The Annals of Applied Statistics
2010, Vol. 4, No. 3, 1365–1382
DOI: 10.1214/10-AOAS340

ANALYSIS OF SPATIAL DISTRIBUTION OF MARKER EXPRESSION IN CELLS USING BOUNDARY DISTANCE PLOTS

BY KINGSHUK ROY CHOUDHURY[1], LIMIAN ZHENG AND JOHN J. MACKRILL[2]

University College Cork

Boundary distance (BD) plotting is a technique for making orientation invariant comparisons of the spatial distribution of biochemical markers within and across cells/nuclei. Marker expression is aggregated over points with the same distance from the boundary. We present a suite of tools for improved data analysis and statistical inference using BD plotting. BD is computed using the Euclidean distance transform after presmoothing and oversampling of nuclear boundaries. Marker distribution profiles are averaged using smoothing with linearly decreasing bandwidth. Average expression curves are scaled and registered by x-axis dilation to compensate for uneven lighting and errors in nuclear boundary marking. Penalized discriminant analysis is used to characterize the quality of separation between average marker distributions. An adaptive piecewise linear model is used to compare expression gradients in intra, peri and extra nuclear zones. The techniques are illustrated by the following: (a) a two sample problem involving a pair of voltage gated calcium channels (Cav1.2 and AB70) marked in different cells; (b) a paired sample problem of calcium channels (Y1F4 and RyR1) marked in the same cell.

1. Introduction. Optical fluorescence microscopy (OFM) offers a high resolution view of the morphology and spatial organization of intact cells and organelles. Various proteins, nucleic acids and metabolites can be individually labeled with different fluorescent colors, giving an in vivo picture of their behavior and role in living cells [Fernandez-Gonzalez et al. (2006)]. The increasing quality and quantity of these images necessitate quantitative, indeed statistical, analysis of the inherent spatial and morphological information. In this paper we consider the problem of mapping the spatial distribution of marker expression in reference to distance from the cell/nuclear boundary.

First, we consider an experiment comparing the distribution of two different voltage-gate calcium channels (VGCC), which play an important role in linking (a) muscle excitation with contraction and (b) neuronal excitation with transmitter

Received June 2009; revised January 2010.
[1]Supported by Research Frontiers Grant 07/MATF/543 from the Science Foundation Ireland.
[2]Supported by Research Frontiers Grant 07/BMIF/548 from the Science Foundation Ireland.
Key words and phrases. Euclidean distance transform, smoothing, functional data analysis, curve registration, image texture.

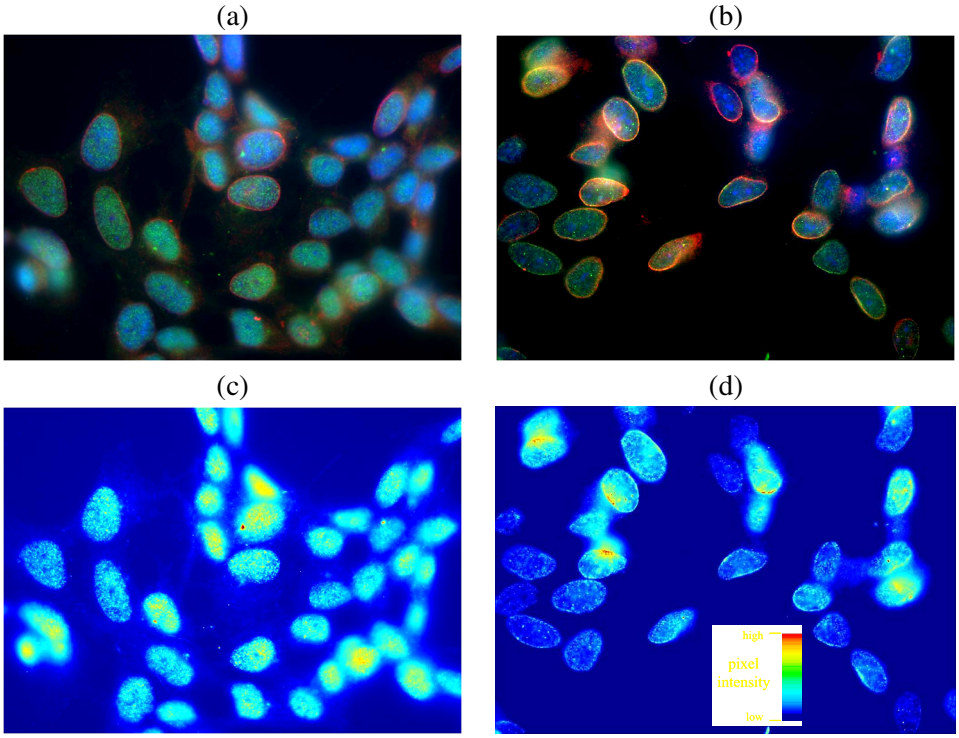

FIG. 1. *Optical fluorescence microscopy image of SH-SY5Y neuroblastoma cells. Images are* 3840 × 3072 *pixels*, 8*-bit discretization*, *with* 0.08 $\mu m \times 0.08$ μm *pixel size*, *acquired using a Nikon Eclipse E*600 *epifluorescent microscope with* 60× *objective. Images are labeled in a blue chromatin marker* (*'DAPI'*), *a red nuclear membrane marker* (*'Emerin'*) *and* (a) *a green marker* (*'AB*70*'*), *selective for either Cav*1.2 *or Cav*1.3 *VGCC's*, (b) *a green marker* (*'Cav*1.2*'*) *selective for only Cav*1.2 *VGCC*. (c)–(d) *Psuedo-color image of green* (*AB*70) *channel of image in* (a). (d) *Pseudo-color image of green* (*Cav*1.2) *channel of image in* (b).

release. New research indicates they may play a role in gene transcription [Gomez-Ospina et al. (2006)]. The VGCC's Cav1.2 and Cav1.3 are studied in the nuclei of the human neuroblastoma cell-line SH-SY5Y. The images (Figure 1) are labeled with three different fluorescent dyes: (i) a blue chromatin marker, 4,6-diamidino-2-phenylindole ('DAPI'), which essentially marks the body of the nuclei; (ii) a red mouse monoclonal antibody recognizing the nuclear matrix protein emerin which lines the nuclear membrane ('emerin'); (iii) a green antiserum which recognizes either (a) both Cav1.2 and 1.3 ('AB70') or (b) only Cav1.2 ('Cav1.2'). Details of the experiment can be found in Callinan et al. (2005). The green markers are used as proxy for presence of the VGCC. Of particular interest is the proximity of the VGCC to the nuclear membrane, which can give clues to its role in signal transmission to/from the nucleus. There appears to be considerable variability in the green marker distribution both within and across cells [Figures 1(c) and (d)]. This

suggests that comparisons across the images can only be accomplished in a distributional or average sense. Since the orientation of nuclei is modified arbitrarily during cell fixation in the slide, any analysis conducted on this data should ideally be orientation invariant.

As a second example, we compare the distributions of an intracellular calcium release channel, the type 1 ryanodine receptor (RyR1) with that of the sarcoplasmic (SR)/endoplasmic reticulum (ER) calcium ATPase (SERCA), an enzyme that pumps Ca^{2+} into the lumen of intracellular Ca^{2+} stores such as the SR and ER. In this experiment JEG-3 trophoblastic cells [Figure 7(a)] were labeled with the following: (i) DAPI (blue marker) to mark the body of the nucleus; (ii) a mouse monoclonal antibody Y1F4 (red marker) recognizing all SERCA subtypes; and (iii) a rabbit polyclonal antiserum recognizing the type 1 RyR subtype only (RyR1, green marker). Analysis of the distribution of RyRs in the trophoblastic cell-line JEG-3 is of interest because the roles of these calcium channels in nonmuscle cell types, such as these placental epithelial cells, have not been extensively characterized. In muscle cells, RyR channels play a pivotal role in coupling extracellular signals to the release of calcium from the SR/ER, which triggers activation of the contractile apparatus [Mackrill (1999)]. We anticipate that RyR1, a channel that releases Ca^{2+} from the SR/ER, would display a similar distribution to that of SERCA, the main pumping system that actively accumulates Ca^{2+} into this organelle.

From a statistical perspective, this problem involves the comparison of marker expression distributions across cells and experimental conditions. When orientation is not of interest, it is convenient to reduce the two-dimensional distribution of markers to one-dimensional profiles, plotted against a common 'distance.' This makes it easier to superimpose and visualize multiple profiles across cells on the same plot. As each nucleus/cell has a different shape and size, measurement of proximity must be adapted to the shape or 'geometry' of each individual nucleus/cell. For instance, when we consider distribution of the boundary marker emerin (red) for a typical nucleus in Figure 1(b), the profile distribution of expression generated by plotting against radial distance (from the center of the cell) appears to have a bimodal distribution [Figure 2(a)]. By contrast, when we use *boundary* distance (BD), that is, the distance of each point to the nearest nuclear boundary (Section 2), the profile distribution for the same nucleus appears to have a single sharp peak at 1 [Figure 2(b)]. The difference is because, unlike radial distance, the level sets of BD are individually adapted to the nuclear/cell boundary [Figures 2(c) and (d)]. BDs can be computed using algorithms such as Euclidean distance mapping (EDM) [Fabbri et al. (2008)] or morphological erosion [Jahne (2005)]. BDs are normalized to a common scale, for example, 1 at center and 0 at the boundary, to allow comparison across cells/nuclei of different shapes and sizes [Knowles et al. (2006)]. In this paper we consider an extension of EDM to compute BD both within and outside the cell/nucleus. We also propose the use of oversampled smoothed boundaries for computation of BD to correct for polygonization of the cell/nuclear boundary during manual identification (Section 2).

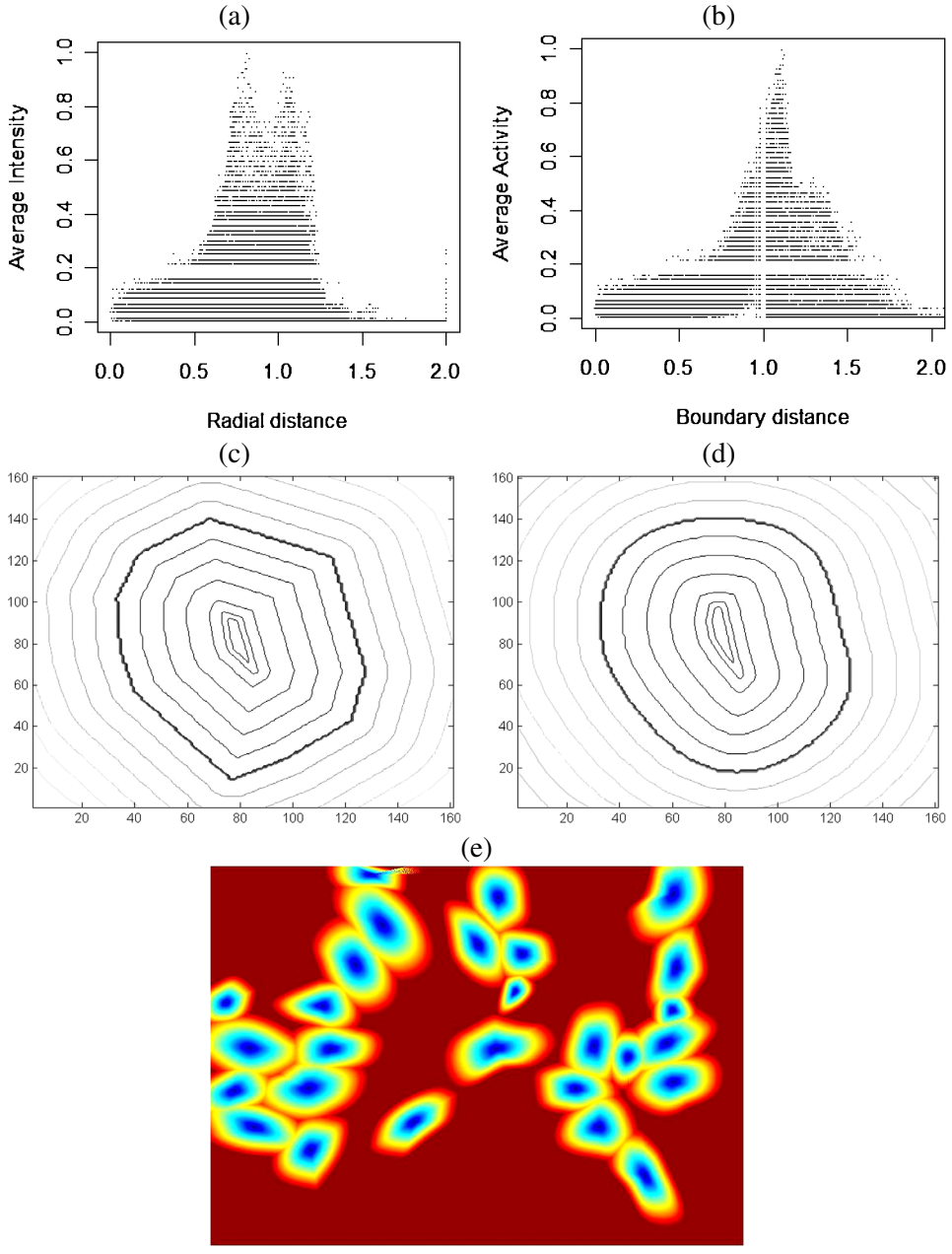

FIG. 2. (a) *and* (b) *Profile distribution of emerin* (*red marker*) *for a nucleus in Figure* 1(b). *Each point in the plots represents the observed red channel intensity at a pixel in the image. Expression of point* (x, y) *is plotted against* (a) *the radial distance from the center of the nucleus*, (b) *distance to nearest point on nuclear boundary*. (c)–(e) *Boundary distance* (*BD*) *maps* (c) *of the polygonal region representing a nucleus*, (d) *same nucleus*, *but smoothed boundary*, (e) *pseudocolor image showing orbits for all nuclei in Figure* 1(b).

Previous analyses of profile distributions from boundary distance plots have been basically descriptive [Bewersdorf, Bennett and Knight (2006); Knowles et al. (2006)]. In this paper we develop methods for improved estimation and statistical inference from profile distributions. For this purpose, we construct smooth average expression curves to summarize the profile distribution for each nucleus. In particular, we show why a linearly increasing bandwidth for smoothing is necessary (Section 3.1). Variations in light intensity across the image are compensated by scaling expression curves (Section 3.1.1). Uneven blue staining near the boundary of the nucleus can cause incorrect boundary identification, which was originally thought to affect only distances near the boundary [Bewersdorf, Bennett and Knight (2006)]. By analyzing this as an errors in variables type problem, we show that estimated BDs are biased upward. To correct for this, we realign average expression curves using an x-axis dilation prior to statistical analysis (Section 3.1.2). Next, we show how methods such as t-tests and penalized discriminant analysis can be used to describe the differences between groups of profile average expression curves (Section 3.2.2). We also use a knot-adaptive piecewise linear model to draw inferences about expression curves and their derivatives within regions of interest in the nuclei (Section 3.3). Section 4 applies this methodology to the second example. Section 5 concludes with a summary of findings and their scientific implication.

The main steps involved in BD analysis are listed below. The sections of the paper where these steps are described in detail are given in brackets:

1. Mark cell/nucleus boundaries and compute BD maps (2.1).
2. Obtain average marker expression curves for each cell (3.1).
3. Align activity curves group by scaling and shifting (unpaired: 3.1.1 and 3.1.2, paired: 4).
4. Comparison across groups using functional data analysis (unpaired: 3.2, paired: 4).
5. Comparison of expression gradients using piecewise linear models (3.3).

2. Computing boundary distances. If we represent a cell/nucleus as a point set R, the Euclidean distance transform (EDT) of a point p within R is defined as $D(R^c, p) = \inf\{d(p, q) \mid q \in R^c\}$, that is, the distance of the p from the nearest point in the complement of R. Let $d_m = \sup\{D(R^c, p), p \in R\}$ denote the 'maximal distance' from the boundary. To obtain a scaled BD map that is 0 at the 'center' of R and 1 at the boundary and extends continuously for outside R, we define

$$\mathrm{BD}(p) = \begin{cases} 1 - d_m^{-1} D(R^c, p), & p \in R, \\ 1 + d_m^{-1} D(R, p), & p \in R^c. \end{cases} \tag{2.1}$$

A number of efficient algorithms for computing the EDT have appeared over the last decade or so [Fabbri et al. (2008)] and many of these are available in standard image analysis packages such as the freely available ImageJ (http://rsbweb.nih.

gov/ij/). Contours of the BD function resemble the cell boundary for points near the boundary, but not necessarily for points deep in the interior [Figures 2(c) and (d)].

2.1. *Boundary smoothing*. To construct the BD, we first need to identify cell/nucleus boundaries using either automated segmentation methods [Jahne (2005)] or hand drawing. When segmented regions are polygonal, so too are the contours of the resulting BD map [Figure 2(c)], whereas we know that nuclear membranes have a much smoother shape. We use periodic smoothing splines to smooth the boundary curve [Wahba (1975)]. The fitted curve appears to circumscribe the polygon defined by the original boundary points [Figures 2(c) and (d)]. The fitted curve is sampled at a large number of points (1000) to generate the smooth boundary. The resulting contours are typically visually more satisfactory. When BD is computed for an image with multiple nuclei, a decision rule is required to assign points to 'orbits' of particular nuclei. In Figure 2(e) an *orbit* of a particular nucleus consists of all points whose BD is smallest relative to distances to other nuclei. Nuclei/cells that lie on the boundary of the image are ignored from subsequent analysis.

3. Statistical analysis of profile distributions. Let $h(a|r)$ be the profile distribution of expression a at BD r. For analysis across nuclei, we summarize h by its conditional expectation $g(r) = \mathrm{E}[h(a|r)]$. We call g the average expression curve. If we assume g to be a smooth function of r, for a wide class of distributions h, g can be estimated by nonparametric regression of the point cloud $h(a|r)$ as a function of r [Figure 3(a)] [Silverman (1985)]. Its computation is described in Section 3.1. We then present methods for alignment and analysis of these curves across nuclei.

3.1. *Estimating average expression curves*. Assuming that the nucleus is approximately elliptical, its contours have circumference $0.5\pi er$, where r is length of the minor axis and e is the eccentricity of the ellipse. Thus, the number of points on a given constant BD contour increases linearly as a function of BD. Although the nuclei are not exactly elliptical, the accuracy of this approximation for a typical nucleus in Figure 1(b) is empirically borne out in Figure 3(b). For distances beyond the boundary ($r = 1$), this relationship need not hold, as orbits may compete for points.

In situations where the density of points, $f(r)$, is not constant, smoothing with a fixed bandwidth can pose problems. In Figure 3(c), the fixed bandwidth Nadaraya–Watson estimator $\hat{g}^{\mathrm{N-W}}(r) = \{\sum_i K(b^{-1}(r - r_i))\}^{-1}\{\sum_i K(b^{-1}(r - r_i))h(a_i|r_i)\}$ appears to be more variable near the center (low density) and have more bias near toward the boundary (high density) [Figure 3(c)]. By contrast, the 'optimal' variable bandwidth kernel smoother has bandwidth proportional to $f(r)^{-0.2}$ [Silverman (1984)]. The smoothing spline estimator $\hat{g}(r) = \arg\min\{\sum_i (h(a_i|r_i) -$

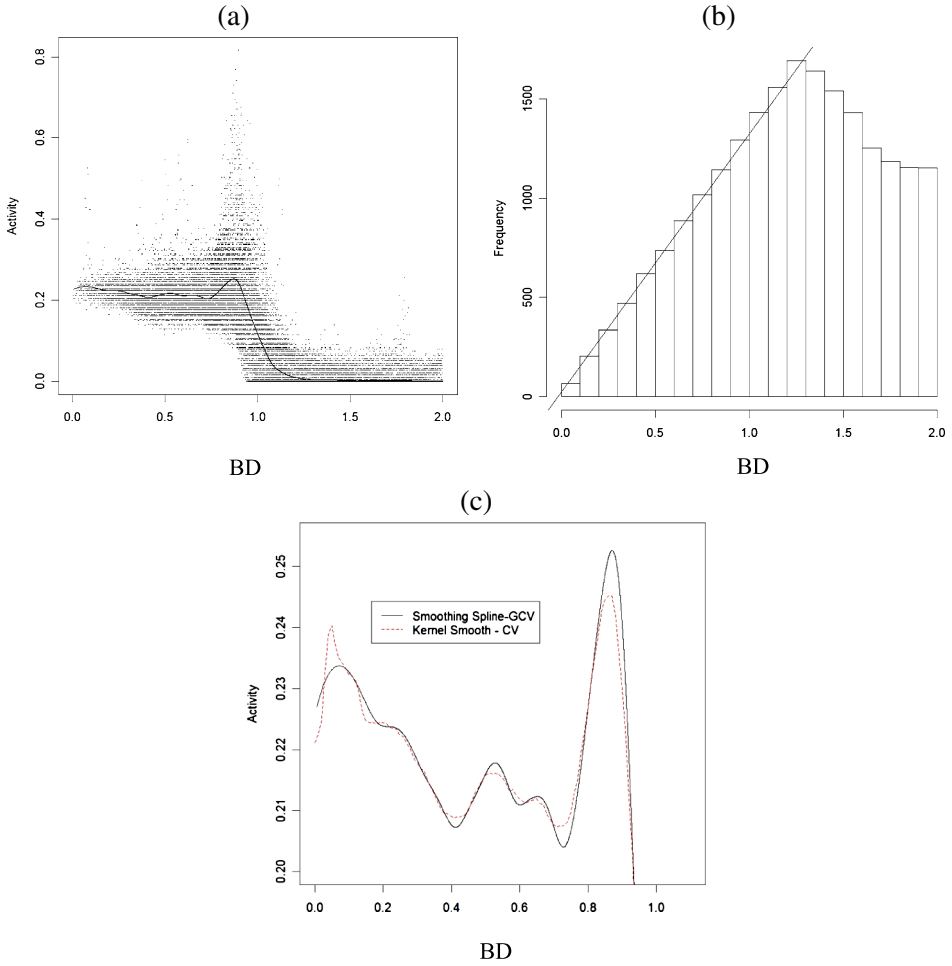

FIG. 3. (a) *Distribution of green* (*Cav*1.2) *expression against boundary distance* (*BD*) *for one nucleus in Figure* 1(b). *Line shows local average computed by smoothing spline.* (b) *Histogram of number of pixels sampled as a function of BD for radial plot in* (a). *Fitted line shows approximation to linear trend from center to nucleus boundary.* (c) *Close-up view of local average of expression shown in Figure* 1(b), *showing differences between variable bandwidth smoothing spline and fixed bandwidth kernel estimates.*

$g(r_i))^2 + \lambda \int \{g''(u)\}^2 \, du\}$ has an equivalent 'bandwidth' proportional to $f(r)^{-0.25}$, that is, almost optimal [Silverman (1984)]. It appears to produce a more satisfactory estimate [Figure 3(c)]. We will therefore use it for estimating average expression curves, with λ being chosen by generalized cross-validation. For ease of further computation, each estimated average expression curve is evaluated at a common grid of regularly spaced points $r_i = 0.01i$, $i = 1, 2, \ldots, 200$, in the interval $(0, 2]$.

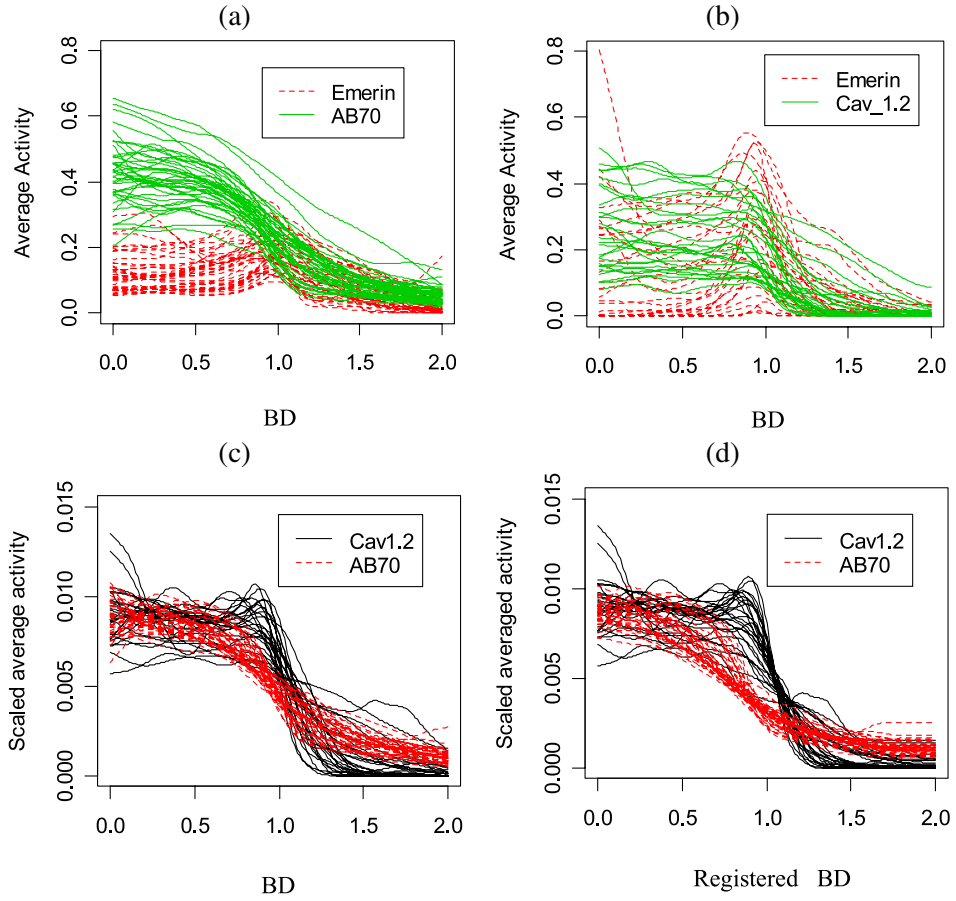

FIG. 4. *Smoothed average expression curves corresponding to*: (a) *Nuclei in Figure* 1(a). *Red denotes emerin expression and green denotes Cav*1.2 *expression*. (b) *Nuclei in Figure* 1(b). *Red denotes emerin expression and green denotes AB*70 *expression*. (c) *Scaled expression curves corresponding to Cav*1.2 (*solid black lines*) *and AB*70 (*dashed red lines*) *expression. The area under each curve equals* 1. (d) *Scaled expression curves corresponding to Cav*1.2 (*solid black lines*) *and AB*70 (*dashed red lines*) *plotted against registered BDs. Curve registration was done by individual dilation of boundary BDs.*

3.1.1. *Scaling expression curves.* In Figures 4(a) and (b) we see that the average expression curves of emerin (red marker) peak near the boundary, whereas the average activities of Cav1.2 and AB70 are high inside the nuclei and low beyond the boundary. However, there appears to be a lot of variation in the amplitudes/scales of these curves. We define scale $s_k = \int_0^2 g_k(r)\,dr$, where g_k is the average expression curve for the kth cell: it is approximated as a Riemann sum, $\hat{s}_k = \sum \hat{g}_k(r_i)$. There appears to be moderate to strong positive correlation between the red and green scale values across cells within each image: for Cav1.2

and Emerin, $\hat{\rho} = 0.77$ and for AB70 and Emerin, $\hat{\rho} = 0.69$. This suggests that at least some of the variation in scaling may be due to uneven illumination across the image, affecting both red and green channels. If, on the other hand, there had been fluoro marker sensitive effects like photo bleaching, it would most likely affect a single channel locally, producing poor correlation. To eliminate this extraneous source of variation, we normalize the curves g_k by dividing by the factor estimated scale factor $\hat{s}_k$. The scaled profiles now all subtend an area of 1: they give us the average 'distribution' of the marker in each nucleus as a function of distance. The distribution curves [Figure 4(c)] show much less intra group variation in the y-direction than the unscaled versions [Figures 4(a) and (b)].

3.1.2. *Dilation based registration*. Uneven DAPI staining leads to errors in identification of the true nuclear boundary. Assuming additive measurement errors in true boundary distances $D(R^c, p)$, we can write expected value of the resulting observed boundary distance $\mathrm{BD}_o(p)$ as follows:

$$\begin{aligned} \mathrm{E}[\mathrm{BD}_o(p)] &= 1 - \mathrm{E}\big[(d_m + \varepsilon_m)^{-1}\big(D(R^c, p) + \varepsilon_p\big)\big] \\ &= 1 - \mathrm{E}\big[(1 + d_m^{-1}\varepsilon_m)^{-1}\big(1 - \mathrm{BD}(p) + d_m^{-1}\varepsilon_p\big)\big]'. \end{aligned} \tag{3.1}$$

We further assume that the measurement errors ε_p and ε_m are i.i.d. $\mathrm{U}[-e, e]$. Taking expectations with respect to the uniform distribution, we get

$$\begin{aligned} \mathrm{E}[\mathrm{BD}_o(p)] &= 1 - \big(\ln(1 + d_m^{-1}e) - \ln(1 - d_m^{-1}e)\big)\big(1 - \mathrm{BD}(p)\big) \\ &\approx 1 - (1 + 3d_m^{-2}e^2)\big(1 - \mathrm{BD}(p)\big) \approx (1 + 3d_m^{-2}e^2)\mathrm{BD}(p). \end{aligned} \tag{3.2}$$

The first approximation in (3.2) follows from a first order Taylor series expansion. The second approximation assumes $e \ll \mathrm{BD}(p)$. Thus (3.2) shows that estimated BD have an upward bias, which can be modeled by a location independent scale factor.

In terms of observed boundary distances, we can thus write a model for the observed expression as $z_k(p) = g_k(\delta_k \mathrm{BD}_o(p)) + \varepsilon(x, y)$. Differential dilation of nuclear boundaries causes misalignment of expression curves across nuclei. To realign them, we first estimate the parameters δ_k by minimizing the within image registration sum of squares:

$$\mathit{WREGSSE} = \sum_{k=1}^{nc} \int_0^2 w(r)\big(g_k(r\delta_k) - \mu(r)\big)^2\, dr. \tag{3.3}$$

Here μ is the (unknown) mean curve across nuclei within the group (Cav1.2 or AB70), nc is the number of nuclei in each image (=27 for Cav1.2, =38 for AB70) and w is a weighting function which reflects the precision of the estimated curves (see Section 3.1). Here $w(r) = r^{0.75}$, $0 < r < 1$; $w(r) = 1$, $0 < r < 1$; $w(r) = 0$ otherwise, based on the fact that the variability of the smoothing spline estimate is proportional to $f(r)^{-0.75}$ [Silverman (1985)], where $f(r)$ is the density derived

in Section 3.1. Minimization of *WREGSSE* can be achieved through a two-step Procrustes type iterative procedure [Ramsay and Silverman (2002)]. Step 1: The group mean μ is estimated by the sample mean of the scaled expression curves g_k. Step 2: Given μ, the criterion (3.3) is separable in the δ_k, each of which can be estimated by a line search procedure. We start with an initial estimate of $\delta_k = 0$ $\forall k$ and steps 1 and 2 are iterated to convergence. In this case, iterating only 1 step of the iterative algorithm resulted in a 15% reduction of *WREGSSE* for the Cav1.2 image and 8% for the AB70 image. Further steps did not result in any significant decrease of *WREGSSE*. The registered curves $g_k(r + \hat{\delta}_k)$ appear to exhibit greater location alignment [Figure 4(d)]. Within image registration yields estimated group mean curves, $\hat{\mu}_C$ and $\hat{\mu}_A$ for Cav1.2 or AB70 respectively. These are then registered to each other by minimizing the sum of squares difference, $BREGSSE = \int_0^2 (\hat{\mu}_A(r\delta_A) - \hat{\mu}_C(r))^2\,dr$. The parameter δ_A is estimated by line search and results in a 53% reduction in *BREGSSE*. The combined dilation for nuclei in the AB70 group is thus given by the product $\delta_A\delta_k$.

3.2. *Functional data analysis* (*FDA*).

3.2.1. *Mean comparison.* Comparison of group means (calculated pointwise) shows slightly different profiles for Cav1.2 and AB70 [Figure 5(a)]. Cav1.2 expression appears to remain constant up to the cell boundary, whereupon there is a sharp drop-off, with little expression beyond the boundary. For AB70, the drop-off is more gradual and some expression appears to extend beyond the boundary. We tested the null hypothesis $H_0 : \mu_A = \mu_C$, against a general alternative using $T(r_i) = (\hat{\mu}_C(r_i) - \hat{\mu}_A(\delta_A r_i))(n_C^{-1}s_C^2(r_i) + n_A^{-1}s_A^2(\delta_A r_i))^{-0.5}$, $i = 1, 2, \ldots, 200$. Here $s_C^2(r_i)$ is the sample variance of the (registered) expression distributions for the Cav1.2 group at r_i and $n_C = 27$ is the number of cells in this group. Similar notation is used for the AB70 group, with $n_A = 38$. The significance of the test statistic was computed by means of repeated randomization: under the null hypothesis, two sets of expression distributions of size n_A and n_C were repeatedly randomly sampled ($N = 5000$ times) without replacement from the combined collection of $n_A + n_C$ pooled expression distributions from both groups. For each sample (permutation), $\sup|T(r)|$ was computed. Approximate 95% simultaneous critical levels were computed as $\pm T_{0.975} = 2.45$, which is the 97.5th percentile of the $\sup|T(r)|$ statistics across permutations. Although the test statistic does not appear to be significant near the center of the nuclei ($r = 0$), we see that the observed test statistic is above the confidence band in a region close to the boundary ($r = 1$), while it is below the confidence band outside the nucleus boundary [Figure 5(b)].

3.2.2. *Penalized discriminant analysis.* To further describe the difference between the groups, we consider the problem of discriminating between the groups using the average expression curves g_k using Fisher's linear discriminant analysis

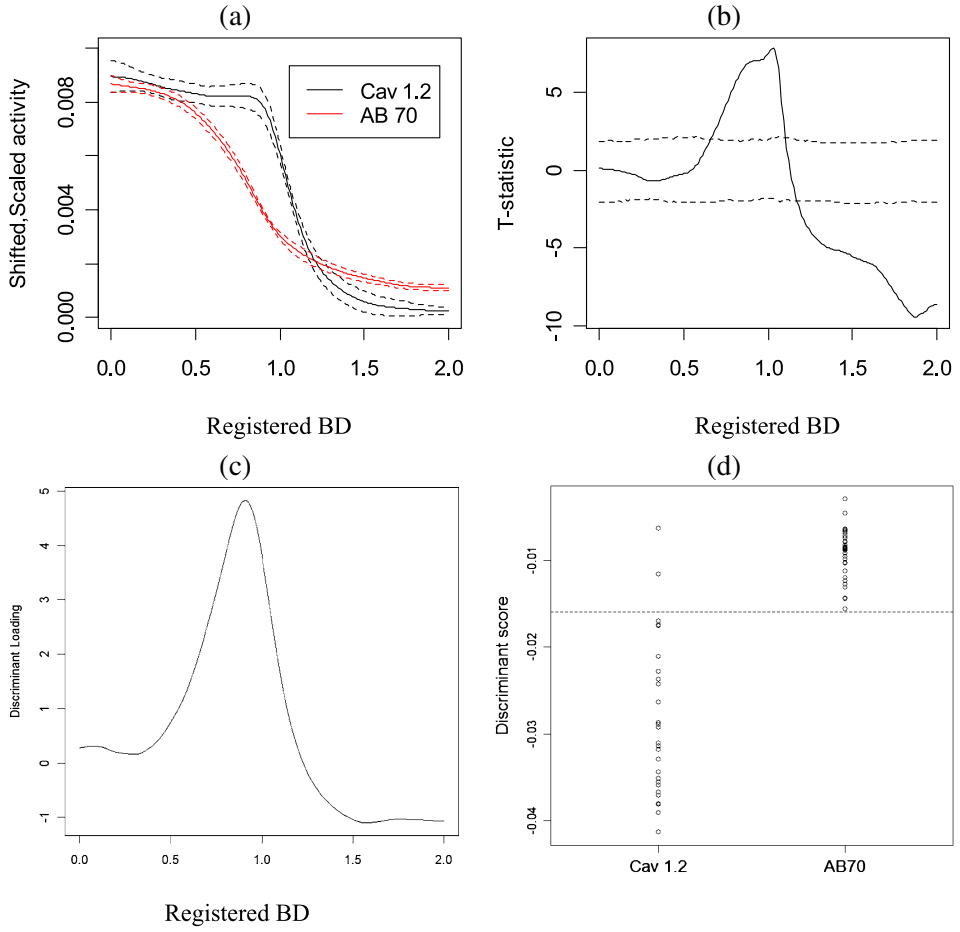

FIG. 5. (a) *Comparison of mean* (*scaled and registered*) *expression curves for Cav*1.2 (*black*) *and AB*70 (*red*) *expression. Curves are obtained by pointwise averaging across nuclei within each image. Dashed lines show pointwise* 95% *confidence intervals*. (b) *Solid line is curve of test statistic for two sample T-test, calculated pointwise, between Cav*1.2 *and AB*70 *expression curves shown in* (a). *Dashed lines denote* 95% *confidence band for test statistic, generated by repeated randomization.* (c) *Plot of coefficients of the discriminant function after regularized linear discriminant analysis between Cav*1.2 *and AB*70 *expression curves shown in Figure* 4(d). (d) *Plot of discriminant scores by group with dashed line showing optimal threshold.*

(LDA). The discriminant $\mathbf{d} = \mathbf{W}^{-1}(\hat{\boldsymbol{\mu}}_\mathbf{A} - \hat{\boldsymbol{\mu}}_\mathbf{C})$, where $\mathbf{W}$ is the within class covariance matrix, $\hat{\boldsymbol{\mu}}_\mathbf{C} = (\hat{\mu}_C(r_1), \ldots, \hat{\mu}_C(r_{200}))$, $\hat{\boldsymbol{\mu}}_\mathbf{A} = (\hat{\mu}_A(\delta_A r_1), \ldots, \hat{\mu}_A(\delta_A r_{200}))$. In classical statistics, $\mathbf{W}$ is estimated by the pooled sample variance covariance matrix, that is, $\mathbf{W} = 0.5\boldsymbol{\Sigma}^\mathbf{A} + 0.5\boldsymbol{\Sigma}^\mathbf{C}$, where $\boldsymbol{\Sigma}^A_{ij} = (27-1)^{-1}\sum_k(\hat{g}^A_k(r_i) - \hat{\mu}_A(r_i))(\hat{g}^A_k(r_j) - \hat{\mu}_A(r_j))$, $i, j = 1, \ldots, 200$, is the sample within group variance covariance matrix for the AB70 group and $\boldsymbol{\Sigma}^\mathbf{C}$ is similarly defined for the Cav1.2

group [Anderson (2003)]. However, the dimension of the expression curves (200) exceeds sample size (65), causing $\mathbf{W}$ to become singular, which causes problems when computing $\mathbf{d}$. To ensure stable inversion, we instead compute a penalized within class covariance matrix $\mathbf{W}_p = 0.5\boldsymbol{\Sigma}^{\mathbf{A}} + 0.5\boldsymbol{\Sigma}^{\mathbf{C}} + \lambda\mathbf{I}$, where $\mathbf{I}$ is a 200×200 identity matrix and λ is a regularization parameter. We use the decision rule: nucleus k belongs to AB70 if $\mathbf{d}_P^T \hat{\mathbf{g}}_k > \tau$, where $\mathbf{d}_p = \mathbf{W}_p^{-1}(\hat{\boldsymbol{\mu}}_{\mathbf{A}} - \hat{\boldsymbol{\mu}}_{\mathbf{C}})$is the penalized discriminant and τ is a predetermined threshold. A leave out one cross-validation (CV) procedure is used to choose a combination of λ and τ which jointly minimize misclassification error [Hastie, Tibshirani and Friedman (2001)]. Using grid search over a range of λ (10 values between 0.0001 and 0.1 equispaced on a logarithmic scale) and τ (10 equispaced values between 0.5 and 1.5), a unique minimum CV error of 2 misclassifications out of 65 (i.e., 3%) was obtained for $\lambda = 0.0007$ and $\tau = 1.17$ [Figure 5(d)]. The optimal penalized discriminant is practically zero from the center out, has a sharp dip near the boundary and an elevated level beyond it [Figure 5(c)]. The use of a Laplacian type penalty, suggested by [Friedman (1989)], instead of $\mathbf{I}$ does not appear to produce a well conditioned matrix $\mathbf{W}_p$ in this case.

3.3. *Flexible parametric modeling.* The analysis of the previous section has demonstrated some differences in the average expression of the two types of VGCC makers near the boundary of the nucleus and possibly beyond. In this section we attempt to better characterize these differences by fitting separate linear models to the expression in three regions: the interior of the nucleus, the nuclear boundary and the exterior, using a piecewise linear model

$$g^P(r) = \mathrm{E}[h(a|r)] = \sum_{i=1}^{3} a_i + b_i r\, I\{\kappa_i < r < \kappa_{(i+1)}\}. \tag{3.4}$$

Here $\kappa = \{\kappa_i, i = 1, 2, 3, 4\}$ are knot points with $\kappa_1 = 0$ and $\kappa_4 = 2$, $\mathbf{a} = (a_1, a_2, a_3)$ are intercepts and $\mathbf{b} = (b_1, b_2, b_3)$ are slopes. The knotpoints κ_2 and κ_3 allow flexibility in choosing the extent of the 'boundary' region. Unlike usual implementations of piecewise models, model (3.4) does not impose continuity across knots. This allows parameters for each piece to be estimated mutually independently, simplifying inference. We use the weighted least squares criterion $L(\kappa, \mathbf{a}, \mathbf{b}) = \sum w(r)(h(a|r) - g^P(r))^2$ for model fitting. The weighting function $w(r)$ is the same as used in (3.3), to account for sampling density. For given knot points, κ_2, κ_3, the criterion $L(\kappa, \mathbf{a}, \mathbf{b})$ can be fit as three separate linear models using weighted least squares. However, when two successive pieces have identical slopes and intercepts, that is, $a_i = a_{(i+1)}$ and $b_i = b_{(i+1)}$, the choice of κ_i is not unique, since any value of κ_i in $[\kappa_{(i-1)}, \kappa_{(i+1)}]$ will yield the same value of $L(\kappa, \mathbf{a}, \mathbf{b})$. To avoid this ambiguity, we instead minimize a penalized weighted least squares criterion of the form

$$L_p(\kappa, \mathbf{a}, \mathbf{b}) = \sum_r w(r)\big(h(a|r) - g^P(r)\big)^2 + \lambda P(\kappa_2 - 1) + \lambda P(1 - \kappa_3). \tag{3.5}$$

Here P is an asymmetric penalty function: $P(x) = \infty$ if $x \geq 0$ and $P(x) = x^2$ if $x < 0$. We note that P is a penalty on knot location, quite different from the smoothness penalty commonly used in function estimation [Hastie, Tibshirani and Friedman (2001)]. It enables the pieces to be kept on the correct side of the nuclear boundary and the boundary piece to be relatively short. To obtain the minimizer of (3.5), we adopt a two-stage procedure:

Step 1: For given κ_2, κ_3, we compute the minimizer of $L(\kappa, \mathbf{a}, \mathbf{b})$ by weighted least squares as $\hat{\mathbf{a}}_\kappa, \hat{\mathbf{b}}_\kappa$. These are computed across a triangular grid of knot points $\kappa_{2jm} = 0.01j$, $\kappa_{3jm} = 0.01m$, $j = 1, \ldots, 100$, $m = j, \ldots, 100$.

Step 2: The criterion $L_p(\kappa, \hat{\mathbf{a}}_\kappa, \hat{\mathbf{b}}_\kappa)$ is computed for all knot points in the grid of κ values using (3.5). The regularization parameter λ is chosen by grid search to be the smallest value which ensures unique estimation of knot points. The global minimum of $L_p(\kappa, \mathbf{a}_\kappa, \mathbf{b}_\kappa)$ is obtained by grid search over κ values.

Fits from piecewise modeling closely match the average expression curves obtained by spline smoothing [Figure 7(a)], with median $R^2 (= 1 - \sum(\hat{g}_k - \hat{g}_k^P)/V(\hat{g}_k))$ values of 0.99 for both AB70 and Cav1.2 groups.

Comparison of intercepts across groups shows no significant differences in any of the three regions [Figure 6(b)]. Comparison of slopes of each piece across groups shows no significant difference for the last piece, which represents expression beyond the nuclear boundary [Figure 6(c)]. The difference in the first pieces near the center of the nucleus is marginally statistically significant, but with very little absolute change in median slope value. The main difference between the Cav1.2 and AB70 groups lies in the middle piece (across the boundary), with Cav1.2 having significantly lower (steeper) slopes. Cav1.2 knots also appear to occur later than AB70 knots on average [Figure 6(d)].

4. Paired analysis of radial maps. In the first example [Figures 1(a) and (b)], we compared radial distributions of markers (Cav1.2 and AB70) from different cells. In the second example [Figure 7(a)], we are interested in comparing two markers (Y1F4 and RyR1) present in the same cell. From a statistical perspective, the first example is a two sample problem, whereas the second example is a paired sample one. In the second example, construct BD maps for individual cells as described in Section 2. Subsequently, we compute average expression curves for each marker as described in Section 3.1 and then scale them as described in Section 3.1.1. A simpler procedure is required for curve alignment, since the pairs of curves within each cell are automatically aligned. We construct a paired registration criterion *WREGSSEP*, where Y and R denote Y1F4 and RyR1 respectively:

$$WREGSSEP = \sum_{k=1}^{nc} \int_0^2 w(r)\{(g_k^Y(r\delta_k) - \mu^Y(r))^2 + (g_k^R(r\delta_k) - \mu^R(r))^2\}\, dr. \tag{4.1}$$

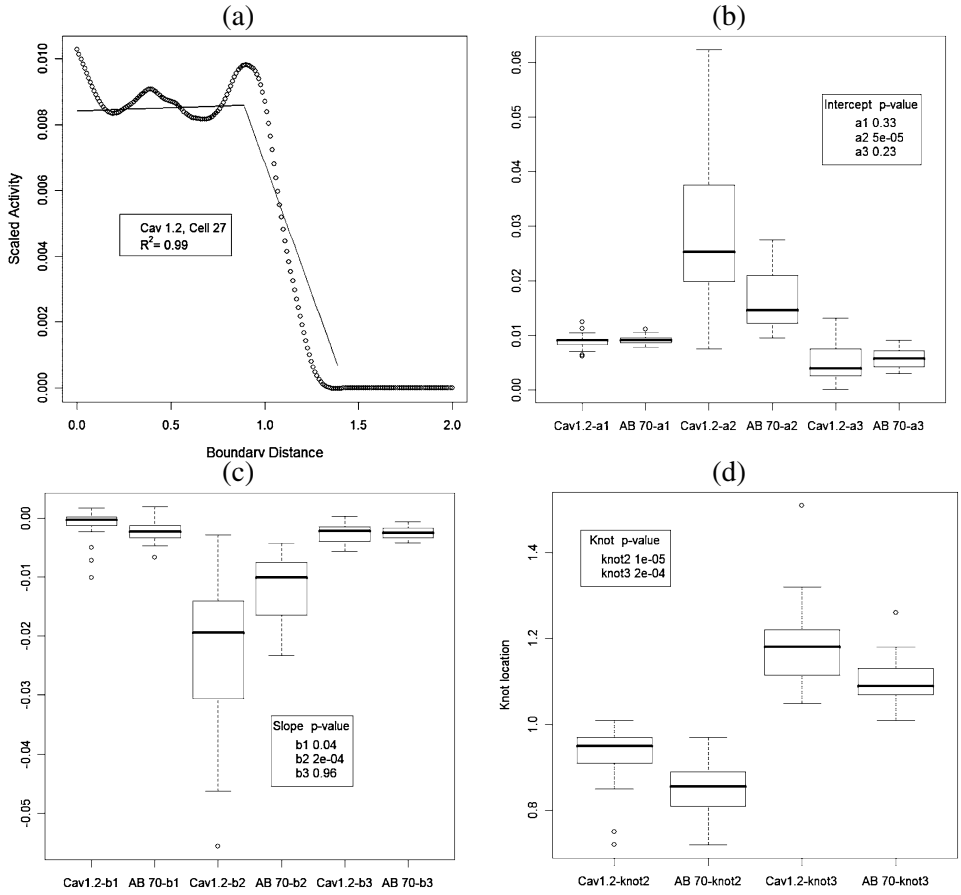

FIG. 6. (a) *Example of adaptive piecewise linear fit to a scaled Cav*1.2 *expression curve. Dots show actual curve and lines show three piece linear fit.* (b)–(d) *Subject specific piecewise linear modeling with p-values from two sample t-tests comparing parameter estimates for Cav*1.2 *and AB*70 *curves*: (b) *intercepts*, (c) *slopes*, (d) *knot locations.*

The other terms are as in (3.3). Minimization of *WREGSSP* and subsequent curve alignment were accomplished for $nc = 17$ curves (from two images) using the iterative algorithm described in Section 3.1.2.

From Figure 7(b), we can see that the RyR1 (green) curves display a coherent pattern: their intensity peaks somewhere beyond the nuclear boundary. Thereafter, their expression remains constant. For the Y1F4 (red) curves, there appear to be two subpopulations. One subpopulation peaks at the nuclear boundary, followed by a sharp decline in average intensity. The other subpopulation plateaus near the nuclear boundary, but then their average intensity increases with increasing radial distance. A paired *t*-test between the two populations [Figure 7(c)] shows significant difference in average expression between the two makers near the nuclear

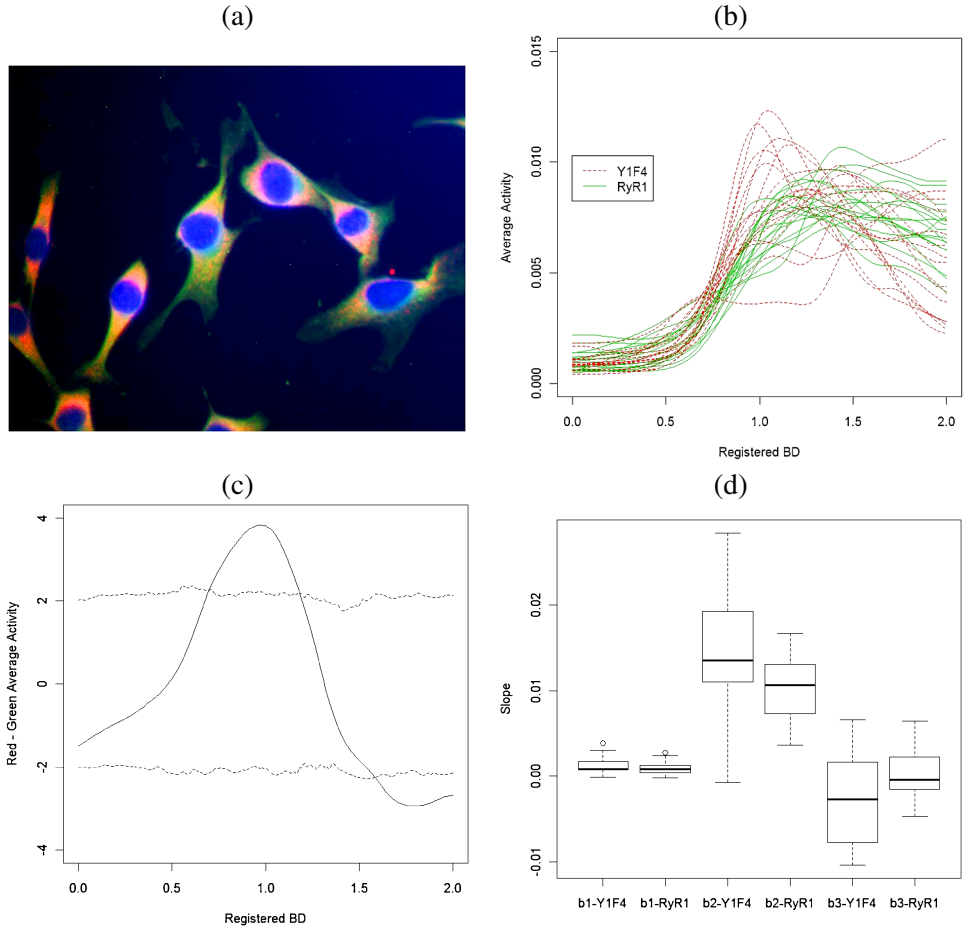

FIG. 7. (a) *Optical fluorescence microscopy image of JEG-3 trophoblastic cells. Images are labeled in a blue chromatin marker (‘DAPI’), a red marker for all forms of sarcoplasmic/endoplasmic reticulum calcium ATPase (‘Y1F4’) and a green marker [the type 1 ryanodine receptor (‘RyR1’)].* (b) *Comparison of scaled and registered average expression curves of Y1F4 and RyR1.* (c) *Solid line is curve of test statistic for paired T-test, calculated pointwise, between Y1F4 and RyR1 for expression curves shown in* (b). *Dashed lines denote 95% confidence band for test statistic, generated by restricted randomization.* (d) *Comparison of estimated slopes for Y1F4 and RyR1 from piecewise linear modeling.*

boundary and at points beyond a distance of 1.6. Confidence bands for the paired test statistic were computed using restricted randomization, that is, each pair of Y1F4 and RyR1 average expression curves was randomly reassigned to one of two groups each. The null distribution of the test statistic was then approximated using the procedure described in Section 3.2.1. The first penalized linear discriminant (not shown) is similar in shape to the paired T-statistic. A minimum CV misclassification error rate of 47% was obtained for this data set (Section 3.2.2).

Finally, the piecewise linear model (3.4) was fitted to individual expression curves. The quality of fit was typically very good (median R^2 of 0.98). Primary interest lies in the intensity gradient (slope) for the third part, which is beyond the nuclear boundary. For RyR1, we see a tight slopes distribution centerd at 0 [Figure 7(d)]. For Y1F4, we see a more dispersed slope distribution, with a preponderance of negative slopes. A paired t-test of mean slope difference shows a significant difference (p-value 0.004) between the markers.

5. Discussion. We have presented a modern statistical approach for the analysis of marker expression distributions under boundary distance mapping. The technical improvements proposed include the following: (i) Extension of the Euclidean distance map to points outside the boundary. (ii) Presmoothing and oversampling of object boundaries for improved estimation of boundary distances. (iii) Variable bandwidth smoothing of marker expression distributions. (iv) Scaling and shifting of average expression curves to account for variations in lighting and incorrect boundary identification. (v) Comparison of average expression curves across experimental conditions using suprema of t-tests and penalized discriminant analysis. (vi) Targeted inference on regionwise group differences by flexible parametric modeling. The methods are illustrated using two experiments involving calcium channels, however, the proposed techniques are general enough to be immediately applicable to other types of experiments, for example, the study of chromatin structure [Bewersdorf, Bennett and Knight (2006)] or other nuclear proteins [Knowles et al. (2006)]. In order to be applicable at larger scales, however, automated methods of image segmentation are required, for example, to identify nuclear/cellular boundaries. The success of automated techniques typically varies, depending on the quality/resolution of imaging as well as the complexity of the field of view [Jahne (2005)]. In the future we also hope to extend boundary distance analysis to more complex features, such as the local structure of marker expression.

The main findings of the analysis of the VGCC experiment are that Cav1.2 appears to have a uniform distribution throughout the nucleus which vanishes outside the nuclear boundary. Conversely, the expression of AB70 appears to gradually decrease as it reaches the periphery of the nucleus and some expression appears to persist beyond the nuclear boundary. The relatively clean separation between these two proteins (misclassification error rate of 3%) may indicate that there is a difference in transmembrane function of the channel proteins recognized by the antibodies Cav1.2 and AB70. The functional consequences of these differences will be the subject of future investigations.

In the JEG-3 cell-line experiment, differences in the distribution of RyR1 and SERCA (Y1F4) are not that clear (misclassification error rate of 47%). This is not unexpected, since both proteins would be expected to be located in the ER of these trophoblasts. Our finding of heterogeneity in the Y1F4 average expression curves suggests that only certain subdomains of the ER within JEG-3 cells could be specialized for SERCA-mediated Ca^{2+} uptake. This possibility is not without

precedent, since the SR of striated muscle is functionally and anatomically divided into subdomains specialized for either Ca^{2+} uptake or for Ca^{2+} release [Mackrill (1999)]. In early video microscopy studies using Ca^{2+}-sensitive fluorophores it was noted that sister cells displayed distinct Ca^{2+} responses to hormonal stimulation [Ambler et al. (1988)]. Epigenetic variations in the abundance (intensity) of Ca^{2+}-signalling components between individual cells in a population could give rise to such differences.

We have proposed a modification of the Euclidean boundary distances [Knowles et al. (2006)] to measure boundary distance for points outside the object boundary. A similar extension for erosion-based distance measurement [Bewersdorf, Bennett and Knight (2006)], using dilation instead of erosion, is straightforward [Bewersdorf, Bennett and Knight (2006)]. Similarly, the methodology described here can extend to 3-d stacks of images in a straightforward manner. However, we note that the methodology described here can be satisfactorily applied only in situations where the orientation of expression/objects is not of interest, since all orientation information is lost in the profile distributions.

The attraction of the FDA approach lies in the fact that it extends standard univariate statistical techniques like ANOVA and *t*-tests to curve data [Ramsay and Silverman (2002)]. However, the necessity of preprocessing curves by registration can mean that some information about differences between groups can be lost. The adaptive piecewise linear approach proposed in Section 3.3 avoids this loss of information. Significant differences in the distribution of knot points across groups indicate that this may indeed be the case. Moreover, piecewise linear modeling also reveals that the difference in the average expression curves may not be in their magnitude, but in their slopes.

Acknowledgments. The first author would like to thank Trevor Hastie and Rob Tibshirani for helpful discussions.

REFERENCES

AMBLER, S. K., POENIE, M., TSIEN, R. Y. and TAYLOR, P. (1988). Agonist-stimulated oscillations and cycling of intracellular free calcium in individual cultured muscle cells. *J. Biol. Chem.* **263** 1952–1959.

ANDERSON, T. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

BEWERSDORF, J., BENNETT, B. and KNIGHT, K. (2006). H2AX chromatin structures and their response to DNA damage revealed by 4Pi microscopy. *Proc. Natl. Acad. Sci. USA* **103** 18137–18142.

CALLINAN, L., MCCARTHY, T., MAULET, Y. and MACKRILL, J. (2005). Atypical L-type channels are down-regulated in hypoxia. *Biochemical Society Transactions* **33** 1137–1139.

FABBRI, R., COSTA, L., TORELLI, J. and BRUNO, O. (2008). 2D Euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys* **40** 2:1–2:44.

FERNANDEZ-GONZALEZ, R., MUNOZ-BARRUTIA, A., BARCELLOS-HOFF, M. and ORTIZ-DE-SOLORZANO, C. (2006). Quantitative in vivo microscopy: The return from the 'omics.' *Current Opinion in Biotechnology* **17** 501–510.

FRIEDMAN, J. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* **84** 165–175. MR0999675

GOMEZ-OSPINA, N., BARRETO-CHANG, O., HU, L. and DOLMETSCH, R. (2006). The C terminus of the L-type voltage-gated calcium channel Ca(V)1.2 encodes a transcription factor. *Cell* **127** 591–606.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer, New York. MR1851606

JAHNE, B. (2005). *Digital Image Processing*, 6th ed. Springer.

KNOWLES, D., SUDAR, D., BATOR-KELLY, C., BISSELL, M. and LELIÈVRE, S. (2006). Automated local bright feature image analysis of nuclear protein distribution identifies changes in tissue phenotype. *Proc. Natl. Acad. Sci. USA* **103** 4445–4450.

MACKRILL, J. J. (1999). Protein–protein interactions in intracellular Ca2+-release channel function. *Biochem. J.* **337** 345–361.

RAMSAY, J. and SILVERMAN, B. (2002). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

SILVERMAN, B. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898–916. MR0751281

SILVERMAN, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Roy. Stat. Soc. Ser. B* **47** 1–52. MR0805063

WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–29. MR0362682

K. ROY CHOUDHURY
DEPARTMENT OF STATISTICS
UNIVERSITY COLLEGE CORK
IRELAND
E-MAIL: kingshuk@ucc.ie

L. ZHENG
J. J. MACKRILL
DEPARTMENT OF PHYSIOLOGY
UNIVERSITY COLLEGE CORK
IRELAND