Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry

Saurav Basu [*] [†], **Soheil Kolouri** [*] [†], **Gustavo K. Rohde** [†] [‡]

[*]denotes equal contribution,[†]Center for Bioimage Informatics, Department of Biomedical Engineering, and [‡]Department of Electrical and Computer Engineering, Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Edited by Zvi Kam, Weizmann Institute of Science, Rehovot, Israel, and accepted by the Editorial Board January 15, 2014 (received for review October 24, 2013)

Modern microscopic imaging devices are able to extract more information regarding the sub cellular organization of different molecules and proteins than can be obtained by visual inspection. Pre-determined numerical features (descriptors) often used to quantify cells extracted from these images have long been shown useful for discriminating cell populations (e.g. normal vs. diseased). Direct visual or biological interpretation of results obtained, however, is often not a trivial task. We describe an approach for detecting and visualizing phenotypic differences between classes of cells based on the theory of optimal mass transport. The method is completely automated, does not require the use of predefined numerical features, and at the same time allows for easily interpretable visualizations of the most significant differences. Using this method we demonstrate that the distribution pattern of peripheral chromatin in the nuclei of cells extracted from liver and thyroid specimens is associated with malignancy. We also show the method can correctly recover biologically interpretable and statistically significant differences in translocation imaging assays in a completely automated fashion.

Hight content screening | Cellular morphometry | Cytometry

Significance

Much of what is currently known about how cells work has been derived through visual interpretation of microscopy images. Computational methods for image analysis have emerged as quantitative alternatives to visual interpretation. We describe an analysis pipeline for cell image databases that combines statistical pattern recognition with the mathematics of optimal mass transport. The approach is fully automated and does not require the use of ad hoc numerical features. It enables the identification of discriminant phenotypic variations, or biomarkers, between sets of cells (e.g. normal vs. diseased) while at the same time allowing for the visualization of meaningful differences. The approach can be used for fully automated high content screening with a variety of microscopic image modalities.

Introduction

Quantitative analysis of cell images is extensively used in several health sciences applications [1]. Scientists wishing to quantify the effects of certain drugs, genes, and other perturbations (e.g. benign vs. malignant cancer cells) routinely make use of numerical software that are capable of evaluating statistical differences between two populations of cells captured under the microscope [2]. Beyond simple automation facilitating the analysis of thousands of cells, the purpose of such software is to attempt to extract information that the human visual system is unable to cope with. A well-known drawback of existing methods is that the visual interpretation of any differences found is usually 'hidden' from the user. The popular numerical features used to quantify and compare cells such as form factor, Gabor and Haralick texture features, color histograms, etc. [7–9], usually do not have a direct biological interpretation. The situation is even more complicated when multiple features are needed simultaneously to characterize differences between cells, given that the physical interpretation of a combination of features with different units is a nontrivial task. Consequently, statistical tests are limited to determining whether or not two or more cell populations are different. Visual interpretation of any obtained result is usually non-intuitive and difficult.

Here we describe a method, which we call the *transport based morphometry* (TBM), that takes as input a database of pre-segmented cell images and outputs a representation for the same data which can be used for simultaneous visualization and quantitative evaluation in commonplace biological domains. An a priori set of numerical features is not needed as all calculations for comparing cells are done using the entire information present in each cell image. Our approach is based on combining a framework for image analysis based on the theory of optimal mass transport [3, 11] together with a modified version of the linear discriminant analysis technique [10], as well as other modifications and extensions as explained below. Mass, in this case, refers to the intensity associated with each particular pixel, which is often linear w.r.t. number of molecules (e.g. proteins) present in that location. The idea is demonstrated in Fig. 1. The segmented images are first normalized so as to eliminate the effect of translation and rotation, and are spatially morphed to a precomputed 'reference' image. A weighted Euclidean distance computed between two transformations approximates the optimal transport of mass (corresponding to image intensity) between each image, and thus defines a linear embedding for the data. The main phenotype variations in a given dataset can then be given by the standard principal component analysis (PCA) technique. PCA, however, can only be used for visualizing variations as a whole. Statistically significant differences between different classes of cells are computed through a modified version of the penalized Linear Discriminant Analysis (*p*LDA) [10] applied to the linear optimal transport (LOT) coordinates computed by the method described in [3]. Given that the utilized embedding technique is invertible the result of any statistical analysis (e.g. PCA,*p*LDA) can be directly visualized in image space.

With respect to the work in [3, 10] we describe the following extensions. We utilize the LOT framework, in combination with a modified version of the *p*LDA method, to derive a linear discriminant subspace for different cell populations by selecting only the dimensions which contain statistically meaningful distances, as described in the methods (and supplement) section. We also describe a method that allows one to obtain a visual and quantitative interpretation of

Reserved for Publication Footnotes

the phenotype variations in a cell population with respect to a chosen independent variable (e.g. time, drug concentration, etc.). Here we also focus on describing how these advances can play a role in eliciting previously unavailable information regarding cell morphology in several problems relevant to cell biology and pathology. Finally, we combine all these advances into a freely available concise software package that can be used by scientists to perform high content screening tasks [17].

In the results section, we describe the application of TBM to discover information in several high content screening applications. The first application is concerned with discovering the principal differences in nuclear chromatin arrangement between normal, benign, and malignant cells extracted from the liver and thyroid of pediatric patients [4]. Nuclear structure has long been a highly used biomarker in image-based pathology. For many malignancies, however, nuclear structure is not utilized given the absence of knowledge related to discriminating structural information. We utilize TBM to uncover statistically meaningful and at the same time easy to interpret differences in nuclear structure in these malignancies.

The second application is concerned with automated screening for cell phenotype changes in imaging-based assays. Such assays are routinely used for a wide variety of applications including drug discovery, functional genomics, chemical probe discovery, etc. In this paper, we detail the application of TBM to quantifying translocation of the Forkhead fusion protein as a function of Wortmannin dosage in stably transfected human osteosarcoma cells (U2OS) [5]. We show that TBM is able to correctly identify the underlying trend of cytoplasm-to-nucleus translocation in a manner which is both statistically significant and intuitive to understand. We note that in contrast to currently available methods [5], the trend does not have to be "assumed" a priori. Rather it is automatically discovered without any human intervention.

Results

Visualizing variations of chromatin patterns in normal and cancerous cells. Exploratory visual analysis is an important part of coming to a comprehensive understanding of the phenotype variability in a given set of cells derived from a particular experiment. It can be used to obtain an understanding of the main trends regarding shape, structure, and texture variation in a given experiment. We applied TBM to visualize the most significant nuclear structure variations present in the thyroid and liver specimens, as well as in the Forkhead fusion protein in the cytoplasm-to-nucleus translocation imaging assay (see Methods). Using the principal component analysis technique [6], in conjunction with the transport approach described in [3] we are able to conclude that the main modes of variation (in order of decreasing corresponding variance) are: nuclear size, elongation, shifts in chromatin concentration, as well as shifts in chromatin concentration accompanied by nuclear envelope protrusions. The first nineteen components of the TBM-enabled PCA analysis of the liver data are shown in Figure 2. These modes correspond to roughly 90% of the variance in the dataset. In a similar manner (Supplement), the top three modes of variation in the thyroid dataset (preserving 90% of the total variation) were the cell size, cell shape (elongated vertically vs. elongated horizontally) and shifts in chromatin concentration (Fig. 3 in section 1.2 of Supplement) . The top six PCA directions preserving 90% of the variations in the U2OS dataset is also shown in Fig. 4 in section 1.2 of Supplement.

Peripheral migration of nuclear chromatin is predominantly responsible for FHB cancer in liver cells. We utilized the TBM approach to discover the most discriminant, while visually interpretable, differences between normal liver and fetal-type hepatoblastoma (FHB) specimens. While the PCA technique is useful for visualizing overall morphology trends in a given population of cells, by itself, it contains no information regarding which morphology changes are responsible for discriminating two sub populations. To that end, we applied the *p*LDA-based method, described in detail in the methods section. Figure 2, for example, contains no information regarding which modes could be used for differentiating normal vs. cancerous liver cells. Fig. 3 summarizes the visual information uncovered by TBM when investigating FHB cancer in liver cells. We note this result substantially differs from the result obtained by simply applying our earlier work on discriminant subspace selection [10] to the LOT embedding described [3] in that it provides a description of differences which are more statistically significant (see Supplement section 1.5 for more details). The horizontal axis is plotted in units of standard deviation of the chromatin spread along the most statistically significant discriminant direction (in transport space [3]) between benign and FHB cells. In this visualization, each bar in both histograms shown corresponds to the relative number of nuclei that most closely resembled the nuclear structure shown right below. The representative images corresponding to each histogram coordinate are shown below the horizontal axis. The *p*-value of the histogram separation (computed using cross validation) is zero within numerical precision, and therefore the separation of the normal and cancer sub-populations is highly significant. In this case TBM discovers that as the axis of chromatin discrimination slides from left to right, typical nuclear chromatin migrates from peripheral bands in the nucleus to being more and more concentrated at the center of the nucleus. In other words, significant concentration of nuclear chromatin in the center of the nucleus, as opposed to its concentration around the nuclear boundary, can suggest possible FHB condition in the liver cells.

Patterns of circumferential bands in chromatin identifies progression of cancer from normal to FA through FTC in the thyroid. Interesting insights regarding visual differences between the normal, follicular adenoma of the thyroid (FA) and follicular carcinoma of the thyroid (FTC) populations were discovered in Fig. 4 when the thyroid dataset consisting of three sub-populations of normal, FA and FTC cells were input to the TBM pipeline. Fig. 4 demonstrates the difference between the normal, FA and FTC sub-populations, computed using the methods described below. The horizontal axis represents the highest level of visual difference inside the subpopulations, i.e. is directed along the most significant direction of difference between normal, FA and FTC cells. The representative images are generated between every unit of standard deviation from the mean image in the dataset along the discriminant direction. Similar to the previous experiment, the positions of alternate circumferential bands of chromatin concentration is revealed to be a possible biomarker for identifying and distinguishing FA and FTC from the normal case. Whereas one can detect the existence of peripheral and central chromatin concentrations in the benign case, the FTC case seems to have a more uniform chromatin spread across the nucleus and the FA subpopulations are distinguished by a single circumferential concentration band approximately halfway between the periphery and the center. In order to facilitate a clearer inspection of the pairwise differences between the three classes (normal, FA and FTC), Fig. 6 in the Supplement section 1.3 shows the pairwise histogram projections on the most discriminant direction found in Fig. 4.

Gradual translocation of the Forkhead protein (FKHR-EGFP) in the nucleus of transfected human osteosarcoma cells (U2OS) with variation of dosage of Wortmannin is visually verified by TBM. The third dataset containing Wortmannin injected assays of U2OS cells to affect translocation of the Forkhead protein in the nucleus serve as a verification tool for the statistical and representational veracity of TBM. As shown in Fig. 5(A), the Forkhead protein (FKHR-EGFP) gradually translocates from the cytoplasm towards the nucleus of U2OS cells (left to right) with increasing dosage of the drug Wortmannin [12]. Note that although there are four realizations of twelve assays in the dataset [12] with the first assay being the negative control (no Wortmannin) and the twelfth and last assay being the positive control (maximum Wortmanni-n of 250 nM added), we have only shown six equispaced assays in Fig. 5(A) from the first realization [12].

TBM can be used to automatically recover the pattern of translocation of FKHR-EGFP in the U2OS cells through a representational axis signifying the most significant pattern of variation of FKHR-EGFP in the cytoplasm. Fig. 5(B) bundles all the realizations of the first six assays into a positive control and the last six assays into a negative control, and the discrimination visualization step of the TBM is applied to the two sub-populations to verify the FKHR-EGFP translocation. It can be seen from Fig. 5(B) that the projections of the positive control (0.00nM-7.81nM) in cyan is clearly separated from the projections of the negative control (15.63nM-250nM) in red along the most visually discriminant direction estimated by *p*LDA. Moreover, the horizontal axis (again plotted in units of standard deviation of the FKHR-EGFP variation) has been tagged with synthesized images that statistically express the average FKHR-EGFP translocation along the dosage increase. It can be observed from the synthesized images in Fig. 5(B) that the FKHR-EGFP translocation has been accurately captured.

In addition, the projections of all U2OS cells in twelve individual assays along the discriminating direction in Fig. 5(B) are shown in Fig. 5(C). All cells in a particular assay have been given a unique color. It can be observed from Fig. 5(C) that the projection histograms of individual assays shifts from the negative control towards the positive control with increase in Wortmannin dosage, with a rather abrupt change occurring between 7.8 and 15.6 nM. This serves to verify that the negative to positive discrimination direction in Fig. 5(B) estimated by TBM in fact gives reasonable progression of the translocation.

Maximally correlated images with respect to Wortmannin dosage shows quantifiable translocation of the FKHR-EGFP from the nucleus towards the periphery of the average cell boundary. TBM also provides considerable insight into the response of the U2OS cells to Wortmannin dosage variation with respect to FKHR-EGFP translocation. Using the TBM pipeline we computed the direction in LOT space that is most correlated with dosage values (see methods). Fig. 6 demonstrates the dosage response curve. The horizontal axis represents the logarithm of the Wortmannin dosage values (log(0.977nM)-log(250nM)), note that 0.00nM is not included in this experiment. The vertical axis represents the normalized projected value of the data on the maximally correlated direction, described above. These normalized projected values serve as a measure of Wortmannin activity where zero corresponds to negative control and one corresponds to positive control. The images corresponding to $0\% - 100\%$ activity of the Wortmannin have been shown along the vertical axis. The presented curve matches with the one reported in the product specification of FKHR Redistribution Assay provided by Thermo Fisher Scientific Inc in 2008.

Summary and Discussion

We have described TBM (see [17] for download) as a method for decoding morphology differences in cell populations. The method builds on previous work related to image processing using optimal transport [3], as well as *p*LDA [10], by adding the capability to construct a quantitative, discriminant, linear subspace of cell phenotypes that can be directly visualized. We note the proposed LOT-based discriminant subspace described by our TBM approach is substantially different than our previous work in [10] in that it provides a more reliable description of statistically significant differences between two cell populations (see Supplement section 1.5 for more details). In addition, we also describe how to use the framework for visualizing the morphological variations most correlated with a given independent variable (e.g. drug dosage). This visual analysis of structural differences can lead to improved understanding of inter-relationships between cellular structure and functions. We believe TBM is the first systematic approach of a totally automated visual exploratory tool in cell image analysis that offers both the benefits of observation as well as involved statistical tests on cell image databases containing one or more phenotypes.

We applied TBM to discover statistically significant and visually interpretable differences of nuclear chromatin configuration in normal vs. cancerous cells. The analysis shows that, on average, the most discriminative information in these was how much chromatin is present in the center vs. periphery of these. Malignant cancerous cells were shown, on average, to have more chromatin concentrated and packed at the center of the nuclear envelope, a finding consistent with the biology of cancer cells [15]. We believe TBM could be used in numerous pathology and cytology applications to recover visually interpretable differences between normal, benign, and malignant cells.

We would like to note here that there are few methods for direct application of statistical analysis tools, such as PCA and LDA, to the image pixel intensities directly. For example, a direct and naive application of the PCA to the pixel intensities following [16] can lead to the discovery of biologically meaningless artifacts as evidenced in section 1.4 in the Supplement. Pixel intensity variations inside real images are highly non-linear and any linear interpolation of pixel intensities in the image space leads to presence of aliasing artifacts inconsistent with real images.

In addition, we have also used TBM to blindly recover known information regarding nuclear to cytoplasm protein translocation in a screening assay. In the U2OS dataset, TBM not only confirms the translocation of FKHR-EGFP visually, establishing the veracity of the functionality of TBM, but it also outputs a visual variant of the dosage response curve of the drug Wortmanin that clarifies the representative effect of the dosage increase on the average U2OS cell.

In conclusion, we anticipate important use of TBM in cell phenotype analysis in part due to the visual exploration component that provides intuitive insight into the structural modes of cell construction. In addition, TBM is fully automated and relieves the end user of manual and tedious definition as well as selection of arbitrary image features, potentially leading to more accuracy and promises of generative modeling of cells. It can be applied to analyze segmented cell images where the cell content (intensity or texture) can be viewed as a distribution of 'free mass'. In this work we applied TBM to two dimensional fluorescence and transmitted light microscopy images, but other microscopy imaging modalities (e.g. Coherent Anti-Stokes Raman Scattering) could also benefit. The framework is also amenable to three dimensional images (albeit at an increase in computation cost). Finally, the TBM framework was presented here in the context of scalar images. It can also be used to analyze the relationship between multiple protein distributions, obtained through multiple fluorescence labels or spectral imaging modalities, in cell populations. This topic will be the subject of future work.

Materials and Methods

Datasets. In order to demonstrate the ability of TBM to discover visual information hitherto impossible with the standard feature based approaches, we have identified three cell image datasets that include cell texture variation either due to functional difference (cancer versus non-cancer) or drug infusion (drugs inhibiting protein translocation in cells). The first two datasets have been obtained from the archives of the University of Pittsburgh Medical Center. The first dataset contains microscopy images of liver tissue samples obtained from ten different subjects including five cancer patients suffering from fetal-type hepatoblastoma (FHB), with the remaining images from the liver of five healthy individuals. The second dataset contains microscopy images of resection specimens of thyroid from twenty different subjects. The first ten subjects provide images of normal thyroid tissue, patients diagnosed with follicular adenoma of the thyroid (FA) provide the next five cases and the last five cases belong to patients diagnosed with follicular carcinoma of the thyroid (FTC). The acquisition of these datasets is described in detail in [11]. Briefly, images were stained with the Feulgen technique to tag DNA content and scanned at 0.074μm per pixel resolution.

The third image dataset demonstrates cytoplasm to nucleus translocation of the Forkhead (FKHR-EGFP) fusion protein in stably transfected human osteosarcoma cells, U2OS [12]. In this assay, the images are obtained from 48-well plates of cells incubated with 12 different dosages of Wortmannin. The images have a resolution of 0.6μm per pixel.

Image Segmentation. Prior to being analyzed with our TBM approach described below, each morphological exemplar (DNA pertaining to one nucleus, or protein distribution from one cell) was first segmented using standard approaches. The nuclear datasets were segmented as described in [4]. The liver dataset were segmented to have 500 nuclei with an average of 50 nuclei per patient. The Thyroid dataset consisted of 2053 cell images with an average of 102 images per patient.

The U2OS cell dataset was segmented using Cellprofiler [5] with the exact pipeline described in [5]. The DNA channel is used to locate the nuclei and consequently mark the seed points of a growth step where every seed point grows into a closed curve that encircles its respective cell boundary. On an average 730 images of individual cells were extracted from images of the wells incubated with the same dosage of Wortmannin, leading to 8756 images of cells with 12 different classes (12 different dosages of Wortmannin). A detailed discussion of the segmentation procedure is presented in [5].

Transport-based cell morphometry. The aim of our TBM approach is to take as input segmented morphological exemplars and output 'coordinates' for each exemplar that can be used for both visualizing the the main modes of variation of a dataset as well as the main ways in which two or more groups of morphological exemplars differ from one another. To that end, the images are normalized following a similar approach as described in [14], the linear optimal transportation embedding is calculated from the normalized images [3], and a modified version of the penalized Linear Discriminant Analysis (*p*LDA) method [10] is utilized to capture the most statistically significant discriminant direction. Finally, we used the well-known Kolmogorov-Smirnoff test [13] for assessing significance when comparing distributions over a chosen linear subspace.

Preprocessing

Each segmented structure is first normalized so that the sum of its intensities equals one, so as to remove differences in staining procedures during imaging. This normalization limits us to investigating relative changes in overall mass distribution while all information regarding absolute amounts is lost. In addition, due to computational complexity considerations, each segmented structure is also approximated with 'point masses' using the algorithm described in [3]. Briefly, a weighted K-means clustering algorithm is utilized, in conjunction with the available image intensities, so as to best approximate the input image. The number of point masses is chosen so as to keep the computation time within a reasonable range (e.g. 30s per image pair). Details of the algorithm used are available in [3]. When it comes to visualizing images from particle approximations (processed as explained below), bilinear interpolation is used to distribute the masses onto the image grid, and a Gaussian function of small variance is used to render a more realistic visualization of the discrete particles. In the end, each morphological exemplar is represented by $\sum_{j=1}^{N} p_j \delta_{\vec{y_j}}$ where N is the number of masses being used, $\vec{y_j}$ are the 2D Cartesian coordinates of the j^{th} particle, while p_j is its mass here $\delta_{\vec{y_j}}$ is an unit impulse function placed at the location $\vec{y_j}$). In this study, we utilized $N = 600$ masses per structure. Following this procedure each point mass approximation is translated so that its center of mass is at the center of the field of view, and its main axis of orientation, whose computation includes its mass distribution, is aligned with the vertical axis. For cells whose mass distribution is perfectly circular and uniform, such an axis is impossible to define, as they are identical under all possible rotations.

Optimal Transport and Linear Embedding

The main idea in TBM is to quantify similarities between morphological structures in each dataset (all three datasets are processed independently) by measuring the amount of effort (quantified as mass times distance that it must be transported) that would have to be spent to re-arrange the particle approximation of one structure onto another [11]. Here we use the linearized version of this metric constructed based on a tangent space approximation of the underlying Riemannian manifold [3]. The idea is to first compute a reference structure and then compute the optimal transport between each image in the database and the reference structure. As in [3] we compute an average structure by running the particle approximation algorithm on the Euclidean average of the input digital images (computed after normalization for rotation, translation, and intensity, as described above). Let $\sigma = \sum_{k=1}^{N_\sigma} q_k \delta_{\vec{z_k}}$ be the representation of the reference structure for the given dataset Let $\mu = \sum_{i=1}^{N_\mu} m_i \delta_{\vec{x_i}}$ be a sample structure from the dataset. The optimal transport between μ and σ is computed by

$$d^2{}_{OT}(\sigma,\mu) = \min_{f\in\Pi(\sigma,\mu)} \sum_{k=1}^{N_\sigma}\sum_{i=1}^{N_\mu} |\vec{z_k}-\vec{x_i}|^2 f_{ki} \qquad \mathbf{[1]}$$

subject to $f_{ki} \geq 0, \sum_{i=1}^{N_\mu} f_{ki} = q_k, \sum_{k=1}^{N_\sigma} f_{ki} = m_i$. Here $\Pi(\sigma,\mu)$ is the family of all transport plans from σ to μ, and f denotes the optimal 'transport plan'. Thus the transport plan is simply an assignment function that states what proportion of the mass (intensity) of the particle at $\vec{z_k}$ has moved to the particle at $\vec{x_i}$ [3]. Since these can take fractional values between 0 and 1, splitting and joining of particles is possible. These are often rare as in most cases whole (either 0 or 1) assignments are made. Similarly, let g be the optimal transport plan between structure σ and $\nu = \sum_{j=1}^{N_\nu} p_j \delta_{\vec{y_j}}$. A linear embedding for structures μ and ν can then be computed by

$$\bar{x}_k = \frac{1}{\sqrt{q_k}} \sum_{i=1}^{N_\mu} f_{ki}\vec{x_i} \text{ and } \bar{y}_k = \frac{1}{\sqrt{q_k}} \sum_{j=1}^{N_\nu} g_{kj}\vec{y_j}, \quad k = 1, \cdots, N_\sigma. \qquad \mathbf{[2]}$$

while the linear optimal transport between μ and ν is given by:

$$d_{LOT,\sigma}(\mu,\nu) = \sum_{k=1}^{N_\sigma} |\bar{x}_k - \bar{y}_k|^2. \qquad \mathbf{[3]}$$

Thus, after pre-processing, the linear embedding for each morphological structure in a database of images is computed from equation **[2]**. The approximate transport distance between any pair of images in the database is computed from equation **[3]**. It can be noted the resultant dimensionality of the linear embedding is $600 \times 2 = 1200$.

Visualizing principal phenotypic variations

Given the LOT embedding computed from equation **[2]**, we utilize the standard principal component analysis (PCA) [6] technique for data visualization. The covariance matrix for the LOT embedded image set is $S_T = \frac{1}{M}\sum_m (\mathbf{x}_m - \bar{\mathbf{x}})(\mathbf{x}_m - \bar{\mathbf{x}})^T$ with $\bar{\mathbf{x}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}_m$, where $\mathbf{x}_m$ is the m^{th} vectorized LOT embedded image. The principal components are given by the eigenvectors of S_T, and can be used to explain, and in this case, visualize the main modes of phenotypic variation in a dataset. As customary we have retained the top k PCA directions that preserve 90% of the variation of the original LOT embeddings. We have plotted the texture variation within cell populations by simply reconstructing intermediate LOT embeddings along the PCA axes.

Detecting and visualizing phenotype differences.

The LOT embeddings can also be used for visualizing the discriminating modes of cell texture variation between cell subpopulations. To that end we apply the pLDA technique described in [10] to compute the most discriminant components explaining the differences between two subsets (classes) of a given dataset. The penalized LDA direction that denotes the direction along which the projections of the C classes are maximally separated (in the LDA sense) is given by the solution to the optimization problem

$$\mathbf{w}_{pLDA} = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \frac{\mathbf{w}^T S_T \mathbf{w}}{\mathbf{w}^T(S_W + \alpha \mathbf{I})\mathbf{w}} \qquad \textbf{[4]}$$

where $S_W = \sum_c \sum_{n\in c} (\mathbf{x}_n - \bar{\mathbf{x}}_c)(\mathbf{x}_n - \bar{\mathbf{x}}_c)^T$ represents the 'within class scatter matrix'. The penalty weight α signifies a trade-off between the traditional LDA direction and the topmost PCA directions of the same dataset (see [10] for a motivation). The precise methodology for the determination of α for this work is based on fitting an exponential decay model to a metric that measures how far two consequent subspaces are from one another (see Supplement section 1.5 for details). As in the case of PCA, several mutually orthogonal pLDA directions can be found from an augmented version of eqn. (6) which give consecutive directions which show maximal residual discrimination between the populations. Consistent with existing literature, we have retained the topmost pLDA direction that shows significant statistical difference ($p \leq 0.05$ in a Kolmogorov-Smirnov test between distributions) between the projected histograms of the cell subpopulations along the direction. An important distinction here is that in our method we choose to report only the statistically meaningful directions computed using cross validation (using a portion of the data held out from the training process). We make note that, as a whole, the method described in this subsection yields a linear discriminant subspace that significantly differs from the procedure described in our earlier work [10]. Comparisons between the method described here and our earlier method [10] are available in Supplement section 1.5 and show the method can provide more reliable information regarding difference between cell populations.

Computing morphology variations most correlated with an independent variable

Define vector $\mathbf{v} = [v_1, \ldots, v_M]^T$ such that v_i is a scalar attribute of the i'th image (i.e. dosage of Wortmannin). We are able to search in the LOT space to find a direction, which is most correlated with $\mathbf{v}$. Hence, we are able to visualize the statistical effect of that specific attribute in the dataset. The most correlated direction, $\mathbf{w}_{corr}$, is found as follows,

$$\mathbf{w}_{corr} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\mathbf{w}^T X \mathbf{v}}{\sqrt{\mathbf{w}^T \mathbf{w}}} = \frac{X\mathbf{v}}{\sqrt{\mathbf{v}^T X^T X \mathbf{v}}} \qquad \textbf{[5]}$$

where $X = [\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_M - \bar{\mathbf{x}}]$ is a matrix, which contains the vectorized and mean subtracted, LOT images as its columns. The given direction can then be visualized by plotting $\mathbf{w} = \bar{\mathbf{x}} + \lambda \mathbf{w}_{corr}$, with λ a chosen length along the projection (in units of standard deviation of the projected data along $\mathbf{w}_{corr}$).

A Note on Numerical Implementation

All computations of TBM were performed in a highly parallel distributed computing cluster in the ECE department at Carnegie Mellon, and average computation time for generation of results in each dataset extended to a couple of hours. Computer code in the Matlab language is available [17].

ACKNOWLEDGMENTS. The authors would like to acknowledge informative discussions with Drs. Wei Wang, John A. Ozolek, and Dejan Slepcev. The authors acknowledge support from NIH grant GM090033.

1. Editorial. The quest for quantitative microscopy. *Nature Methods*, 9(7):627–627, 2012.
2. L. Yang, W. Chen, P. Meer, G. Salaru, L. A. Goodell, V. Berstis, and D. J. Foran (2009) Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. *Information Technology in Biomedicine, IEEE Transactions on*, 13(4):636-644.
3. W. Wang, D. Slepcev, S. Basu, J. A. Ozolek, and G. K. Rohde (2013) A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Computer Vision*, 101(2):254-269.
4. W. Wang, J. Ozolek, and G. K. Rohde (2010) Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry Part A*, 77(5):485-494.
5. 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 (2006) Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.
6. T. W. Anderson (1963) Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34(1):122-148, 1963.
7. P. H. Bartels, T. Gahm, and D. Thompson (1997) Automated microscopy in diagnostic histopathology: From image processing to automated reasoning. *International journal of imaging systems and technology*, 8(2):214-223.
8. C. Demir and B. Yener (2005) Automated cancer diagnosis based on histopathological images: a systematic survey. *Rensselaer Polytechnic Institute, Tech. Rep. TR-05-09*, Troy, NY.
9. K. Rodenacker, E. Bengtsson (2002) A feature set for cytometry on digitized microscopic images. *Analytical Cellular Pathology*, 25(1):1–36.
10. W. Wang, Y. Mo, J. A. Ozolek, and G. K. Rohde (2011) Penalized fisher discriminant analysis and its application to image-based morphometry. *Pattern Recognition Letters*, 32(15):2128–2135.
11. W. Wang, J. A. Ozolek, D. Slepcev, A. B. Lee, C. Chen, and G. K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE Trans. Med. Imag.*, 30:621–631, 2011.
12. V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter (2012) Annotated high-throughput microscopy image sets for validation *Nature methods*, 9(7):637-637.
13. R. Boddy, G. Smith (2009) Non-Parametric Statistics, Statistical Methods in Practice: for Scientists and Technologists *John Wiley & Sons*, 129-138.
14. G. K. Rohde, A. J. S. Ribeiro, K. N. Dahl and R. F. Murphy (2008) Deformation-based nuclear morphometry: Capturing nuclear shape variation in HeLa cells. *Cytometry A*, 73(4): 341-350.
15. D. Zink, A.H. Fischer, J.A. Nickerson (2004) Nuclear structure in cancer cells. *Nature reviews cancer*, 4(9):677-687.
16. M. A. Turk and A. P. Pentland (1991) *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp 586-591.
17. <http://www.andrew.cmu.edu/user/gustavor/software.html> To be made available upon acceptance.

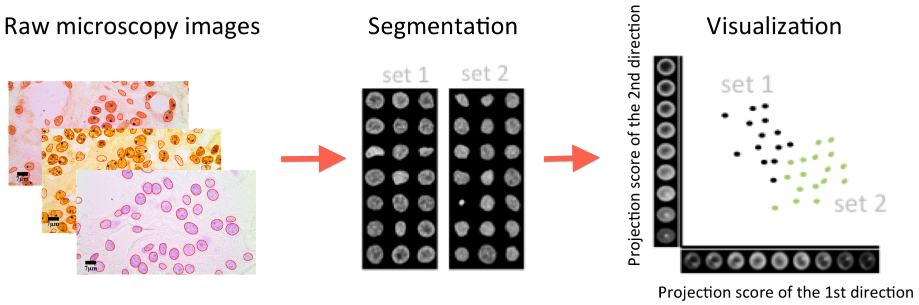

Raw microscopy images
Segmentation
set 1
set 2
Visualization
Projection score of the 2nd direction
set 1
set 2
Projection score of the 1st direction

Fig. 1. Schematic of the TBM pipeline. Microscopy images (on the left) are first segmented to obtain individual cells, the scale bars on the images denote $7\mu m$. Individual images are then normalized to eliminate translation and rotation. Finally, the embedding of the corresponding image is computed in a linear space that is both discriminative and visualizeable.

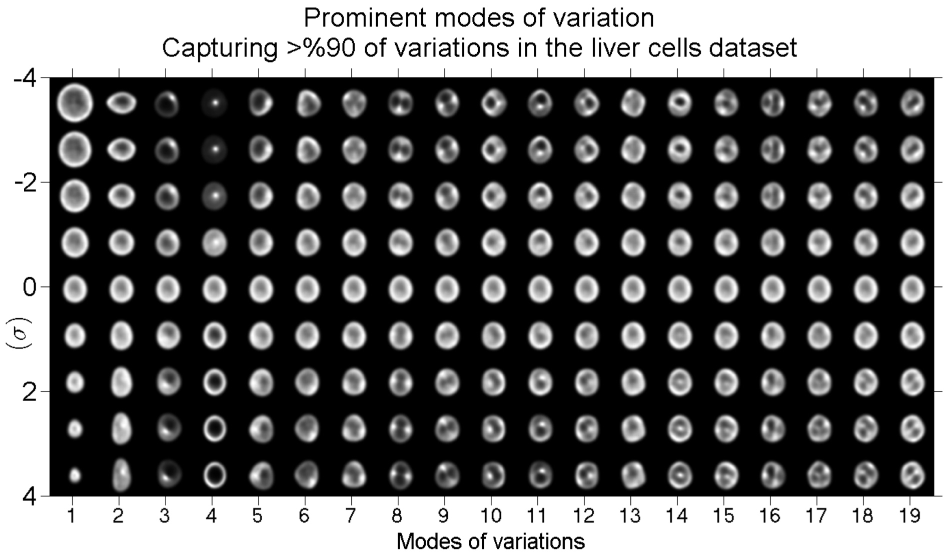

Prominent modes of variation
Capturing >%90 of variations in the liver cells dataset
-4
-2
0
2
4
(σ)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Modes of variations

Fig. 2. The top nineteen modes of variations obtained from applying PCA to the linear transport embedding of the liver dataset. The vertical axis represents units of standard deviation of chromatin variation in a particular mode of variation. The first mode of variation demonstrates that the overall size of the nucleus is the dominant variation in the data set. The second mode suggests that the second most abundant variation in the data set is elongation. The remaining variations correspond to the protrusion of cells and shifts in the center of mass of chromatin distribution. The demonstrated variations explain roughly 90% of the variance in the dataset.

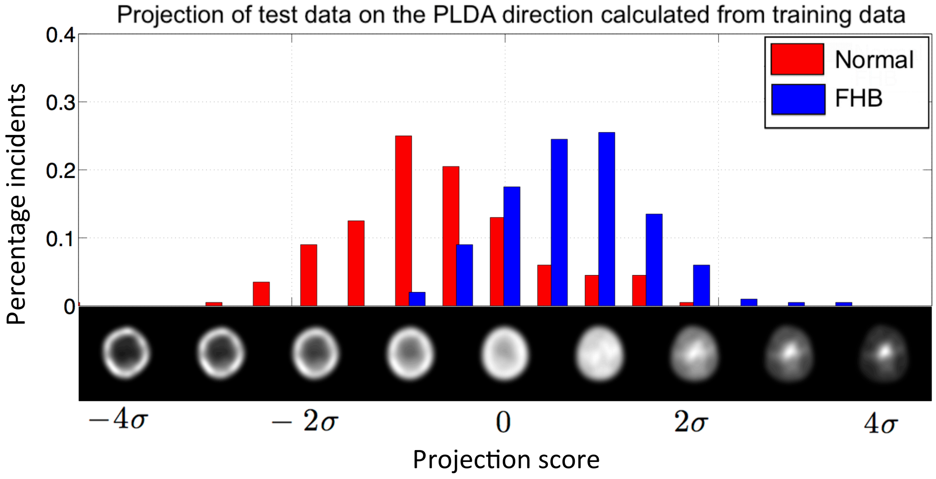


Fig. 3. The histograms of the projections of the coordinates of images of FHB and normal liver cells on the most discriminant direction. The representative images correspond to histogram coordinate which is in units of standard deviation of the projection. The p-value of the histogram separation (in a two sample Kolmogorov-Smirnov test [13] for statistical difference in distributions) is zero within numerical precision. It can be seen that the peripheral migration of nuclear chromatin is predominately responsible for FHB cancer in liver cells.

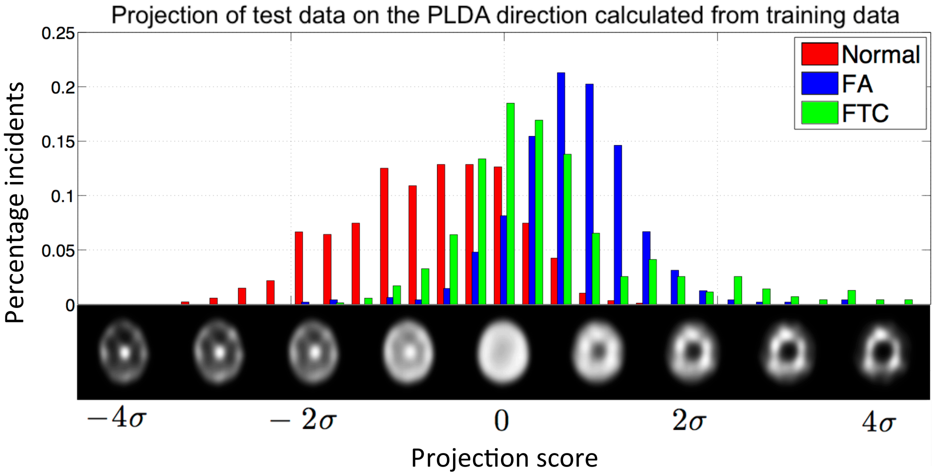


Fig. 4. The histograms of the projections of the coordinates of images of the normal, follicular adenoma of the thyroid (FA) and follicular carcinoma of the thyroid (FTC) sub-populations in the thyroid dataset on the most discriminant direction. The representative images correspond to histogram coordinate which is in units of standard deviation of the projection. The pairwise p-value of the histogram separation (in a two sample Kolmogorov-Smirnov test [13] for statistical difference in distributions) is zero within numerical precision. It can be seen that the pattern of the radial bands of chromatin concentration in the nuclei constitute a discriminating factor.

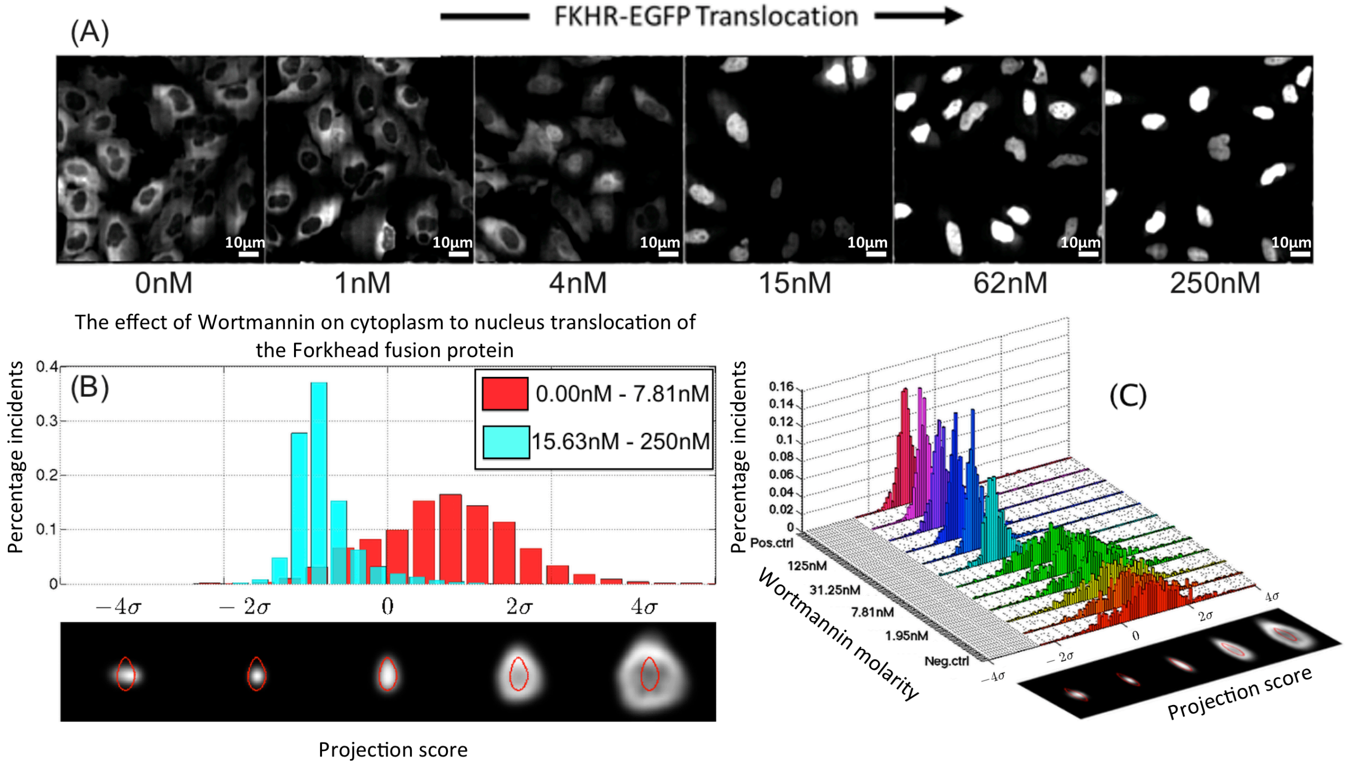


Fig. 5. (A) 6 equispaced assays from the first realization with injected Wortmannin dosage equal to from left to right 0nM, 0.977nM, 3.91nM, 15.63nM, 62.5nM, and 250nM, the scale bars on the images correspond to $10\mu m$.

(B) The projection of data on the most discriminant direction which separates the groups of assays with 0.00nM-7.81nM and 15.62nM-250nM dosage of injected Wortmannin which is shown in red and cyan, respectively. The presented images correspond to histogram coordinate which is in units of standard deviation of the projection, and the red contour represents the mean size of the nuclei throughout the data set. (C) the projection of the data on the most discriminant direction which separates the assays with different amount of injected Wortmannin with corresponding images. Each histogram represents the projection of an assay with specific dosage of Wortmannin.

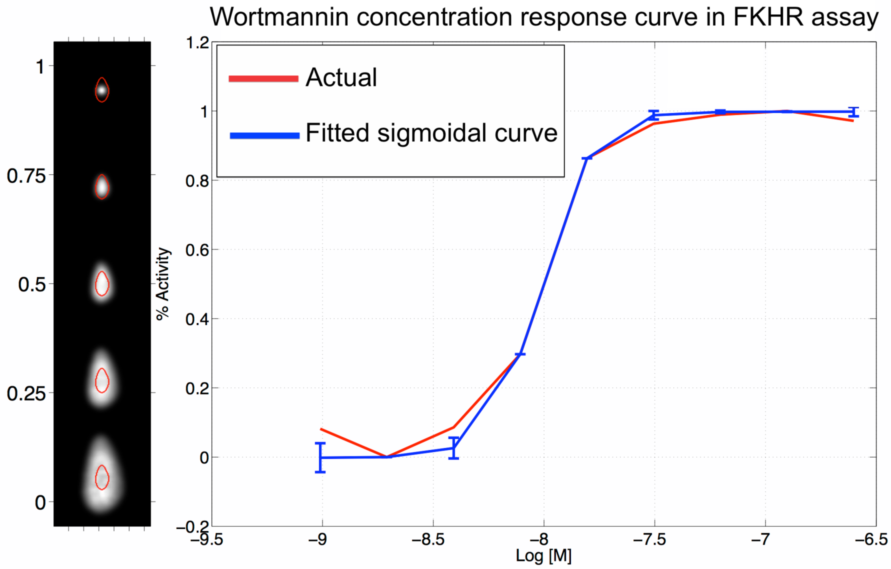


Fig. 6. The Wortmannin concentration response curve in FKHR assay. The horizontal axis shows the logarithmic dosage of injected Wortmannin (log(0.977nM)-log(250nM)). The maximally correlated direction with the dosage values is found and the vertical axes shows the normalized projection of the data on this direction. The presented images correspond to the projected values and alternatively to the activity of the Wortmannin drug. The red curve is the Wortmannin concentration response curve calculated from the data, while the blue curve shows a sigmoidal function fitted to the actual curve. Note that the presented images are contrast stretched for visualization purposes.

Supporting Information
Basu et al. 10.1073/pnas.1319779111

SI Text

Datasets. Sample images showing cells from each of the phenotype groups for liver and thyroid cells are depicted in Figures (7) and (8), respectively. It can be seen that the phenotypic differences between groups are highly complex and cannot be easily determined by visual inspection.

Complete set of PCA images for the datasets. As noted before, the last step of TBM is to apply linear statistical analysis techniques such as the (1)PCA to compute the first k directions of texture variation in the embedded image set that preserve at least 90% of the original variance, and (2) pLDA to compute the most statistically significant discriminant directions (p value ≤ 0.05 w.r.t. Kolmogorov-Smirnov test). In the main body of the paper, we omitted showing all the relevant PCA directions for ease of visualization and understanding. However, for a complete understanding of the TBM pipeline and to establish the uniformity of the TBM pipeline as such, we provide the figures that visualize all the relevant PCA directions for all the three datasets we have used in this paper. For the liver dataset, Fig. 9 shows all the relevant PCA directions whereas Fig. 10 shows all relevant PCA directions for the U2OS dataset.

Complete set of LDA images for the datasets. As noted earlier in the Results section, TBM can also be used to generate representative images that span discriminative directions that are a combination of one or more top pLDA directions to provide a deeper insight into the total discrimination and their relative amounts present in the dataset. To further clarify out point, after obtaining the linearly embedded images, we applied pLDA to the embedded image database to obtain the two most discriminant directions between the three sub-populations. This simply means that the directions are ordered by their relative discrimination power, and as we move along any particular direction, we can expect to see chromatin pattern changes that represent the three sub-populations progressively.

Fig.11 (A) demonstrates the visual difference between the normal, FA and FTC sub-populations. The horizontal axis represents the highest level and the vertical axis represents the second highest level of visual difference respectively. The representative images are generated between every unit of standard deviation from the mean image in the dataset along both directions. Any cell image in Fig. 11 (A) represents a linear combination of the two discriminating modes and signifies how the cell phenotype texture changes as we move between the sub-populations. The highly negative axes values in Fig. 11 (A) represent the normal case, and the images gradually move towards the positive axes through the FA class and ultimately attains highly positive values in both the axes for the FTC class. Fig. 11 (A) shows that as one moves from the normal to FTC through the FA class, the thyroid nuclei predominantly differed in their concentration of chromatin in peripheral bands and their shape. Greater departure from the normal to the FTC through FA gradually shifted the chromatin in radial bands farther and farther away from the nuclear centre, whereas the nuclear shape became thinner.

Fig. 11 (B) shows the corresponding representation of the image data in Fig. 11 (A) as projections along the visually discriminant directions. It can be seen from Fig. 11 (B) that the FA and FTC cases overlap quite a lot and it is difficult to differentiate only through visual examination, as is confirmed by Dr. John A. Ozolek, assistant professor of pathology at University of Pittsburgh.

Additionally, the pairwise discrimination between the projected normal, FA and FTC histograms on the top discriminant direction in the thyroid dataset is shown in Fig. 12.

An alternative pixel based analysis technique to TBM. A comparison of TBM with a straightforward application of linear statistical analysis tools to image pixel intensities is presented in this section. For all images in the dataset having the same dimensions, centered and aligned, the pixel intensities can be included in a giant one-dimensional vector with each index of the vector representing a distinct pixel location. These giant vectors can be treated as feature vectors representing the image itself. Although application of the PCA and LDA to these feature vectors provides a visual exploration of the variation within the dataset, yet, intrinsic nonlinearity of the pixel distributions lead to discovery of information, which strays far from any real biological phenomenon as evidenced in Fig. 13 and Fig. 14.

Discriminant directions. In this Section, we explain our method for selecting the discriminant subspace and highlight the differences between the proposed method and the one introduced in Wang et al. , IJCV 2013 [3]. We show that these methods lead to significantly different lower dimensional discriminant subspaces. The proposed method is different from that of Wang et al. in two major ways. First, we use a different approach for finding parameter α compared to the method used in Wang et al.. Second, in our approach we select those discriminant directions that provide statistically significant discrimination in the dataset, while Wang et al. use the top pLDA directions. Below we explain the mentioned differences, and the motivation behind them in detail.

1. **Identifying parameter α:**
   Both methods utilize the formulation of the pLDA discriminant subspace,

$$\mathbf{w}_{pLDA}(\alpha) = \underset{\|\mathbf{w}\|=1}{\text{argmax}} \frac{\mathbf{w}^T S_T \mathbf{w}}{\mathbf{w}^T (S_W + \alpha \mathbf{I})\mathbf{w}}, \qquad [6]$$

   where the regularization parameter, α, in the pLDA objective function provides a trade-off between PCA and LDA. For $\alpha = 0$ the objective function of the penalized LDA is clearly equal to the objective function of LDA. However, as $\alpha \to \infty$ the objective function of penalized LDA becomes equal to that of PCA. Therefore one expects to see a gradual change in the penalized LDA space with increasing α from zero to infinity. This fact is the motivation for choosing α in both methods.
   Wang et al. increase the value of α and measure the stability of the first pLDA direction using the norm of the difference of the two consequent discriminant directions. They pick $\alpha = \alpha^k$ if,

$$\|\mathbf{w}_{pLDA}(\alpha^k) - \mathbf{w}_{pLDA}(\alpha^{k-1})\|_2 \leq \epsilon, \qquad [7]$$

   where ϵ is a predetermined small value that is set to $\epsilon = 0.005$, and $\alpha^0 < \alpha^1 < \cdots$ are the sequence of α's.
   In our approach, however, we increase the value of α and measure the stability of the pLDA subspace (more than one direction) for two consequent α values using the projection metric. The projection metric between two orthonormal matrices, $\mathbf{X}$ and $\mathbf{Y}$, of size $D \times M$ is defined as,

$$d_P(\mathbf{X}, \mathbf{Y}) = (\sum_{k=1}^{M} sin^2(\theta_k))^{\frac{1}{2}}, \qquad [8]$$

Reserved for Publication Footnotes

where $0 \leq \theta_1 \leq \ldots \leq \theta_M \leq \frac{\pi}{2}$ are the principle angles between to subspaces $span(\mathbf{X})$ and $span(\mathbf{Y})$, and are defined as

$$\begin{aligned} cos(\theta_k) = \quad & \max_{\mathbf{u}_k \in span(\mathbf{X})} \max_{\mathbf{v}_k \in span(\mathbf{Y})} \ \mathbf{u}_k^T \mathbf{v}_k \\ & s.t. \ \|\mathbf{u}_k\|_2 = \|\mathbf{v}_k\|_2 = 1, \\ & \mathbf{u}_k^T \mathbf{u}_i = \mathbf{v}_k^T \mathbf{v}_i = 0, \{i = 1, \ldots, k-1\} \end{aligned} \quad [9]$$

Given $\mathbf{W}(\alpha^k) = [\mathbf{w}^1_{pLDA}(\alpha^k), \mathbf{w}^2_{pLDA}(\alpha^k), ..., \mathbf{w}^N_{pLDA}(\alpha^k)]$, where $\mathbf{w}^i_{pLDA}(\alpha^k)$ is the i'th discriminant direction calculated with parameter α^k, we calculate the projection metric between $\mathbf{W}(\alpha^k)$ and $\mathbf{W}(\alpha^{k-1})$. Figure 15 shows the projection metric of two consequent subspaces as a function of α. At last, we fit an exponential function to the calculated curve and set $\alpha = \alpha^k$ that corresponds to the half life of the fitted function.

Finally, Figure 16 shows the projection of the data (Liver cells) onto the first *p*LDA direction calculated from the α parameter identified with the approach in Wang et al. [10] and our approach. It can be clearly seen that our proposed method leads to a more visible separation between the distributions.

2. **Statistical significance ranking:**

Another important feature of the method proposed in this paper is to rank the *p*LDA directions based on their statistical significance in the sense of Kolmogorov-Smirnov. While the generalized eigen-values in Equation 6 designate an ordering on the discriminant directions, such that the direction corresponding to the i'th largest eigenvalue is the i'th most discriminant direction in the sense of Fisher score, LDA and consequently *p*LDA are best suited for discrimination between Gaussian distributions. Hence, if the data distribution does not follow a Gaussian-like distribution, the ordering provided by the eigenvalues is not reliable. This is the motivation behind reranking the discriminant directions by their statistical significance using a nonparametric statistical test like Kolmogorov-Smirnov test. We employ a nonparametric test to avoid any assumptions on the distribution of the data.

In order to show the importance of the statistical significance ranking, we projected the data (Liver cells) onto the subspace calculated from the method proposed by Wang et al. and our proposed method. Figure 17 shows the projection of the data on the 2D subspaces. It can be clearly seen that the data is more separable in the discriminant subspace calculated by our proposed method.

To show that the distribution of the data on the subspace provided by our method provides more discrimination compared to that of Wang et al., we ran a multivariate two-sample test on the projected data shown in Figure 17. We employed the multivariate nonparametric test proposed by Friedman and Steppel (1974) [S1, S2]. The method counts the number of points among the k nearest neighbors (kNN) of each point (say in class 1) that belong to the same class (class 1). From these counts, separate frequency distributions can be compiled for each class. Finally a nonparametric univariate two-sample test is applied to the frequency distributions of the two classes to calculate the statistical significance of discrimination provided by the subspace. We applied the Kolmogorov-Smirnov test to the frequency distributions (Normal and FHB cells) of the projected points into the subspaces show in Figure 17. Figure 18 shows the frequency distributions and their corresponding p-values for the discriminant subspace obtained from the two methods. P-values for the statistical tests are shown in the caption of the figures. It can be seen that our proposed method leads to significantly lower p-value compared to the method in Wang et al [10].

It is worthwhile to mention that the same procedure is applied to other datasets, namely the thyroid dataset and the cytoplasm to nuclei translocation datasets. In all cases, our proposed discriminant subspace provided significantly lower p-values (data not shown for brevity). In the thyroid dataset the pairwise p-values between the three nuclei populations, namely normal, FA, and FTC are calculated. For the normal vs. FA challenge the p-value of the projected data on our subspace is $p = 0.0232$ while it is equal to $p = 0.1349$ for the method introduced in [10]. For the two other challenges (i.e. normal vs. FTC and FA vs. FTC) the p-values are zero to numerical precision for both methods. In the cytoplasm to nuclei translocation dataset, similarly, the p-values between the two populations, namely 0.00-7.81nM and 15.63-250nM, are calculated to be zero to numerical precision for both methods.

S1. F. J, Friedman, S. Steppel, JW. Tuckey. A nonparametric procedure for comparing multivariate point sets. *Stanford Linear Accelerator Center Computation Research Group Technical Memo*, 153, 1973.

S2. M. F. Schilling Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81: 799-806, 1986.

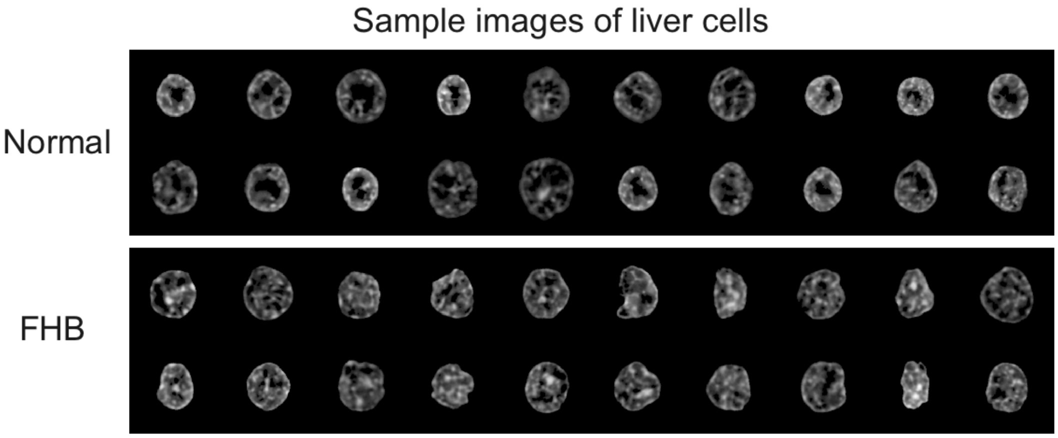


Fig. 7. Sample images of liver cells, showing normal (top) and FHB (bottom) phenotypes.

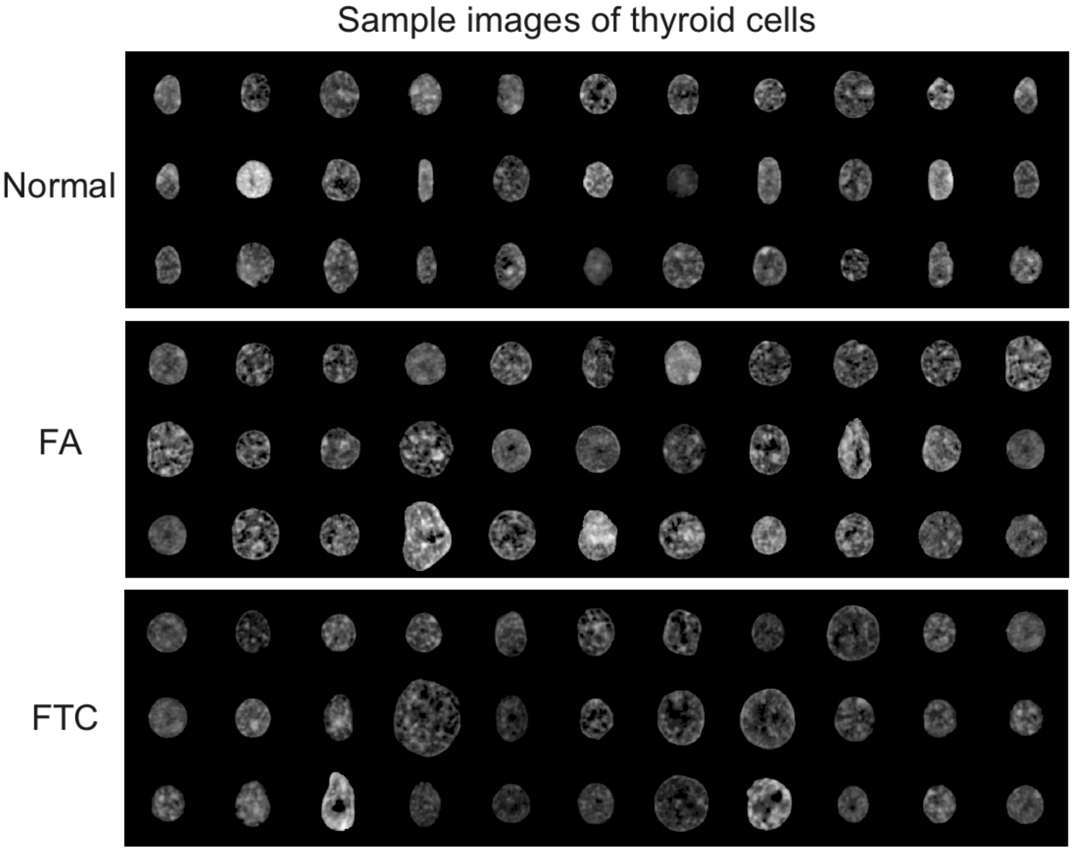


Fig. 8. Sample images of thyroid cells, showing normal (top), FA (middle), and FTC (bottom) phenotypes.

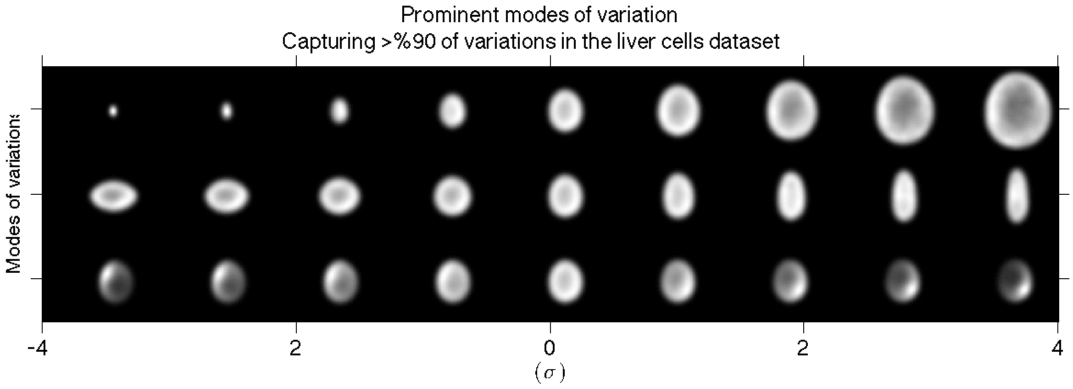


Fig. 9. The top three modes of variations obtained from applying PCA to the linear transport embedding of the thyroid dataset. The horizontal axis represents units of standard deviation of chromatin variation in a particular mode of variation. The first mode of variation demonstrates that the overall size of the nucleus is the dominant variation in the data set. The second mode suggests that the second most abundant variation in the data set is elongation. The third variation corresponds to the shifts in the center of mass of chromatin distribution. The demonstrated variations captures roughly 90% of the total variations in the dataset.

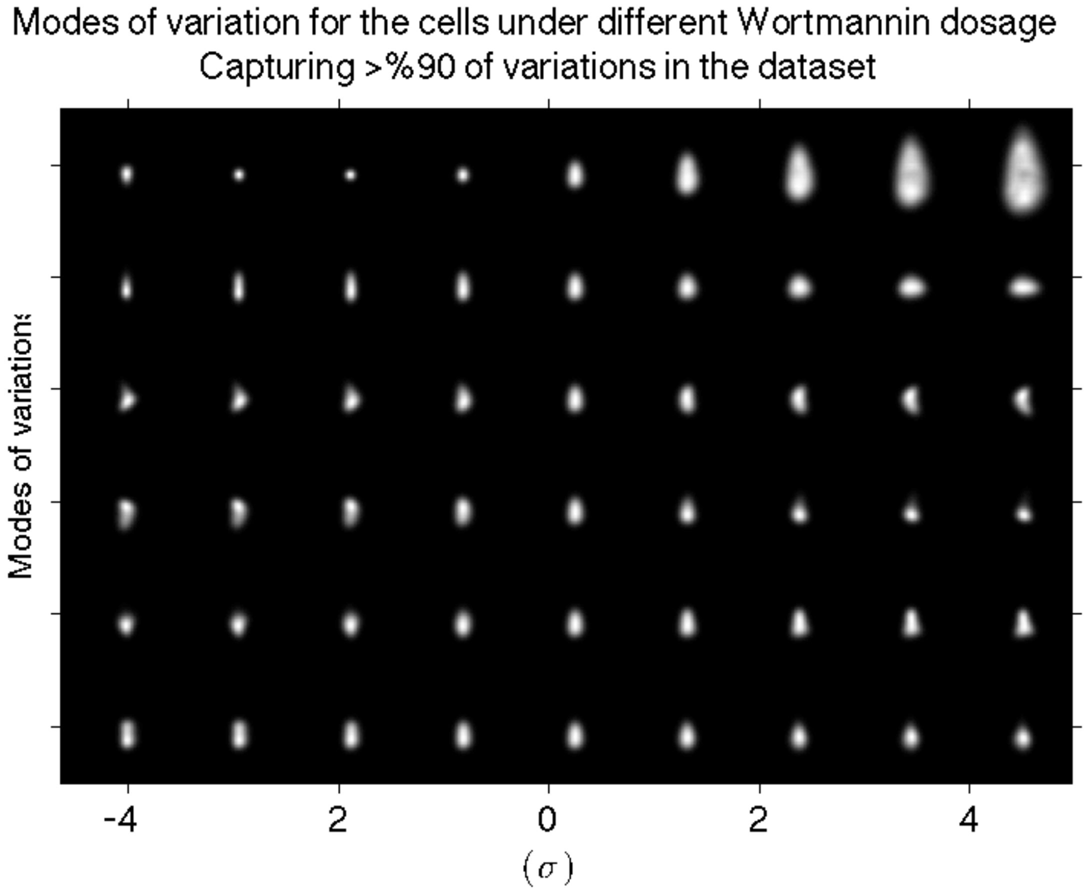


Fig. 10. The top six modes of variations obtained from applying PCA to the linear transport embedding of the U2OS dataset. The horizontal axis represents units of standard deviation of chromatin variation in a particular mode of variation. The first mode of variation demonstrates that the overall size of the nucleus is the dominant variation in the data set. The demonstrated variations explain roughly 90% of the variance in the dataset.

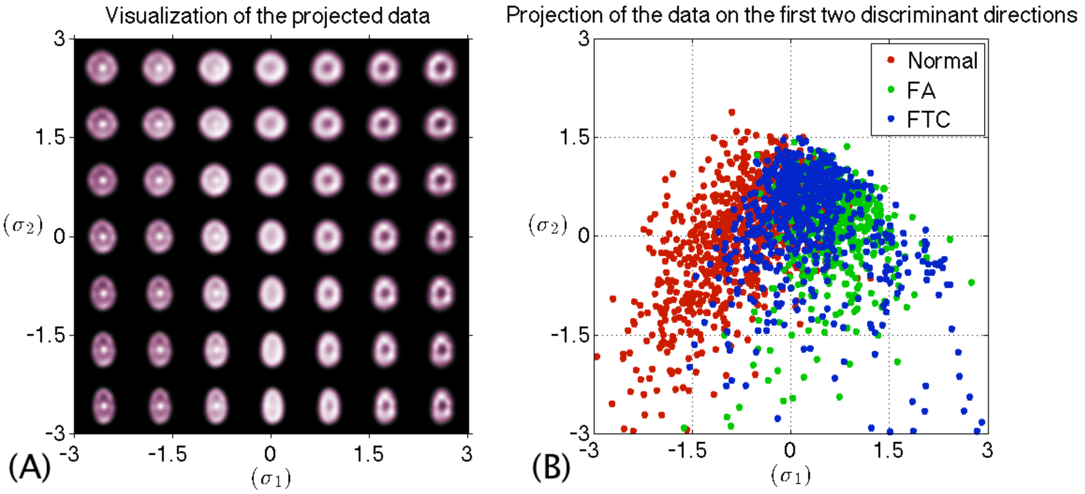


Fig. 11. (A) Demonstration of the visual difference between the normal, FA and FTC sub-populations. The horizontal and vertical axes represent the first and second most discriminant directions, respectively. The axes are in units of standard deviation of the projection on the corresponding discriminant direction. (B) Corresponding representation of the image data in (A). It can be seen that the FA and FTC cases overlap quite a lot and the visual discrimination of these cases based on chromatin distribution is difficult.

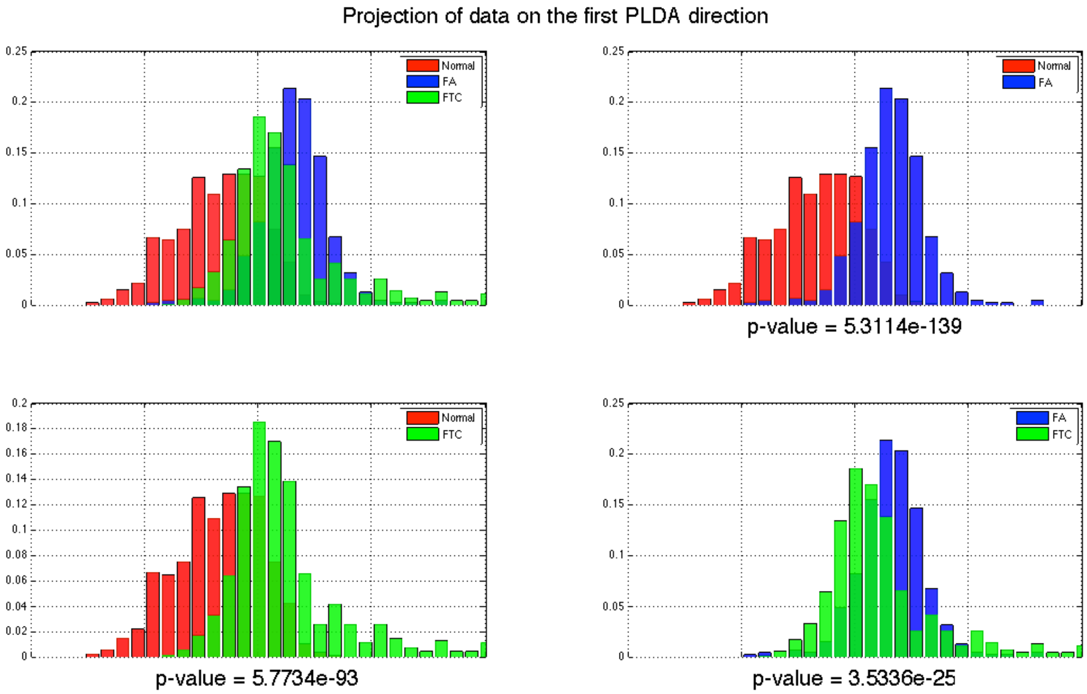


Fig. 12. A pairwise discrimination between the histograms in the three classes in the thyroid datasets. The discriminant direction is the top discriminant direction as presented in Fig. 4

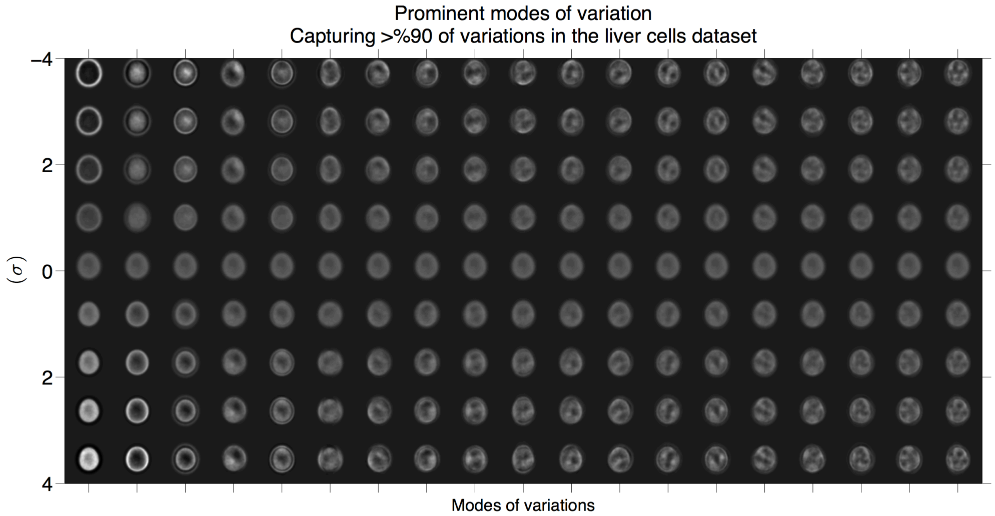

Prominent modes of variation
Capturing >%90 of variations in the liver cells dataset
−4
2
(σ) 0
2
4
Modes of variations

Fig. 13. The top nineteen modes of variations obtained from directly applying PCA to images in the liver dataset. The vertical represents units of standard deviation of chromatin variation in a particular mode of variation. The demonstrated variations explain roughly 90% of the variance in the dataset. As seen from the figure, the variation of size and shape is completely missed here, with only bands of aliasing artifacts explaining the variation.

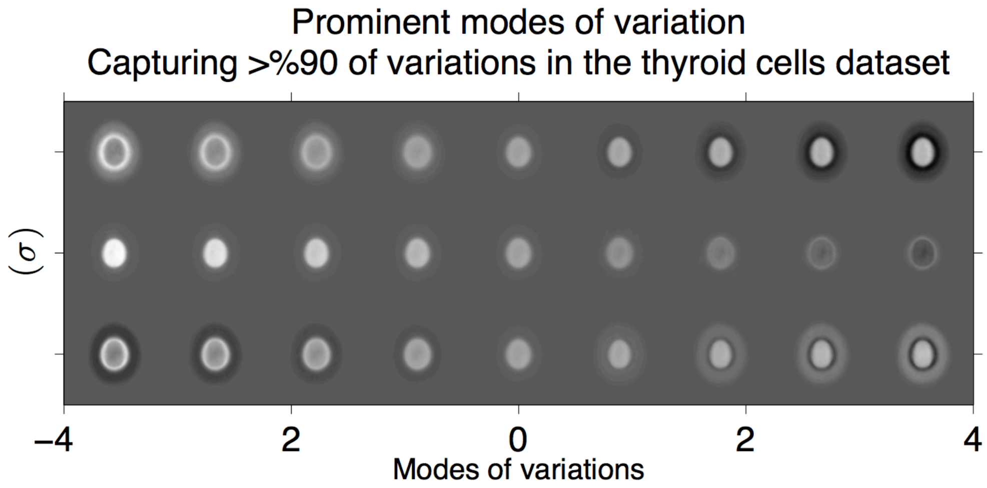

Prominent modes of variation
Capturing >%90 of variations in the thyroid cells dataset
(σ)
−4
2
0
2
4
Modes of variations

Fig. 14. The top three modes of variations obtained from directly applying PCA to the images in the thyroid dataset. The vertical represents units of standard deviation of chromatin variation in a particular mode of variation. The demonstrated variations explain roughly 90% of the variance in the dataset. Again, the band-like artifacts are completely meaningless with respect to capturing any real biological phenomena.

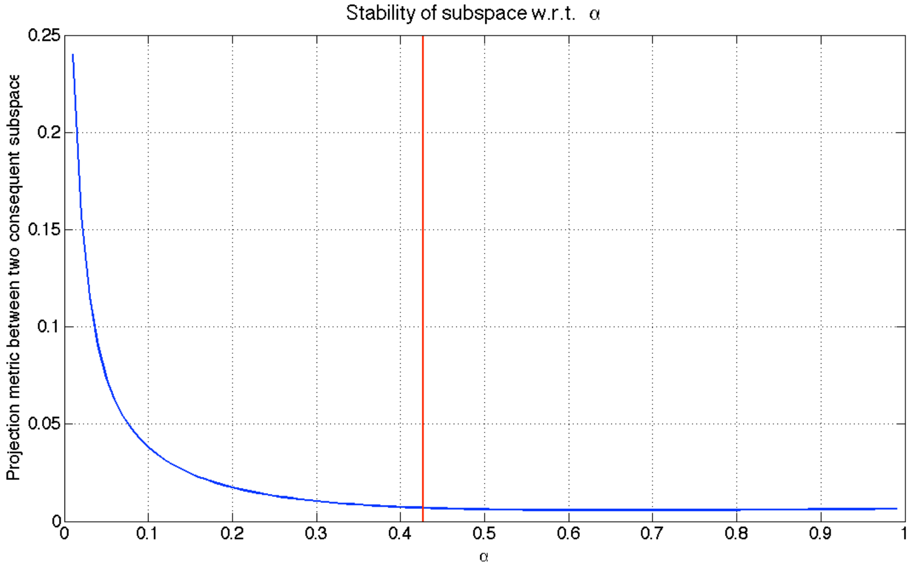
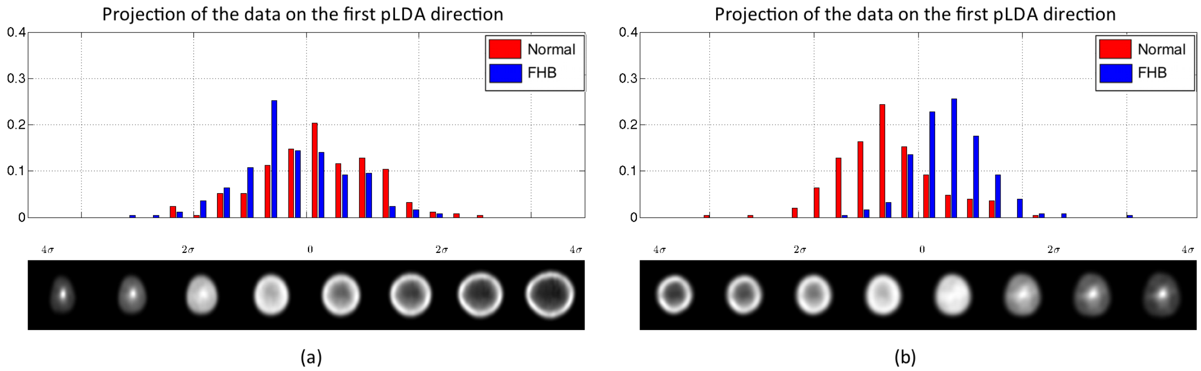


Fig. 15. The projection metric of the consequent subspaces as a function of α. Note that the blue curve is the projection metric which shows that increasing α will lead to stabilizing the penalized LDA subspace. The red line shows the α corresponding to the half life of the blue curve.



(a)

(b)

Fig. 16. The projection of the data onto the pLDA direction calculated from (a) the method proposed by Wang et al. (b) our proposed method.

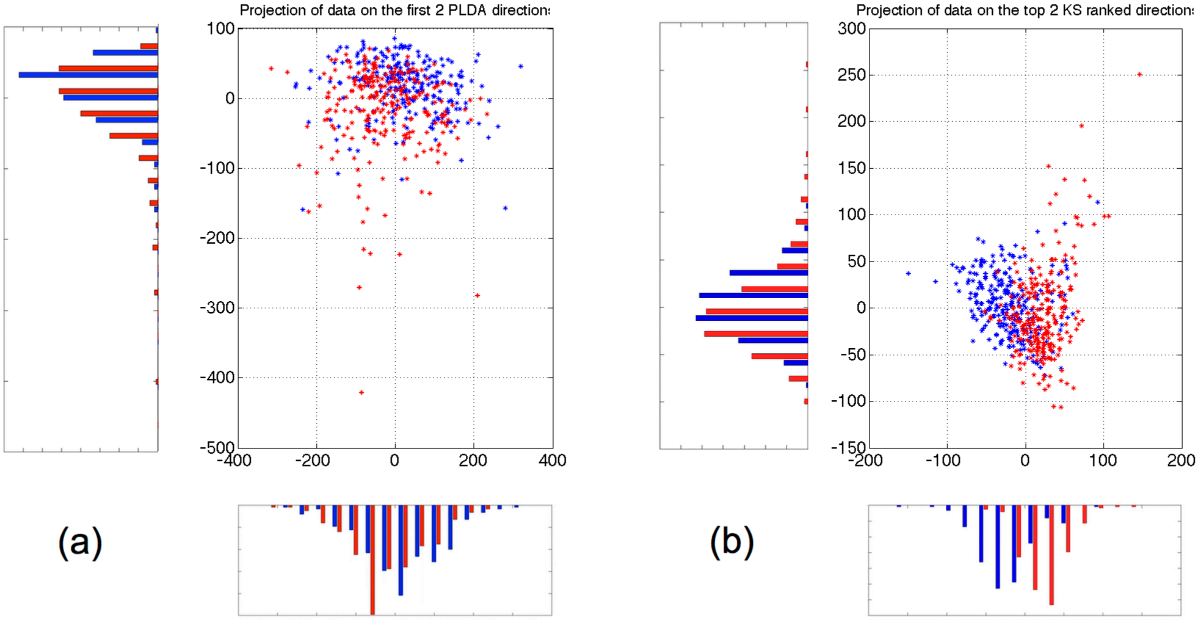


Fig. 17. The projection of the Liver data onto the pLDA subspace calculated from (a) the method proposed by Wang et al. (b) our proposed method. Red corresponds to FHB and blue corresponds to normal.

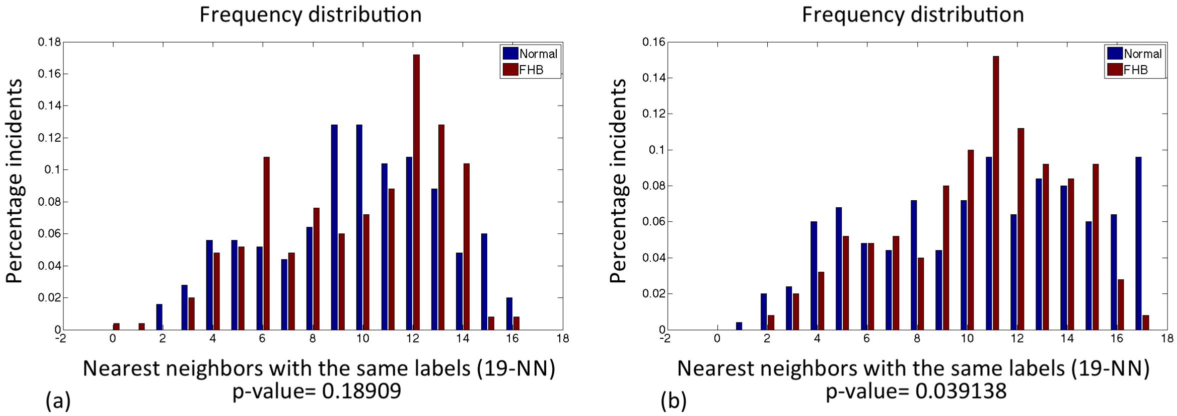


Fig. 18. The frequency distributions calculated from the projected data (Liver) onto pLDA subspace calculated from (a) the method proposed by Wang et al. (b) our proposed method. Red corresponds to FHB and blue corresponds to normal.