R100 (2006).
13. D. L. Rubin, K.-h. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Re and M. Snyder, *Nature Communications* **7** (2016).
14. K. Sirinukunwattana, S. E. A. Raza, Y.-w. Tsang, D. R. J. Snead, I. A. Cree and N. M. Rajpoot, *IEEE Transactions on Medical Imaging* **35**, 1196 (2016).
15. J. Xu, L. Xiang, Q. Liu, S. Member, H. Gilmore, J. Wu, J. Tang and A. Madabhushi, *IEEE Transactions on Medical Imaging* **35**, 119 (2016).
16. F. Xing, Y. Xie and L. Yang, *IEEE Transactions on Medical Imaging* **35**, 550 (2016).
17. A. Janowczyk and A. Madabhushi, *Journal of Pathology Informatics* **7** (2016).
18. A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. C. Shih, J. Tomaszewski, F. A. González and A. Madabhushi, *Scientific Reports* **7**, p. 46450 (apr 2017).
19. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, TensorFlow: A System for Large-Scale Machine Learning, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, (USENIX Association, GA, 2016).
20. M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt and N. E. Thomas, A Method for Normalizing Histology Slides for Quanititative Analysis, in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009.
21. E. D. Gelasca, B. Obara, D. Fedorov, K. Kvilekval and B. S. Manjunath, *BMC Bioinformatics* **10** (2009).
22. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).
23. J. Yao, D. Ganti, X. Luo, G. Xiao, Y. Xie, S. Yan and J. Huang, Computer-Assisted Diagnosis of Lung Cancer Using Quantitative Topology Features, in *6th International Workshop on Machine Learning in Medical Imaging, MLMI'15*, 2015.