Supporting Information for High-Resolution Raman Microscopic Detection of Follicular Thyroid Cancer with Unsupervised Machine Learning

J. Nicholas Taylor[1], Kentaro Mochizuki[2], Kosuke Hashimoto[3], Yasuaki Kumamoto[3], Yoshinori Harada[3], Katsumasa Fujita[2,4,5], Tamiki Komatsuzaki[1,6,7]

[1] Research Institute for Electronic Science, Hokkaido University, Kita 20, Nishi 10, Kita-ku, Sapporo 001-0020, Japan.

[2] Department of Applied Physics, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

[3] Department of Pathology and Cell Regulation, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kajii-cho, Kawaramachi-Hirokoji, Kyoto, 602-8566, Japan

[4] Advanced Photonics and Biosensing Open Innovation Laboratory, AIST-Osaka University, Yamadaoka, Suita, Osaka 565-0871, Japan

[5] Transdimensional Life Imaging Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Yamadaoka, Suita, Osaka 565-0871, Japan

[6] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University Kita 21 Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan

[7] Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS-Univ. Bourgogne Franche-Comt, 9 Av. A. Savary, BP 47 870, F-21078, Dijon Cedex, France

1 Supplementary Methods

Spectral Preprocessing. All computations were performed using custom software written in MATLAB version R2017b (Mathworks, Natick, MA, U.S.A.). Raman images were first preprocessed to remove electronic bias and cosmic rays. Electronic bias was removed by subtracting the value of 520 counts as provided by the manufacturer. Cosmic rays were treated with a recursive outlier detection scheme in which intensities larger than the mean intensity of the image at a particular wavenumber plus 8 standard deviations were replaced with the mean of the intensities in the cube surrounding the outlying pixel. Singular-value decomposition (SVD) was then performed to reduce detection noise. It was found that $\sim$95% of the signal information was contained within the 16 most significant singular values; SVD was reversed with these 16 singular values being retained. Baseline correction was performed with a recursive polynomial fitting algorithm [1]. A 7^{th} degree polynomial was used to fit each spectrum over the entire spectral range (200-3130 cm^{-1}), and the recursion continued until less than 5% of the spectrum was negative after baseline subtraction. Intensities in the silent region (1800-2800 cm^{-1}), and at extreme wavenumbers, i.e., those less than 500 cm^{-1} or greater than 3030 cm^{-1}, were set to zero before area normalization over the remaining wavenumbers (500-1800 cm^{-1}, 2800-3030 cm^{-1}).

Superpixels. Prior to normalization, for the purposes of further increasing signal-to-noise ratio and decreasing the number of spectra used in additional computations, adjacent pixels were averaged over spatially-local regions within the images, producing spectra representing groups of pixels, or superpixels. Rather than using a simple binning scheme on a rectangular grid of predetermined size and location, we instead choose simple linear iterative clustering (SLIC) [2] to better preserve the spatial characteristics of the cells in the Raman images. SLIC is a pixel clustering method based on spatial proximity and similar characteristics. Given the total number of pixels of the image N, the i-th pixel, denoted by γ_i,

is represented by $[I_i(1), I_i(2), ..., I_i(W), x_i, y_i]$, where $I_i(w)$, x_i, y_i denote the Raman intensity prior to normalization at wavenumber w and the x and y coordinates of the i-th pixel. So that superpixels contain groups of proximate individual pixels having similar underlying intensity, the mean of the intensities over the wavenumber dimension, denoted by $\hat{I}_i = \frac{1}{W}\sum_w I_i(w)$, and the coordinates x_i and y_i were input to the algorithm. Each superpixel, Γ_α, in the set of all superpixels $\mathbf{\Gamma} = \{\Gamma_1, ..., \Gamma_{N_{\text{sp}}}\}$, is then defined as the set of pixels minimizing a distance to that considers both the average Raman intensities and positions.

$$d(\gamma_i, \Gamma_\alpha) = \sqrt{\left(\frac{d_s(\gamma_i, \Gamma_\alpha)}{\max_{\Gamma_\alpha \in \mathbf{\Gamma}}\left[\max_{\gamma_i \in \Gamma_\alpha} d_s(\gamma_i, \Gamma_\alpha)\right]}\right)^2 + \left(\frac{d_c(\gamma_i, \Gamma_\alpha)}{\max_{\Gamma_\alpha \in \mathbf{\Gamma}}\left[\max_{\gamma_i \in \Gamma_\alpha} d_c(\gamma_i, \Gamma_\alpha)\right]}\right)^2} \tag{S1}$$

Here, the spatial distance $d_s(\gamma_i, \Gamma_\alpha)$ is the Euclidean distance between the position (x_i, y_i) of pixel γ_i and the center position (x_α, y_α) of superpixel Γ_α,

$$d_s(\gamma_i, \Gamma_\alpha) = \sqrt{(x_i - x_\alpha)^2 + (y_i - y_\alpha)^2}. \tag{S2}$$

The center position of superpixel Γ_α, (x_α, y_α), is obtained as the mean location of the set containing the N_α pixels assigned to Γ_α,

$$\begin{aligned} x_\alpha &= \frac{1}{N_\alpha}\sum_{\gamma_i \in \Gamma_\alpha} x_i. \\ y_\alpha &= \frac{1}{N_\alpha}\sum_{\gamma_i \in \Gamma_\alpha} y_i. \end{aligned} \tag{S3}$$

Similarly, the intensity distance $d_c(\gamma_i, \Gamma_\alpha)$ is the Euclidean distance between the average intensity $\hat{I}_i$ at pixel γ_i and the average intensity $\hat{I}_\alpha = \frac{1}{N_\alpha}\sum_{\gamma_i \in \Gamma_\alpha} \hat{I}_i)$ of superpixel Γ_α,

$$d_c(\gamma_i, \Gamma_\alpha) = \sqrt{\left(\hat{I}_i - \hat{I}_\alpha\right)^2}. \tag{S4}$$

The number of superpixels per image, N_{sp}, was chosen such that the superpixels contain an average of 16 individual pixels. For the 2-dimensional, average Raman intensity image containing N pixels, the SLIC algorithm was initialized by placing the N_{sp} superpixel centers on a square grid with spacing $\sqrt{N/N_{\text{sp}}}$. Each pixel was then assigned to the superpixel center nearest its spatial location. Each subsequent iteration assigned pixels to superpixels based on the minimum distance (Eq. S2) over superpixels having centers that are within the square spacing interval $2\sqrt{N/N_{\text{sp}}}$. The set of pixels $\{\gamma_i\}$ belonging to each superpixel Γ_α are treated as invariant entities defined by the previous iteration until the present iteration is completed. The average superpixel positions typically converged within a 1% tolerance in 10 iterations or fewer, so the algorithm was iterated 10 times or until the superpixel locations had converged, whichever happened first. Isolated single pixels $\{\gamma_i\}$ are merged to the closest superpixels $\{\Gamma_\alpha\}$ after convergence. After the individual pixels have been assigned to superpixels, the average spectrum over the locations of each superpixel are used for further analyses.

Cell Segmentation. Spectra in an image were first partitioned into 10 groups with the k-means clustering algorithm using the L_1 distance (Eq. 3) in the high wavenumber spectral region (2800-3030 cm^{-1}). The spectra belonging to the 6 most intense clusters were retained as pixels containing cellular regions, as determined by visual comparison to the cell locations in the Raman images. After the regions associated with cells were identified, areas associated with particular cells were localized through graph theoretical methods [3]. A binary, 2-D image, containing ones at pixels associated with cells and zeros at pixels associated with non-cell regions, was constructed and treated as a connected graph. From the binary image, connected regions of ones, containing cells, were identified with a flood-fill algorithm. Isolated connected regions containing less than 100 pixels ($< 10\mu\text{m}^2$) are too small to be cells and were discarded. So that

only complete, or nearly complete, cells were included in the analyses, connected regions having significant external distance ($\geq$ 50% of the total external distance) along the boundaries of the images were also excluded. Intensity maps of remaining regions were then examined in wavenumber regions associated with DNA (780-800 cm^{-1}, 1090-1100 cm^{-1}, 1370-1380 cm^{-1}), so as to identify regions containing 2 or more obvious cell nuclei. In such regions, additional barriers were created along local minima of the intensity gradient between the two nuclei. Remaining connections between nuclei were removed manually. 49 connected regions remained, and were taken to represent cells in the Raman images.

Rate-Distortion Theory. As mentioned in the main text, the RDT algorithm solves directly for the conditional probabilities $p(c_k|s_i)$, producing a classification for each spectrum s_i.

$$p(c_k|s_i) = \frac{p(c_k)}{z(s_i;\beta)} e^{-\beta\, D(c_k\,;s_i)} \tag{S5}$$

Here, Eq. 3 in the main text is evaluated for a particular class and spectrum, $D(c_k\,;s_i) = \sum_{j=1}^{N} p(s_i|c_k)\, d_{ij}$, producing a mean distance of the spectrum s_i from the class c_k, and $z(s_i\,;\beta)$ is a normalization function. The joint probability $p(c_k, s_i) = p(c_k|s_i)\, p(s_i)$ is marginalized to obtain the marginal probability of the class c_k.

$$p(c_k) = \sum_{i=1}^{N} p(c_k|s_i)\, p(s_i) \tag{S6}$$

The algorithm [4, 5] is initialized through random assignment of $p(c_k|s_i)$ for all spectra and classes. After normalization, Eqs. S5 and S6 are alternatively computed until self-consistency is reached and Eq. 1 converges within a specified tolerance. A class γ_i is then assigned to each spectrum s_i by taking the maximum probability class given the spectrum, i.e., $\gamma_i = \mathrm{argmax}_k\, p(c_k|s_i)$. The parameter β from Eq. 1 in the main text was chosen based on the maximum pairwise L_1 distance over all spectra, i.e., $\beta = 100 \times \max_{i,j} d_{ij}$, which was sufficiently large to produce hard clustering assignments, i.e., $\frac{1}{N}\sum_{i=1}^{N} \max_k p(c_k|s_i) > 0.99$.

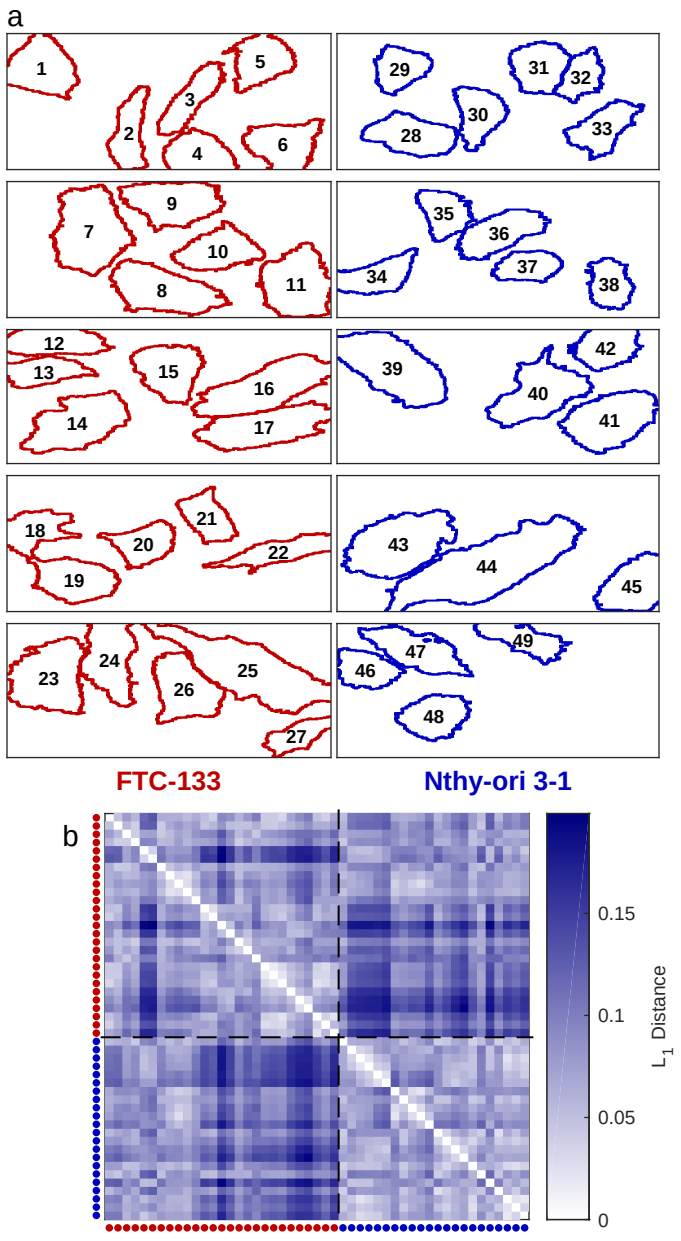

Figure S1: Cell labels and distances matrix among averaged cellular spectra. a) The pixels associated with each cell are outlined with red for FTC-133 cells and blue for Nthy-ori 3-1 cells. An associated index is placed in the center of each cell outline. b) The pairwise distance matrix among averaged cellular spectra. Red markers along the vertical and horizontal axes indicate FTC-133 cells and Nthy-ori 3-1 cells, respectively.

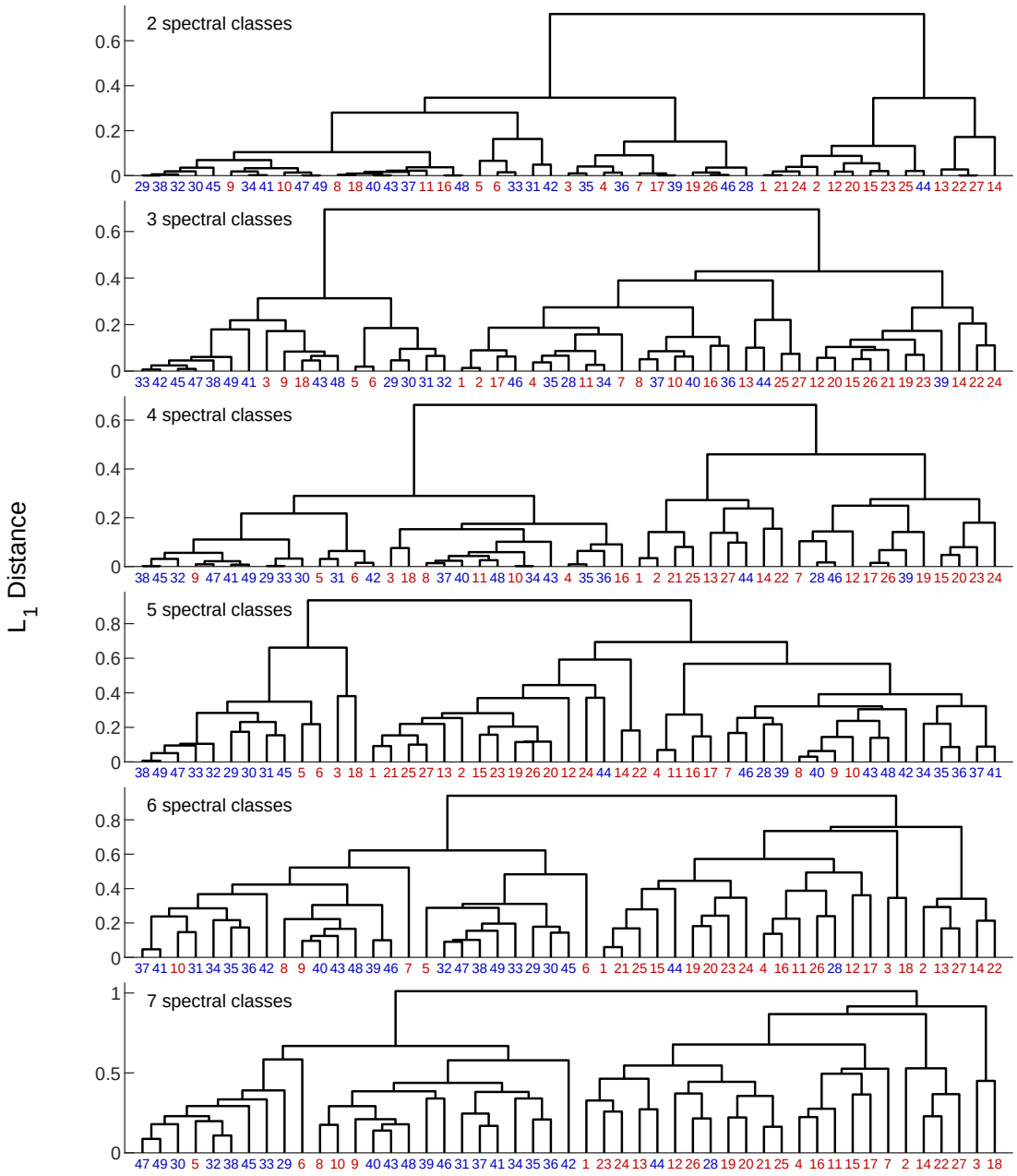

Figure S2: Dendrograms for class density classification with 2-7 spectral classes.

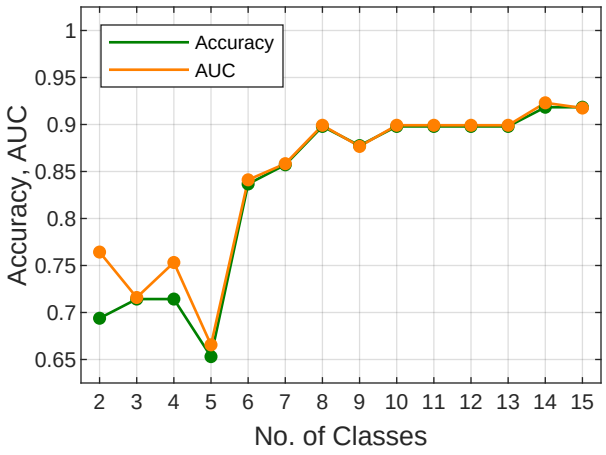

Figure S3: Accuracies and L_1 distances between FTC-133 and Nthy-ori 3-1 class densities for 2 to 15 classes.

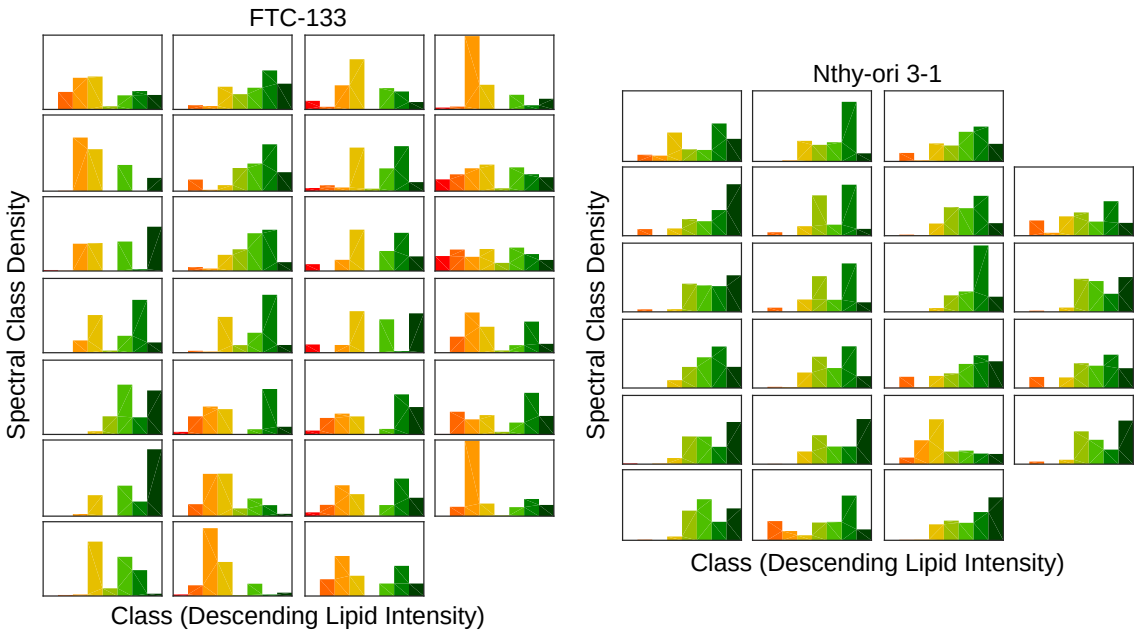

Figure S4: Class densities with 8 spectral classes for all 49 cells.

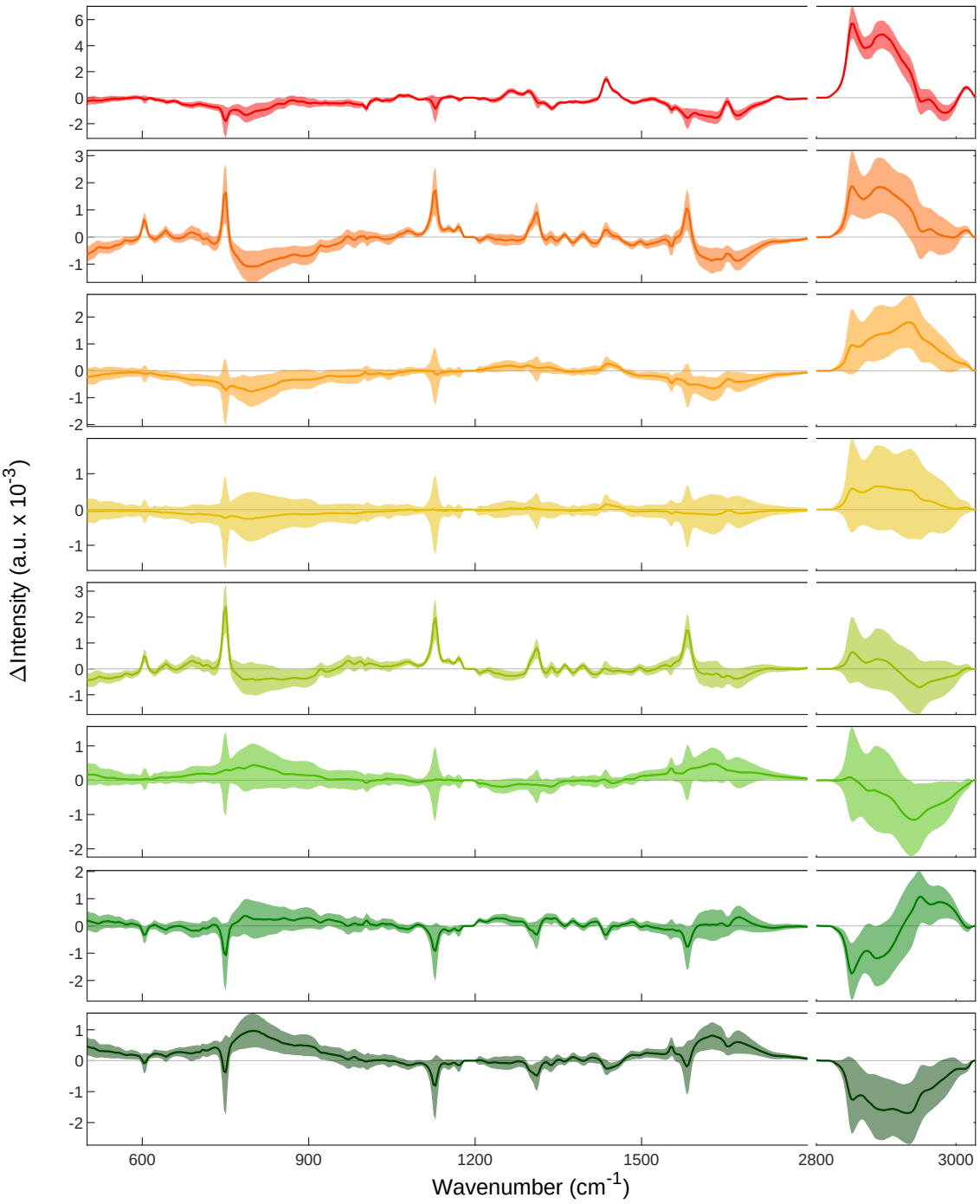

Figure S5: Mean distance spectra. Colors associated with each spectrum are the same as those used in Fig. 4 in the main text. Axes have been rescaled from Fig. 6 so that details can be more easily observed. Shaded regions indicate 95% confidence limits across all pairwise difference spectra.

DNA-Specific Intensity Maps and Class Maps

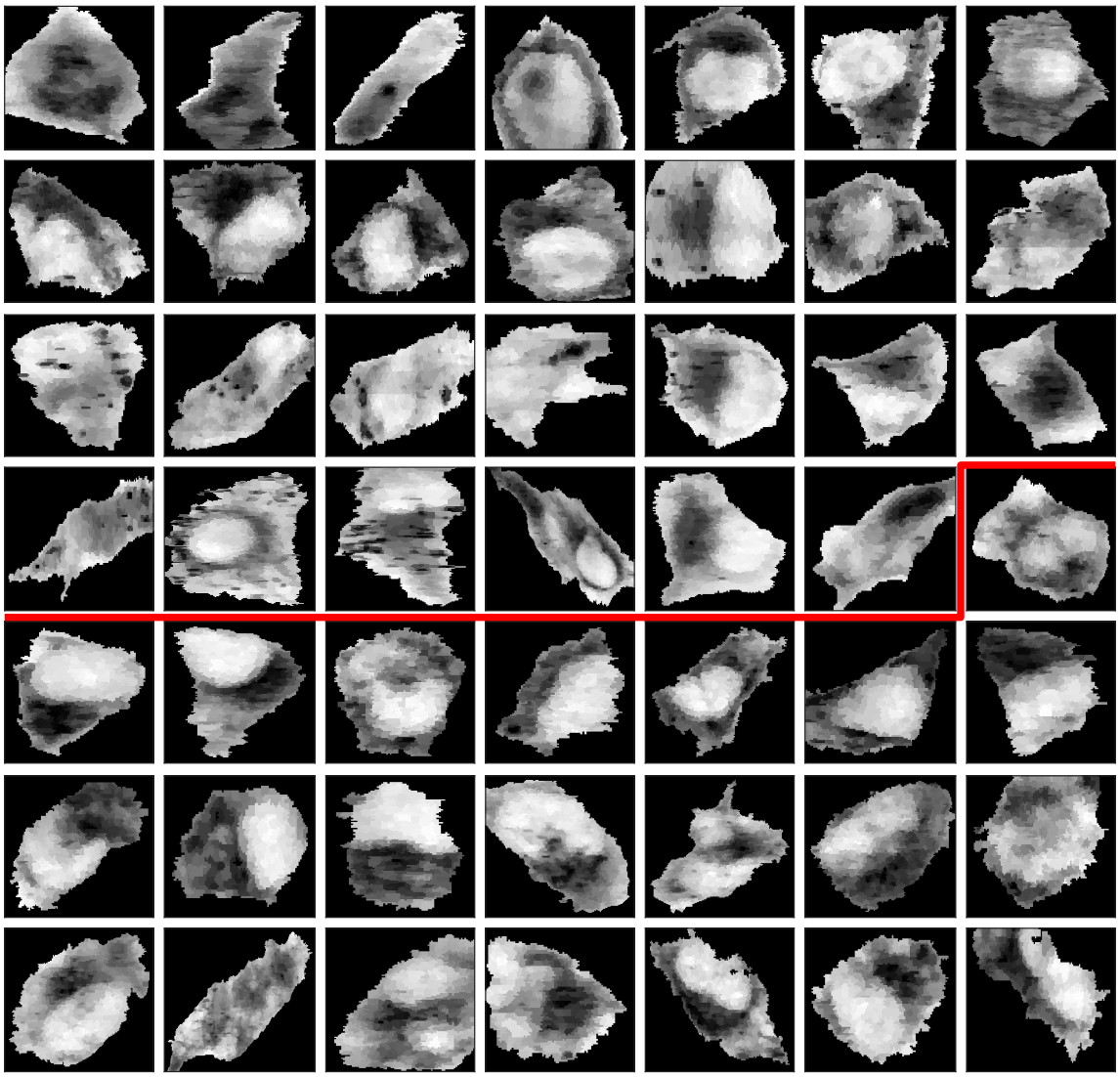

Figure S6: Intensity maps of each cell have been summed over wavenumbers ranges of 780-800 cm^{-1}, 1090-1100 cm^{-1}, 1240-1260 cm^{-1}, 1370-1380 cm^{-1} and 1655-1675 cm^{-1} so that the cell nuclei can be more easily identified. Cells above the thick red dividing line are FTC-133, while those below it are Nthy-ori 3-1.

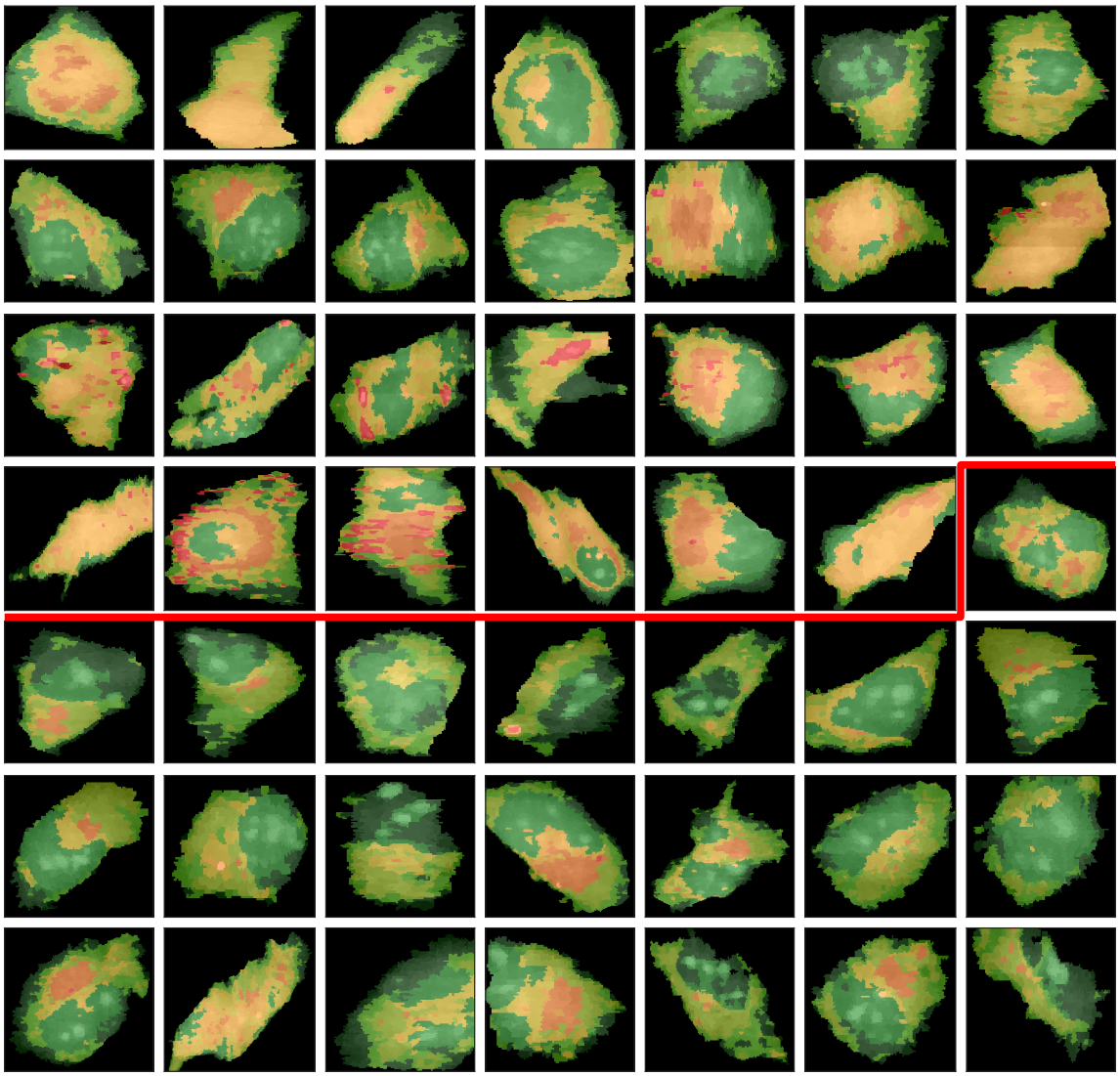

Figure S7: High wavenumber region intensity maps overlaid with class maps for all cells. Cells above the thick red dividing line are FTC-133, while those below it are Nthy-ori 3-1.

References

(1) Lieber, C. A.; Mahadevan-Jansen, A. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. *Appl. Spectrosc.* **2003**, *57*, 1363–1367.

(2) Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Susstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Patt. Anal. Mach. Intell.* **2012**, *34*, 2274–2282.

(3) Gross, J. L.; Yellen, J., *Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications)*; Chapman & Hall/CRC: 2005.

(4) Blahut, R. Computation of Channel Capacity and Rate-Distortion Functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473.

(5) Arimoto, S. An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 14–20.