Multidimensional profiling of cell surface proteins and nuclear markers

Ju Han, *Member, IEEE,* Hang Chang, *Member, IEEE,* Kumari Andarawewa, Paul Yaswen, Mary Helen Barcellos-Hoff, and Bahram Parvin, *Senior Member, IEEE*

***Abstract*— Cell membrane proteins play an important role in tissue architecture and cell-cell communication. We hypothesize that segmentation and multidimensional characterization of the distribution of cell membrane proteins, on a cell-by-cell basis, enable improved classification of treatment groups and identify important characteristics that can otherwise be hidden. We have developed a series of computational steps to (i) delineate cell membrane protein signals and associate them with a specific nucleus; (ii) compute a coupled representation of the multiplexed DNA content with membrane proteins; (iii) rank computed features associated with such a multidimensional representation; (iv) visualize selected features for comparative evaluation through heatmaps; and (v) discriminate between treatment groups in an optimal fashion. The novelty of our method is in the segmentation of the membrane signal and the multidimensional representation of phenotypic signature on a cell-by-cell basis. To test the utility of this method, the proposed computational steps were applied to images of cells that have been irradiated with different radiation qualities in the presence and absence of other small molecules. These samples are labeled for their DNA content and E-cadherin membrane proteins. We demonstrate that multidimensional representations of cell-by-cell phenotypes improve predictive and visualization capabilities among different treatment groups, and identify hidden variables.**

***Index Terms*— Multidimensional profiling; evolving fronts; Voronoi tessellation; iterative scalar voting; E-cadherin; ionizing radiation.**

I. Introduction

Cell surface proteins, such as E-cadherin, regulate cell-cell interactions and physical properties of tissues. E-cadherin is a calcium-dependent cell adhesion molecule that influences differentiation and tissue structure, forming adherens junctions between epithelial cells and communicating with the actin cytoskeleton through associated intracellular proteins. As an endpoint, E-cadherin has been studied extensively, since it appears to function as a barrier to metastasis. Loss of E-cadherin has been associated with (i) increased motility, (ii) cancer progression and metastasis, and (iii) increased resistance to cell death [1]. Since down-regulation of E-cadherin is an important endpoint for quantitative systems biology, we hypothesize that detailed quantitative representation of the E-cadherin signals provides important clues for understanding the effects of different biological perturbations. Furthermore, we reason that representation of E-cadherin on a cell-by-cell basis, coupled with morphological and structural features obtained by other imaging probes, will provide a multidimensional representation that can be mined to improve predictive capability. This paper introduces a novel method for characterizing the E-cadherin signal on a cell-by-cell basis and demonstrates that multidimensional representation of imaging data is advantageous for (i) characterizing heterogeneity, (ii) identifying features that are not visually obvious to human observers, (iii) reducing the number of imaging probes that are needed for differentiating phenotypes associated with different sets of experimental treatments, and (iv) visualizing multidimensional representation of spatial data in the same way that expression data are presented.

With the exception of [2], few studies have been published quantifying membrane signals for high-content screening. Even in this case, few details of the methodology are provided. Furthermore, biological samples are limited to HeLa cells that are known to have well-behaved size and shape features. Complexities and challenges associated with quantifying cell surface protein patterns originate from (i) variation in background, (ii) signal discontinuity, (iii) nonuniformity in the width and strength of the signal, and (iv) nonuniformity in nuclear size and shape features. Although others [3], [4], [5], [6] have proposed multidimensional phenotypic representations, their analyses do not include plasma-bound features.

Our computational protocol avoids traditional ad hoc steps in favor of model-based geometric methods to delineate subcellular regions, associate cell surface protein signals with particular cells, and drive a multidimensional representation of each cell for further analysis. The choice of a particular computational methodology originates partly from the imaging instrument, and partly from the image signature across a wide range of data sets. In our experiment, samples were imaged with wide-field microscopy as opposed to a confocal system. As a result, out-of-focus light and signal accumulation along the Z-axis demand a computational method that is invariant to (i) signal strength, and (ii) the thickness of a membrane-bound protein up to a scale. The main advantages of wide-field microscopy are high-throughput imaging and the fact that it is routinely available. However, because of out-of-focus light, wide variations in signal strength, the three-dimensional structure of membrane proteins, threshold selection, which is difficult and unreliable, has to be performed locally, and the size of the neighborhood has to be dynamically adjusted for robust performance. Furthermore, an E-cadherin signal may not be continuous and maintain gaps in its signature. As a result, a desirable methodology should be able to complete perceptual gaps to retain continuity for subsequent readouts of fluorescent signals.

In this paper, the E-cadherin signal is coupled with labeled nuclear regions so that information about each cell can be preserved. The protocol consists of five major steps: nuclear segmentation, detection of E-cadherin signals on a cell-by-cell basis, feature extraction, feature selection, and discriminant analysis, as shown in Fig. 1. First, each nucleus is identified using an edge-based method, and then grouped subject to convexity and continuity. Unlike thresholding, edge-based methods have an improved immunity to nonuniformity in the fluorescent signal, thus providing a more robust and accurate delineation of nuclear boundaries. Second, the E-cadherin signal is inferred by performing an iterative voting method along a continuous boundary along the cell boundary. A unique aspect of this technique is in the topography of the voting kernel, which is iteratively refined and reoriented. However, unlike our earlier paper [7], where voting was performed along a radial direction with a kernel topography

Ju Han, Hang Chang, Paul Yaswen and Bahram Parvin are with Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA (email: parvin@media.lbl.gov).

Kumari Andarawewa is currently with University of Virginia, Charlottesville.

Mary Helen Barcellos-Hoff is currently with New York University Langone School of Medicine.

The Research was supported by the Low Dose Radiation Research Program, Office of Biological and Environmental Research, U.S. Department of Energy, and the National Aeronautics and Space Amdinistration Grant No. T6275W, NASA Specialized Center for Research in Radiation Health Effects, Grant No. DE-FG03-01ER63240.

designed to infer center of mass, voting for this paper was performed along tangential direction (e.g., normal to the gradient), and the kernel topography was designed to enhance continuity; this method has an excellent noise immunity and is tolerant to perturbations in scale. Furthermore, the membrane signal is registered with the corresponding nuclear region by an evolving front. Third, approximately 400 features corresponding to morphology (e.g., size, aspect-ratio, bending energy of a contour), structure (e.g., texture features), intracellular organization (e.g., fluorescent intensity and its derived features), and intercellular organization (e.g., relationship between cells represented as an attributed graph) are computed for each cell. These measurements are stored in a schema that captures their relationship, order, and cardinality. Fourth, a minimal subset of features from the full multidimensional representation is selected to maximize class separability, whereby classes correspond to different treatment groups. Finally, the discriminating and predictive capability of an optimal feature subset is evaluated using the *Holdout* method and discriminant analysis. Both linear and non-linear methods of discriminant analysis have been employed for comparative analysis [8], [9].

II. Materials and Methods

A. Cell culture

MCF10A (ATCC, Manassas,VA) were cultured in serum-free medium. 0.4 ng/ml recombinant human TGFβ ($R\&D$ Systems) was added at the time of plating and was irradiated 4-5 hours postplating. Cells were grown for 6 days.

B. Radiation treatment and sample preparation

γ radiation treatments were performed with a 5600 curie source of 137-Cs for γ-radiation over a dose range of 0-2 Gy. For high-LET irradiation, MCF10A was irradiated with 1 GeV/amu Fe^{56} ions (from the NASA Space Radiation Laboratory of the Brookhaven National Laboratory) over a dose range of 0-1 Gy. For comparative reasons, dosages of equivalent relative biological effect (RBE) were used as determined through toxicity studies.

C. Immunostaining and image acquisition

Cells were grown on LabTek 8-well chamber slides, fixed with 80% methanol and stained for E-cadherin (BD Biosciences). Nuclei were counter-stained with DAPI (4',6-Diamidino-2-Phenylindole) using 0.5 ng/ml. Samples were imaged using a wide-field epifluorescent microscope with a 40X objective.

D. Nuclear segmentation

Nuclear segmentation enables the context (e.g., reference) for quantifying structural and morphological features on a cell-by-cell basis. However, as a result of sample preparation and fixation, fluorescent signals of adjacent nuclear regions overlap and form a clump. It is important to quantify the phenotypic signature at the individual cell level by partitioning a clump of cells. First, our computational protocol delineates isolated nuclear regions through edge detection, boundary completion (closure), and convexity testing. Next, it partitions touching cells by applying a series of geometric constraints [10]. The basic idea is that nuclear geometry is almost convex, and that at the intersection of the overlapping boundaries, folds (points of maximum curvature) are formed. Thus, by grouping folds that are formed by a closed contour, a convex partition can be inferred. This technique is iterative, and has been shown to be effective in segmenting touching nuclei.

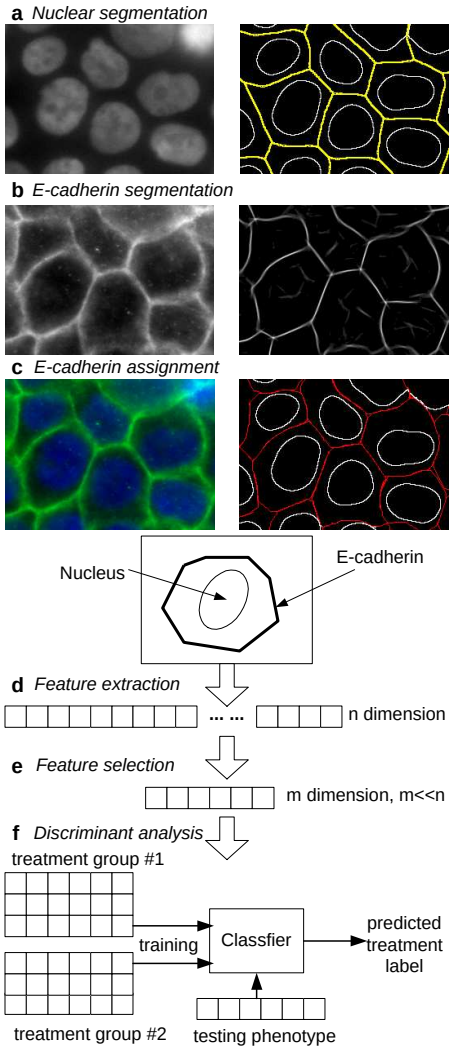


Fig. 1. Multidimensional representation of nuclear and E-cadherin responses for discriminant analysis of iron and gamma irradiation. (a) Each nucleus is segmented using an edge-based technique with a geometric convexity optimization approach for improved reliability. Voronoi region of each cell is then established (yellow boundary). (b) E-cadherin signal is inferred through an iterative voting method. (c) E-cadherin signal is assigned to the corresponding cell. (d) Each cell is represented with morphological and structural features. Furthermore, spatial organization of these features is also captured. (e) An optimal subset of features is selected for maximizing class separability. (f) The classifier is designed with the training feature data and used to predict the treatment condition for any given phenotype at the single cell level.

E. Segmentation of the E-cadherin signals

Two critical steps are involved in segmenting the membrane-bound protein. In the first step, the membrane-bound protein is accentuated and enhanced. In the second step, the membrane-bound protein is assigned to individual nuclear features. The first step utilizes iterative tangential voting to remove noise, enhance signal, and complete perceptual gaps. The second step utilizes evolving fronts for the assignment process. Iterative voting, which was introduced in our earlier paper [7], successfully identified approximate locations corresponding to the center of mass for each round object. In this paper, radial voting has been extended to tangential voting with a set of new topographic kernels and a policy for orienting kernel direction for best performance. The main advantage of iterative voting is a superior noise immunity and completion of perceptual gaps so

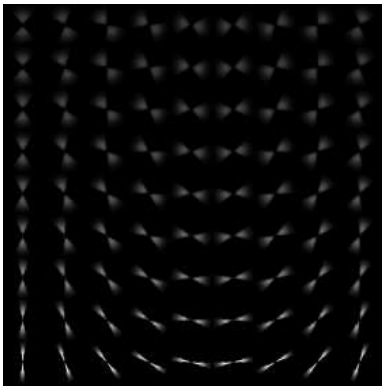


Fig. 2. A sample of kernel topography: Oriented kernels for inference of continuity are bidirectional, and their energy dissipates as a function of distance. Initially, the energy is dispersed (top row), but becomes more focused (bottom row).

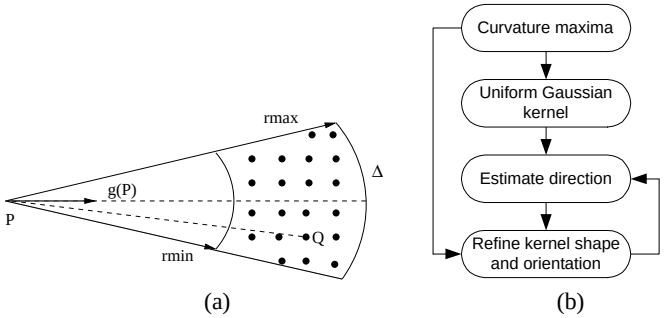


Fig. 3. (a) Redirection of the kernel in the next iteration assuming that Q is the maximum response; and (b) general fbw of the algorithm.

that membrane proteins remain continuous along the cell boundary. However, membrane signals may be absent or retain a very low level of fluorescence signal. The evolving front assures that (i) in the absence of sufficient signal levels, smoothness of membrane protein around the nuclear is preserved, and (ii) corresponding membrane-bound proteins are assigned to each nucleus in a systematic fashion. By starting from an initial condition that is derived from Voronoi tessellation between neighboring nuclei and an enhanced membrane signal as the external force, the evolving front is constrained to converge to a smooth continuous boundary that is overlaid on top of the membrane signal.

• *Iterative tangential voting:* The membrane signals correspond to the negative curvature maxima at a given scale within the image space. But curvature features are noisy and may suffer from undesirable artifacts. The process is initiated by voting with a Gaussian kernel at each image feature point. Let $F(x_o, y_o)$ be the curvature feature at location (x_o, y_o) in the image. Let (x_n, y_n) be a point in the neighborhood of (x_o, y_o) that can be influenced with a kernel applied at position (x_o, y_o). The initial voted image is then represented as

$$V(x_n, y_n) = \sum_{(x_n, y_n) \in Neighbor(x_o, y_o)} \{F(x_o, y_o) * G_{(x_o, y_o)}(\sigma)\} \tag{1}$$

The refinement of the voted image is iterative, involving the application of a more focused kernel at the next iteration along the α direction.

$$\alpha = \arctan \frac{V_{yy} - K_{max}}{V_{xy}} \tag{2}$$

where V_{yy} and V_{xy} are the local derivatives of the voted image, and K_{max} is the maximum curvature computed from the Hessian of the voted image. The shape of the kernels, shown in Fig. 2, indicates whether the energy distribution of the kernel is focused or dispersed. Initially, the energy is dispersed; however, at each consecutive iteration, the energy becomes more focused and at the same time the kernel orientation is redirected along the direction of maximum response, as shown in Fig. 3a. The entire process is shown in Fig. 3b.

These voting kernels are pre-computed and indexed for rapid retrieval.

$$V(x_n, y_n) = \sum_{(x_n, y_n) \in Neighbor(x_o, y_o)} \{F(x_o, y_o) * Kernel(\sigma, \theta, \alpha)\} \tag{3}$$

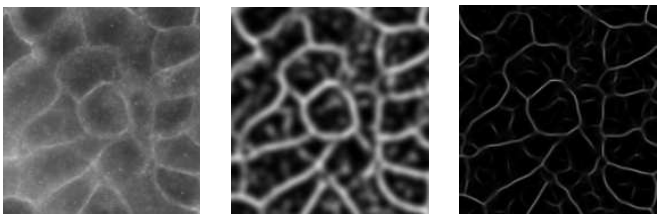
Iterative voting shares a common thread with variational methods in imaging [11], which rely on establishing proper geometric constraints and then regularizing the solution. In this case, geometric constraints are expressed in the *shape* of the voting kernel, and the regularization is embedded in the smoothness of kernels. The iterative process leads the solution to its local minima by searching for the maximum response in a local neighborhood.

Iterative Voting

1) *Initialize the parameters:* Initialize $r_{\max}$, which represents the maximum distance that a feature value is projected spatially, as shown in Figure 3. It has a maximum value in the neighborhood of pixel p, and decays as a function of its proximity to its origin. Initialize $\Delta_{\max}$ and a sequence $\Delta_{\max} = \Delta_N < \Delta_{N-1} < \cdots < \Delta_0 = 0$, which correspond to the spread of the voting kernel, as shown in Figure 3. As shown in Figure 2, Δ has an initial large spread (top row), and becomes more focal (at the bottom). Set $n := N$, where N is the number of iterations.
2) *Initialize the saliency feature image:* Define the feature image $F(x, y)$ to be the local external force at each pixel of the original image. The external force is set to the maximum (negative) curvature, which corresponds to the membrane signal.
3) *Initialize the voting magnitude:* Apply the isotropic voting of Equation 1.
4) *Update the voting direction:* Compute the Hessian of the voted landscape and construct an orientation map based on Equation 2.
5) *Refine the angular range:* Let $n := n - 1$, apply Equation 3, and repeat steps 4-5 until $n = 0$.

• *Behavior of Iterative Voting:* Figure 4 shows a small region of an image, which has been cropped to demonstrate the behavior of iterative voting in more details. It is clear that iterative voting (i) removes ambiguities associated with the location of the peak along the ridge, (ii) sharpens and enhances the ridge topography, so that the top of the ridge is no longer flat and has a dominant curvature, (iii) diffuses noise and spurious structures while enhancing the actual signal, and (iv) improves continuity, in some cases through the coarse-to-fine applications of kernels.

• *Evolving fronts:* The next step of the computational process is to design an initial condition, and define additional constraints for robust segmentation. The initial condition is derived from the region of the space identified by the segmented DNA stain, presented earlier. This is based on region-based Voronoi tessellation of the nuclear mask, which generates a curvilinear partition of the image space, shown in Fig. 1a. Standard Voronoi tessellation divides the space between neighboring vertices with linear boundaries. On the other hand, the region-based Voronoi tessellation divides the space between neighboring blobs along curvilinear boundaries. Formally, let N_i



(a) (b) (c)

Fig. 4. Iterative tangential voting: (a) a small region of the original image, (b) voting results after one iteration, and (c) voting results after 9 iterations. Iterative voting thins the location of the signal, reduces spurious noise, and improves continuity.

correspond to the $i^{th} \in [0, K)$ nuclei in the image, $q \in N_i$, and p be a point in the image. Then the Voronoi region is defined by $V_i = \{p|dist(p, N_i) < dist(p, N_j)j \in \{0, 1, \ldots, K-1\}$ and $j \neq i\}$, where $dist(p, N_i) = \min_{q \in N_i} |p - q|$. Computationally, Voronoi regions are computed from two distance maps (e.g., chamfer images) corresponding to (i) an individual target nucleus, and (ii) nuclei in the immediate neighborhood of the target nucleus. The zero difference between the two distance maps indicates the Voronoi region.

Initiating from the Voronoi region, assignment of the cell surface protein is computed by optimizing an evolving front where external forces are defined by the gradient vector field [12]:

$$E = \int_0^1 \frac{1}{2}(\alpha|X'(s)|^2 + \beta|X''(s)|^2) + E_{ext}(X(s))ds \quad (4)$$

where, $X(s) = [x(s), y(s)], s \in [0, 1]$, is the curve representation. The first and second terms ensure smoothness through stretching and bending. The third term attracts the curve towards a derived representation of the cell surface protein marker, which is a function of the voted image. The evolving front corresponds to

$$X(s,t) = \alpha X''(s,t) - \beta X'''(s,t) - \nabla E_{ext} \quad (5)$$

where, $\nabla E_{ext} = -V$ and $V(x, y) = (u(x, y), v(x, y))$ is the gradient vector flow that minimizes the energy functional

$$\epsilon = \int\int \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2|V - \nabla f|^2 dxdy \quad (6)$$

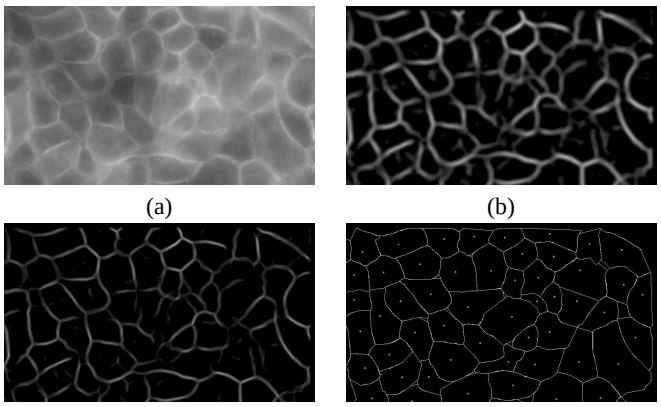
where $f(x, y)$ is a skeletonized representation of the voted image.

An example was selected from our data set to demonstrate the behavior of iterative tangential voting; the results are shown in Fig. 5. In addition to iterative tangential voting, segmentation through evolving fronts is also demonstrated.

F. Evaluation of the segmentation method

From a functional perspective, the choice and design of the algorithms reflect how the solution is constrained. For example, (i) region-based tessellation constrains membrane boundaries to reside within a small ribbon around the nuclei; (ii) voting along the tangential direction fills in potential gaps that can leak information from one cell to its neighbor; and (iii) evolving fronts initiated from Voronoi tessellation are attracted to the voted membrane boundaries while ensuring continuity, which would be lost otherwise. Let's assume that the membrane signal is partially absent, i.e., it is lost as a result of a specific treatment. Then evolving front, from the initial Voronoi tessellation, will remain stationary because there is no attracting force.

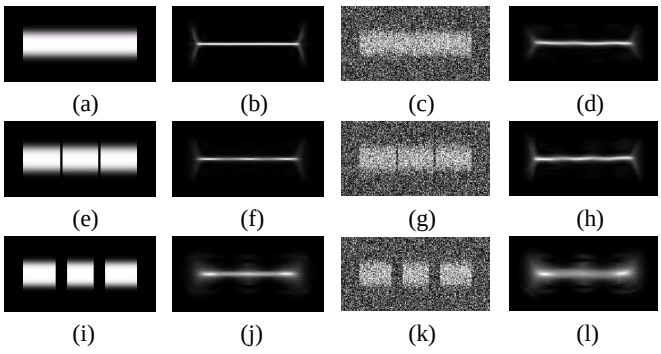
From a quantitative and error analysis perspective, segmentation of membrane-bound protein is evaluated with synthetic data, where



(a) (b) (c) (d)

Fig. 5. Delineation of membrane-bound protein for a 2-D cell culture model: (a) original image; (b) initial voting landscape; (c) final voting results corresponding to the enhanced boundaries; and (d) assignment of surface protein signals to each nuclear region.

ribbon-shaped objects with gaps have been created and noise added. Such an approach provides controlled conditions for direct evaluation, as shown in Fig. 6. In these examples, a ribbon emulating the signature of a membrane-bound protein is used for visual analysis, which indicates excellent noise immunity and gap-filling properties. Furthermore, quantitative analysis of the synthetic images of Fig. 6(a,c), tabulated in Tables I and II, indicates that iterative voting has an improved delineation accuracy. However, the true test of any approach remains to be tested on real data, in which our evaluation can only be qualitative. According to our observations, a membrane signal is always correctly segmented if the sample preparation and fixation is flawless. In some cases, following sample preparation and fixation, multiple cells may be stacked on top of each other, rendering any kind of analysis – even manual – invalid in any case.



(a) (b) (c) (d)
(e) (f) (g) (h)
(i) (j) (k) (l)

Fig. 6. Performance of the iterative voting method with added noise at $SNR = 0$ and perceptual boundaries: (a-d) signal delineation in noise free and noisy synthetic object; (e-h) signal delineation in an object (without and with added noise) with small perceptual gaps; and (i-l) signal delineation in an object (without and with added noise) with large perceptual gaps.

G. Multidimensional representations of cellular features

Phenotypic signatures are computed from every imaging probe that labels an organelle or expression of a specific protein. In this case, three distinct feature sets of morphology, structure, and fluorescence signals are extracted from each marker. For example, in the case of a marker associated with the nuclear region, morphological features of shape such as area, aspect ratio, orientation are computed. Some

TABLE I

AVERAGE ERROR IN ESTIMATING MEDIAL AXIS OF A PERCEPTUAL RIBBON INDICATES AN EXCELLENT NOISE IMMUNITY. THE ERROR IS MEASURED WITH RESPECT TO GROUND-TRUTH AS A FUNCTION OF DISTANCE TO THE MEDIAL AXIS.

	width = 10	width = 20	width = 30
SNR = 0	1.18	1.19	1.0
SNR = 10	0.31	0.28	0.33
SNR = 20	0.09	0.15	0.11

TABLE II

KURTOSIS OF THE VOTED SIGNAL AS A FUNCTION OF NOISE AND THE WIDTH OF THE RIBBON INDICATES ITERATIVE VOTING SIGNIFICANTLY IMPROVES SPATIAL ACCURACY.

	width = 10	width = 20	width = 30
un-voted signal	5.1	0.9	1.1
SNR = 0	20.3	22.0	13.7
SNR = 10	34.5	33.6	26.9
SNR = 20	35.7	34.9	30.0

of these features are computed by fitting an ellipse to the shape features for a more accurate representation. Other shape features correspond to bending energy at multiple scales, which are computed from bounding contours. Structural features correspond to textural attributes that are detected from first-, second-, and third-order derivatives of oriented Gaussian filters [13]. These oriented filters capture responses of inherent image features at multiple scales:

$$\begin{aligned} G_1^\theta &= G_x \cos\theta + G_y \sin\theta \\ G_2^\theta &= G_{xx} \cos 2\theta + 2G_{xy} \sin\theta \cos\theta + G_{yy} \sin 2\theta \\ G_3^\theta &= G_{xxx} \cos 3\theta + 3G_{xxy} \sin\theta \cos 2\theta \\ &\quad + 3G_{xyy} \sin 2\theta \cos\theta + G_{yyy} \sin 3\theta \end{aligned} \tag{7}$$

where $G_x = \partial G(x,y)/\partial x$, $G_y = \partial G(x,y)/\partial y$, and G(x,y) is a 2-D Gaussian function. These derivatives are computed by varying σ of the Gaussian in both dimensions independently. Finally, the fluorescence signal is quantified at global and local scales. While global representation relies on the average signal within the organelle of interest, local representation characterizes how the fluorescence signal is spatially distributed within the nuclear mask. An example of this representation is the change in the fluorescence signal along the radial direction originating from the center of the mass. Since the texture feature vector is rather large, its dimensionality is reduced through principal component analysis (PCA) for subsequent analysis. We opted not to apply the PCA to the entire representation, since the physical meaning of the feature set will be lost during the projection operation. A total of 324 texture features are computed and, through PCA dimensionality reduction, 8 projected features that account for 99% of the total variance are retained for further analysis. Finally, computed features are normalized with a zero mean value and variance of one across all treatment groups.

H. Feature selection and discriminant analysis

Our feature selection method ranks feature sets based on a measure of class separability. Here, the class separability is defined as the ratio of the determinant of mix-class to within-class scatter matrices. This measure will take a large value when samples are well clustered around their class means and the clusters of different classes are well separated. Mix-class scatter matrix is defined as the covariance matrix of the feature vectors with respect to the global mean, and within-class scatter matrix is defined as the covariance matrix of the feature vectors with respect to the mean of each class [8]. The *Holdout* policy is used to evaluate performance of discriminant analysis, where half of the data is randomly selected for training and testing. The process is then repeated to ensure that the classification performance is not compromised by a specific set of training samples.

I. Computational and implementation

Image analysis and feature extraction software has been developed in C++ with QT graphical user interface. Each image has 1000×1000 pixels with 2 channels at 12-bit resolution per pixel. On a very moderate desktop computer (e.g., 2.4 Ghz CPU and 2G of RAM), nuclear segmentation takes 5 seconds, E-cadherin segmentation requires 45 seconds, and feature extraction requires 6 minutes. The bulk of time for feature extraction is due to the computation of texture fields, which are computed by varying θ and scales in Equation 7. Our software was designed to read OME [14] image formats and should be portable to any other Linux platform. We plan to make our data publicly available following the publication of this paper.

III. RESULTS

A. Experimental design

Our study involved a multifactorial experiment in which radiation of two different qualities (e.g., iron, gamma) were applied separately in combination with Transforming Growth Factor β1 (TGFβ) to cultured MCF10A cells. Radiation qualities were varied to examine whether this parameter influences cellular responses independently of toxicity. TGFβ, which belongs to a family of cytokines and modulates cellular responses to radiation [15], [16], [17], was added to mimic an effect of stromal cells on radiation response in tissues. The Barcellos-Hoff lab has postulated that ionizing radiation (IR) alters cell phenotypes, which in turn contribute directly or indirectly to carcinogenesis [18]. IR activates various, multiple signaling pathways depending on the cell type, radiation dose and cell status [19]. IR also affects the activity or abundance of proteases, growth factors, cytokines, and adhesion proteins that are involved in tissue remodeling [20]. TGFβ is activated following IR, and in turn mediates some cellular and tissue radiation responses [15], [17]. Although TGFβ is considered to be a potent tumor suppressor in nonmalignant and pre-malignant tissue, principally through its ability to arrest growth and trigger apoptosis, numerous reports show that TGFβ can switch to a tumor promoter during neoplastic progression [21]. Additionally, the progeny of irradiated nonmalignant human mammary epithelial cells (HMEC) cultured with TGFβ exhibit compromised morphogenesis, polarity, and growth control when cultured in a reconstituted laminin-rich extracellular matrix [22]. The data set used for this analysis consisted of a total of 13 treatment groups with 20 to 80 images in each group and up to 6000 cells per group. Table III summarizes the different treatment groups.

TABLE III

EXPERIMENTAL VARIABLES

Sham			
TGFβ			
Iron irradiation	0.1Gy + TGFβ 0.2Gy + TGFβ 0.5Gy + TGFβ 1Gy + TGFβ 1Gy	Gamma irradiation	0.03Gy + TGFβ 0.1Gy + TGFβ 0.4Gy + TGFβ 1Gy + TGFβ 2Gy + TGFβ 2Gy

B. Visualization of the phenotypic profiles

Segmentation enables a multidimensional representation of the phenotypic profile on a cell-by-cell basis. These measurements are then aggregated and ranked for comparing and visualizing different treatment groups. However, such a multidimensional representation generates a massive amount of multidimensional data requiring a systematic evaluation, because different views of the data can reveal different insights. There are three complementary components for revealing differences between experimental treatments. The following scripts have been developed to interact with the database to query these viewpoints:

• *Feature ranking* identifies the best feature subset for visualizing features that can discriminate between different treatment groups, as shown in Table IV. This step also paves the way for visualizing features that contribute to classification.

TABLE IV

TOP-RANKED FEATURE COMBINATIONS FOR DISCRIMINATING DIFFERENT CELLULAR PHENOTYPES AMONG ALL TREATMENT GROUPS. PC STANDS FOR PRINCIPAL COMPONENT.

	Features	Discriminating power
1-feature	(1) Mean E-cadherin signal	2.1832
	(2) Total E-cadherin signal	1.7691
	(3) Nuclear texture PC #2	1.5306
	(4) Variance of E-cadherin signal	1.4724
	(5) Nuclear size	1.2398
	(6) Nuclear texture PC #1	1.2312
	(7) Nuclear texture PC #8	1.1841
2-feature	(1) + (3)	3.3555
	(1) + (4)	2.8346
	(2) + (3)	2.6733
	(1) + (2)	2.6170
	(1) + (7)	2.5832
	(1) + (6)	2.5661
3-feature	(1) + (3) + (4)	4.2184
	(1) + (3) + (6)	4.0437
	(1) + (3) + (7)	4.0333

• *Feature distribution* enables the visualization of a heterogeneous feature subset (representing a biophysical property) through single or multidimensional histograms, as shown in Fig. 7, 8, and 9. As a result, distribution shift, tightness of the distribution (kurtosis), distribution uniformity (e.g., number of modes in the distribution), and overlap of computed features in two or more treatment groups can be observed. For example, Fig. 7 shows loss of E-cadherin in irradiated samples, which is accompanied by an increase in the peakedness (kurtosis) of the distribution. The relationship between the loss of E-cadherin and heterogeneity of the membrane signal is expected; however, Fig. 7 also indicates that at equivalent radiation doses in the presence of TGFβ, loss of heterogeneity with Fe is higher than with γ radiation. This is an observation that can only be quantified through detailed analysis on a cell-by-cell basis, and appears to be dependent on the presence of TGFβ.

Multidimensional distributions of Fig. 8 and 9 indicate that treatment groups can be differentiated. It is quite interesting to note that phenotypic profiles of control, γ, and Fe form unique clusters (Fig. 9) in the 3-D space. Fig. 9 indicates that texture features corresponding to the chromatin structure are differentially expressed between two irradiation qualities; it also illustrates how hidden features may lead to the formation of a new hypothesis for new experimental design and mechanistic understanding.

• *Heat maps* provide a massive reduction and summarization of a large data set for quick understanding of the phenotypic responses.

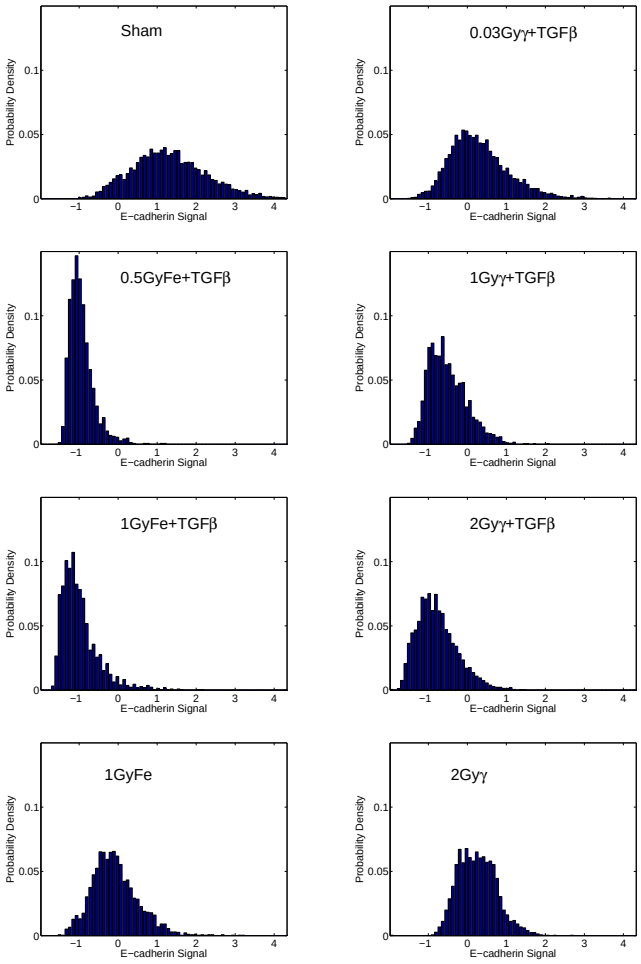


Fig. 7. Distribution and dose response of E-cadherin signal on a cell-by-cell basis: Samples were treated with indicated doses of iron and gamma irradiation to examine relative biological effects on E-cadherin expression. This feature is normalized for a zero mean value and standard deviation of 1 across treatment groups.

Such maps are often used in genome-wide visualization of the mRNA data; however, their utility is extensible. Heat maps enable an immediate view of a specific feature being up- or down- regulated as experimental factors are varied. Fig. 10 summarizes mean and individual cell responses in each treatment group, for 348 images. This map indicates that nuclear size is increased as a result of TGFβ treatment; texture features corresponding to the chromatin remodeling are differentially expressed for γ and Fe irradiation, and variations of the E-cadherin signal is more down-regulated in Fe than in γ. The latter is potentially due to the localized damage induced by Fe irradiation.

In summary, through the complementary utility of the above components, heterogeneity and hidden variables (e.g., size, texture) can be systematically identified and visualized.

C. Differences between treatment groups

In order to evaluate whether different treatment groups can be separated, three experiments are performed to determine (i) how well a treatment group can be discriminated against the control; (ii) how well treatment groups at equal toxicity dosage can be discriminated; and (iii) how hidden variables can be identified. The method is based on discriminant analysis so that selected features can then be examined for their biological properties.

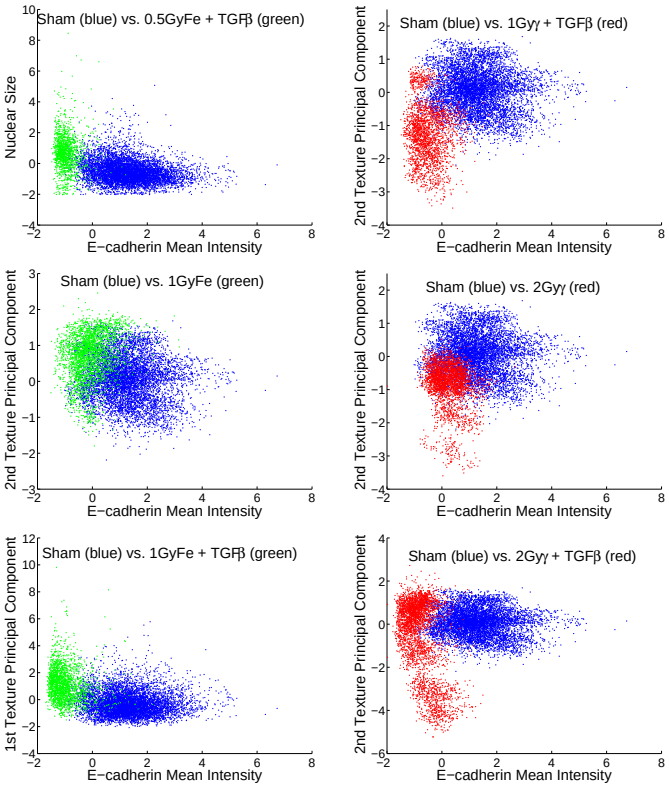


Fig. 8. Scatter plots of feature distribution of single cells in control and different irradiation treatment groups. Each feature is normalized with $\mathcal{N}(0,1)$ across treatment groups.

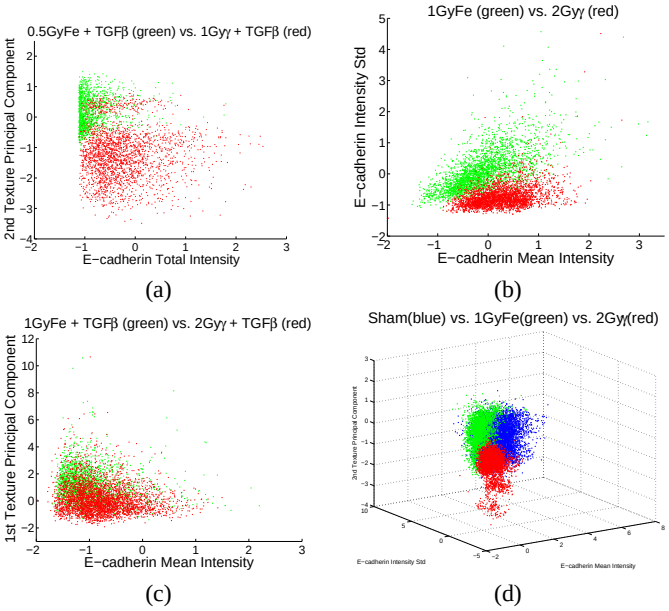


Fig. 9. Scatter plots of feature distribution of single cells in both iron and gamma irradiation at equal toxicity dosage. Each feature is normalized with $\mathcal{N}(0,1)$ across treatment groups.

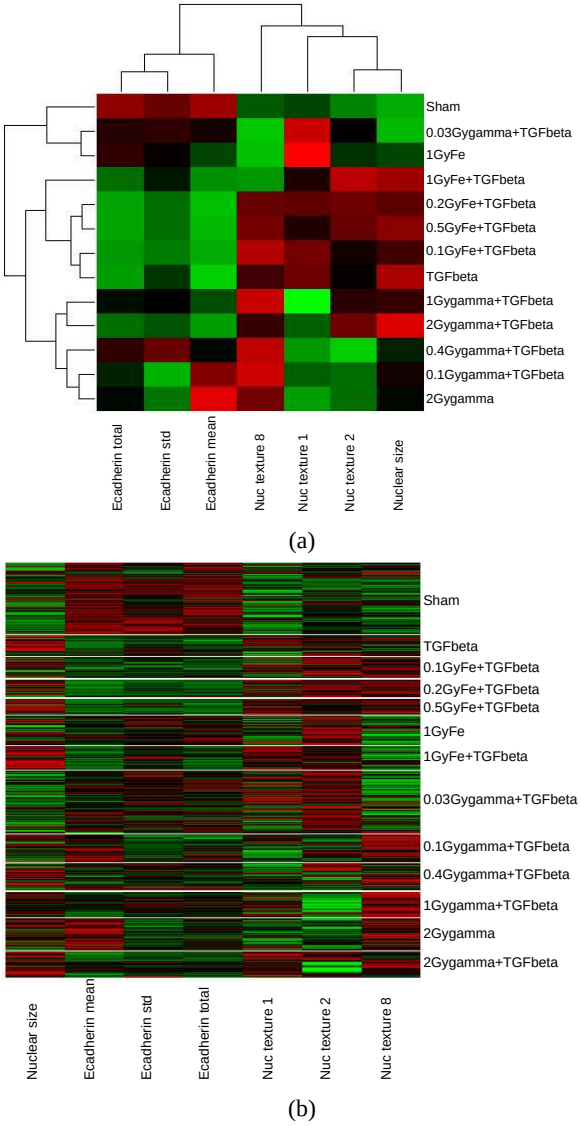


Fig. 10. Heat map of top 7 features with respect to the 13 treatment groups on (a) a group-group basis, and (b) a cell-cell basis.

(I) Table V summarizes classification accuracy between treated and control groups using single or multiple features. In most cases, a single feature is sufficient for discrimination; however, in the absence of TGFβ with a high dosage of γ, additional features can contribute to an improved classification. The number of features is clearly dependent on a specific treatment group; however, the richness of our feature set also contributes to the reduction of the number of needed measurements.

(II) Using the same strategy, we evaluated Fe and γ irradiation at equal toxicity dosage; the results are shown in Table VI. Equal toxicity dosage has been determined through survival analysis. In this case, combining representations based on quantifications of the labeled probe and computed textured features results in an improved predictor for separating treatment groups. For example, variations of E-cadherin signals in each cell are important indicators for 1Gy of Fe versus 2Gy of γ, and global florescence analysis hides the differences in the dose-response curves.

The above two experiments generally indicate a single feature may be sufficient for classification; however, this is not known ahead of time, and it is important from a biological perspective to assess what

TABLE V

SEPARABILITY OF SHAM VERSUS IRRADIATED SAMPLES COMPUTED THROUGH COMBINATORIAL FEATURE SELECTION AND LINEAR DISCRIMINANT ANALYSIS.

Number of features	1	2	3
Sham vs. 0.5GyFe+TGFβ	90%	91%	93%
Sham vs. 1GyFe+TGFβ	91%	93%	94%
Sham vs. 1GyFe	79%	83%	85%
Sham vs. 1Gyγ+TGFβ	86%	93%	95%
Sham vs. 2Gyγ+TGFβ	90%	92%	93%
Sham vs. 2Gyγ	82%	90%	92%

TABLE VI

SEPARABILITY OF Fe VERSUS γ IRRADIATION AT EQUAL TOXICITY DOSAGE.

Number of features	1	2	3
0.5GyFe+TGFβ vs. 1Gyγ+TGFβ	87%	89%	91%
1GyFe+TGFβ vs. 2Gyγ+TGFβ	70%	78%	79%
1GyFe vs. 2Gyγ	91%	95%	97%

additional feature may contribute to classification. This may also be a way to establish hypotheses for subsequent experiments.

(III) Finally, Table VII summarizes separability between different treatment groups by using one single feature at a time so that a hidden variable can be identified through LDA or SVM. In general, SVM provides better classification results at the cost of significantly increased computational time; however, the best discriminating feature is invariant to the type of classifier used. It is clear that while E-cadherin features provide the strongest power for discriminant analysis, texture features perform better when discriminating Fe- and γ- treated samples at equal toxicity dosages established earlier. Since texture features are indicators for chromatin packaging within the nuclear region, new hypotheses may be generated.

IV. DISCUSSION

Our computational protocol generates a coupled multidimensional representation of the spatial features for each nucleus and the membrane-bound proteins exhibiting continuous fluorescent signals along cell surface boundaries. Our protocol is focused on the accurate segmentation of membrane proteins, its assignment to the corresponding cell, and coupled multidimensional representation of the cellular profile for classification. E-cadherin signals are heterogeneous in intensity and scale (e.g., thickness of the membrane signal) and suffers from noise and perceptual gaps. Earlier, we developed a novel method for detecting subcellular compartments through iterative *radial* voting [7]. Here, we have extended iterative voting along the *tangential* direction for enhancing membrane proteins [23], [24]. While the detection of nuclear regions is based on the saliency of the center of mass, membrane protein detection and delineation are based on continuity and inference of perceptual regions (missing signal) along the membrane boundaries. Our method is based on the iterative application of a set of specifically tuned kernels that project pertinent information along a specific direction. These kernels vote iteratively along the radial or tangential direction and project specific features such as spatial gradient or curvature along a specific direction. Next, the regularized membrane-bound delineated signal serves as an external force for an evolving front. The front is initialized by a region-based Voronoi tessellation of the segmented nuclear regions and attracted to the membrane energy. Segmented nuclei and membrane-bound proteins provide the basis for a multidimensional

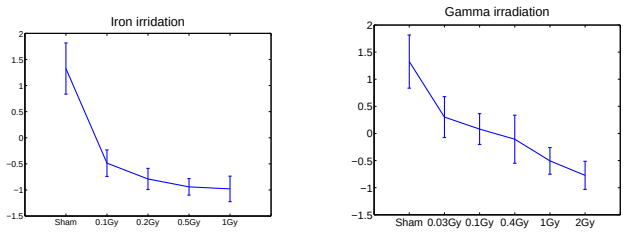


Fig. 11. Dose response of E-cadherin on a cell-by-cell basis indicates a sharper drop in the membrane protein in low dosage as a result of Fe irradiation. The error-bars correspond to the standard deviation of the signal at each dosage.

representation for profiling and classification, which is the most time consuming module in spite of a highly optimized code. More specifically, computational load is concentrated in computing texture features. However, these features are highly parallelizeable and additional speed up can be attained through limited number of scales and orientations in Equation 7. Finally, we demonstrated the application of multidimensional representation for (i) discriminating between treatment and control groups, (ii) discriminating between different treatment groups at an equal toxicity dosage, and (iii) identifying hidden variables. In the first two cases, the number of features is sequentially increased until classification performance levels off. It is important to differentiate between the performance of treated versus control groups and treatment groups at an equal toxicity dosage. In the third case, a single feature that best discriminates between different treatment groups is identified. The significance of this step is that the most dominant feature may not necessarily be the end point (e.g., membrane protein) that the sample was stained for. For example, in Table III, dominant features between iron and gamma, at an equal toxicity dosage (last row), corresponds to the texture feature that is representative of chromatin configuration. Identification of hidden features enables formation of new hypotheses.

REFERENCES

[1] U. Cavallaro and G. Christofori, "Cell adhesion and signaling: implications for tumor progression," *Nat Rev Cancer*, vol. 11, no. 12, pp. 118–32, 2004.

[2] N. Prigozhina, L. Zhong, E. Hunter, I. Mikic, S. Callaway, D. Roop, M. Mancini, D. Zacharias, J. Price, and P. McDonough, "Plasma membrane assays and three-compartment image cytometry for high content screening," *Assay Drug Dev Technol*, vol. 5, no. 1, pp. 29–48, 2007.

[3] L. Loo, L. Wu, and S. Altschuler, "Image-based multivariate profiling of drug responses from single cells," *Nature Method*, vol. 4, no. 5, pp. 445–53, 2007.

[4] M. Lamprecht, D. Sabatini, and A. Capenter, "Cellprofiler: free, versatile software for automated biological image analysis," *Biotechniques*, vol. 42, pp. 71–5, 2007.

[5] R. Murphy, "Systematic description of subcellular location for integration with proteomics databases and systems biology modeling," *Proceedings of IEEE International Symposium on Biomedical Imaging*, pp. 1052–5, 2007.

[6] C. Bakal, J. Aach, G. Church, and N. Perrimon, "Quantitative morphological signatures define local signaling networks regulating cell morphology," *Science*, vol. 22, no. 316, pp. 1753–6, 2007.

[7] B. Parvin, Q. Yang, J. Han, H. Chang, B. Rydberg, and M. Barcellos-Hoff, "Iterative voting for inference of structural saliency and characterization of subcellular events," *IEEE Trans. Image Processing*, vol. 16, no. 3, pp. 615–23, 2007.

[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, 1999.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

TABLE VII
IDENTIFICATION OF HIDDEN VARIABLES BY LDA AND SVM.

Feature	Mean E-cadherin		Variance E-cadherin		Nuclear size		Nuclear Texture 1		Nuclear Texture 2	
Classifier	LDA	SVM	LDA	SVM	LDA	SVM	LDA	SVM	LDA	SVM
Sham vs. 0.5GyFe+TGFβ	90%	97%	76%	91%	82%	81%	72%	75%	53%	81%
Sham vs. 1GyFe+TGFβ	91%	96%	69%	83%	84%	86%	79%	82%	54%	74%
Sham vs. 1GyFe	79%	83%	63%	62%	66%	71%	57%	66%	71%	54%
Sham vs. 1Gyγ+TGFβ	86%	90%	70%	81%	69%	73%	60%	73%	83%	86%
Sham vs. 2Gyγ+TGFβ	90%	92%	78%	86%	76%	76%	60%	66%	60%	75%
Sham vs. 2Gyγ	75%	77%	82%	92%	55%	69%	60%	69%	77%	77%
0.5GyFe+TGFβ vs. 1Gyγ+TGFβ	69%	71%	62%	73%	69%	63%	66%	68%	87%	87%
1GyFe+TGFβ vs. 2Gyγ+TGFβ	59%	63%	62%	65%	63%	56%	70%	70%	54%	65%
1GyFe vs. 2Gyγ	63%	53%	82%	85%	63%	66%	69%	69%	91%	92%

[10] S. Raman, C. Maxwell, M. Barcellos-Hoff, and B. Parvin, "Geometric approach to segmentation and protein localization in cell culture assays," *J Microscopy*, vol. 225, no. 1, pp. 22–30, 2007.

[11] D. Mumford and J. Shah, "Optimal approximation by piecewise smooth functions and associated variational problems," *Communication of Pure Applied Mathematic*, vol. 42, pp. 577–685, 1989.

[12] C. Xu and J. Prince, "Gradient vector fbw: A new external force for snakes," *Proc. Conf. on Computer Vision and Pattern Recognition*, pp. 66–71, 1997.

[13] R. Young, R. Lesperance, and W. Meyer, "The gaussian derivative model for spatial-temporal vision: I. cortical model," *Spatial Vision*, vol. 14, no. 3,4, pp. 261–319, 2001.

[14] G. Goldberg, C. Allan, J. Burel, D. Creager, A. Falconi, H. Hochheiser, J. Johnston, J. Mellen, P. Sorger, and J. Swedlow, "The open microscopy environment (ome) data model and xml files: open tools for informatics and quantitative analysis in biological images," *Genome and Biology*, vol. 6, no. 5, p. R47, 2005.

[15] M. Barcellos-Hoff, "Radiation-induced changes in transforming growth factor e1 and subsequent extracellular matrix reorganization in irradiated murine mammary gland," *Cancer Res*, vol. 53, pp. 3880–6, 1993.

[16] E. Ehrhart, P. Segarini, A. Carroll, and M. Barcellos-Hoff, "Latent transforming growth factor-b activation in situ: Quantitative and functional evidence following lowdose irradiation," *FASEB J.*, vol. 11, no. 12, pp. 991–1002, 1997.

[17] K. Ewan, R. Henshall-Powell, S. Ravani, M. Pajares, C. Arteaga, R. Warters, R. Akhurst, and M. Barcellos-Hoff, "Transforming growth factor-beta1 mediates cellular response to dna damage in situ," *Cancer Res*, vol. 62, no. 20, pp. 5627–31, 2002.

[18] M. Barcellos-Hoff and S. Ravani, "Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells," *Cancer Res*, vol. 60, pp. 1254–60, 2000.

[19] P. Dent, A. Yacoub, J. Contessa, R. Caron, G. Amorino, K. Valerie, M. Hagan, S. Grant, and R. Schmidt-Ullrich, "Stress and radiation-induced activation of multiple intracellular signalling pathways," *Radiat Res*, vol. 159, no. 3, pp. 283–300, 2003.

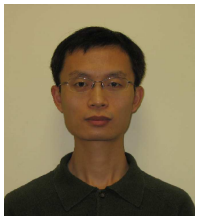
[20] M. Barcellos-Hoff, C. Park, and E. Wright, "Radiation and the microenvironment - tumorigenesis and therapy," *Nat Rev Cancer*, vol. 5, no. 11, pp. 867–75, 2005.

[21] H. Bierie, B. andMoses, "Tumor microenvironment: Tgfbeta: the molecular jekyll and hyde of cancer," *Nat Rev Cancer*, vol. 6, no. 7, pp. 506–20, 2006.

[22] C. Park, R. Henshall-Powell, A. Erickson, R. Talhou, B. Parvin, M. Bissell, and M. Barcellos-Hoff, "Ionizing radiation induces heritable disruption of epithelial cell interactions," *Proc Natl Acad Sci*, vol. 100, no. 19, pp. 10 728–33, 2003.

[23] H. Chang, K. Andarawewa, J. Han, M. Barcellos-Hoff, and B. Parvin, "Perceptual grouping of membrane signals in cell-based assays," *Proc. IEEE Int. Symp. on Biomedical Imaging:from nano to macro*, pp. 532–5, 2007.

[24] J. Han, H. Chang, K. Andarawewa, P. Yaswen, M. Barcellos-Hoff, and P. Parvin, "Integrated profiling of cell surface protein and nuclear marker for discriminant analysis," *Proceedings of IEEE International Symposium on Biomedical Imaging*, pp. 1342–6, 2008.



Ju Han received his Ph.D. degree in the Electrical Engineering Department from the University of California, Riverside in 2005. Since December of 2005, he has been a specialist at the Imaging and Informatics Lab with joint appointment at U.C. Berkeley and Lawrence Berkeley National Laboratory. His research interests are quantitative and integrative modeling of biological processes. Dr. Han is currently working on the development of computational methods for mapping chemical composition at subcellular scales.

Hang Chang received his Ph.D. from the Institute of Automation, Chinese Academy of Sciences in 2008, and he has been with the Lawrence Berkeley National Laboratory since January of 2006. His areas of research are algorithm development for biological image analysis, high performance computing, and development of end-to-end systems for high content screening. Dr. Chang leads the development and validation of quantitative methods for cell-based assays.

Kumari Andarawewa received the B.VSc. degree in Veterinary Medicine and Animal Science from University of Peradeniya, Sri Lanka, in 1998, and the M.S.S degree in Space Studies from International Space University, Strasbourg, France. She received her Ph.D. degree in Molecular Biology from the University Louis Pasteur, Strasbourg, France, in 2004. She was one of the recipients of 2005 - AACR - Women in Cancer Research Brigid G. Leventhal Scholar Award in Cancer Research", which was based on her graduate research. From 2004 to 2007, she was a Postdoctoral Fellow at the Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA in the lab of Dr Mary Helen Barcellos-Hoff. She is currently with University of Virginia, Charlottesville. The goal of her research is to understand how perturbations in the microenvironment lead to neoplasia and how to improve chemo and radiation therapy.

Paul Yaswen received his Ph.D. in Cell and Molecular Biology from Brown U. in 1984, and received post-doctoral training at the Dana-Farber Cancer Institute. He is currently a Staff Scientist in the Dept. of Cancer Biology at the Lawrence Berkeley National Lab. Dr. Yaswen has over 20 years of experience using a tractable human mammary epithelial cell culture system to model processes involved in immortal and malignant transformation of this cell type, thought to be the precursor of most human breast cancers. He has used this model to study cellular responses to specific changes under conditions where other potential variables are controlled, to distinguish local from systemic effects on cellular physiology, and to identify phenotypes that may be uniquely human and thus not amenable to study using animal models. He is currently a member of the Breast Oncology Program at the UCSF Comprehensive Cancer Center, an affiliate of the Berkeley Stem Cell Center, a preceptor in an NIH funded Biology of Aging Training Program at LBNL/UC Berkeley, and a member of the Molecular Oncology Study Section at NIH. Dr. Yaswen has over 50 publications in peer-reviewed journals.

Mary Helen Barcellos-Hoff received an undergraduate degree in Biopsychology from the University of Chicago in 1978 and a doctoral degree in Experimental Pathology from the University of California, San Francisco in 1986. Her graduate research in experimental pathology was conducted with Dr. Dennis F. Deen on determinants of brain tumor cell response to therapy and her postgraduate research concerning extracellular matrix signaling on mammary epithelial functional differentiation training was conducted at Lawrence Berkeley National Laboratory (LBNL) with Dr. Mina J. Bissell. She established her research laboratory at LNBL in 1988 to study breast cancer and ionizing radiation. She is currently Associate Professor of Radiation Oncology at New York University Langone School of Medicine and studies radiation carcinogenesis and mammary biology.

Bahram Parvin is the head of Imaging and Informatics Lab in the Department of Cancer Biology at the Lawrence Berkeley National Laboratory (LBNL), and an adjunct Professor of Electrical Engineering at U.C. Riverside where he teaches Graduate courses on bioimaging and systems biology. His areas of interests are imaging bioinformatics and integrative biology. He received his Ph.D. in Electrical Engineering from the University of Southern California in 1991, has been a member of the organizing and program committee on bioimaging and computer vision conferences, and is a senior member of IEEE. He has 4 patents and over 90 refereed publications.