10-6-2015

Segment and fit thresholding: a new method for image analysis applied to microarray and immunofluorescence data.

Elliot Ensink
Van Andel Research Institute

Jessica Sinha
Van Andel Research Institute

Arkadeep Sinha
Van Andel Research Institute

Huiyuan Tang
Van Andel Research Institute

Heather M. Calderone
Van Andel Research Institute

Authors

Elliot Ensink, Jessica Sinha, Arkadeep Sinha, Huiyuan Tang, Heather M. Calderone, Galen Hostetter, Jordan M. Winter, David Cherba, Randall E. Brand, Peter J. Allen, Lorenzo F. Sempere, and Brian B. Haab

HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2016 May 03.

Published in final edited form as:

Anal Chem. 2015 October 6; 87(19): 9715–9721. doi:10.1021/acs.analchem.5b03159.

Segment and Fit Thresholding: A New Method for Image Analysis Applied to Microarray and Immunofluorescence Data

Elliot Ensink[1], **Jessica Sinha**[1], **Arkadeep Sinha**[1], **Huiyuan Tang**[1], **Heather M. Calderone**[1], **Galen Hostetter**[1], **Jordan Winter**[2], **David Cherba**[1], **Randall E. Brand**[3], **Peter J. Allen**[4], **Lorenzo F. Sempere**[1], and **Brian B. Haab**[1,*]

[1]Van Andel Research Institute,333 Bostwick Ave NE, Grand Rapids, MI 49503

[2]Thomas Jefferson University, 1025 Walnut St Philadelphia, PA 19107

[3]University of Pittsburgh Medical Center, 200 Lothrop St, Pittsburgh, PA 15213

[4]Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065

Abstract

Certain experiments involve the high-throughput quantification of image data, thus requiring algorithms for automation. A challenge in the development of such algorithms is to properly interpret signals over a broad range of image characteristics, without the need for manual adjustment of parameters. Here we present a new approach for locating signals in image data, called Segment and Fit Thresholding (SFT). The method assesses statistical characteristics of small segments of the image and determines the best-fit trends between the statistics. Based on the relationships, SFT identifies segments belonging to background regions; analyzes the background to determine optimal thresholds; and analyzes all segments to identify signal pixels. We optimized the initial settings for locating background and signal in antibody microarray and immunofluorescence data and found that SFT performed well over multiple, diverse image characteristics without readjustment of settings. When used for the automated analysis of multi-color, tissue-microarray images, SFT correctly found the overlap of markers with known subcellular localization, and it performed better than a fixed threshold and Otsu's method for selected images. SFT promises to advance the goal of full automation in image analysis.

Introduction

Many types of scientific experiments use images to collect data. In order to derive information from the image data, it must be interpreted to produce quantitative or semi-quantitative information. If the user simply needs semi-quantitative evaluation from a small number of datasets, the user could visually inspect and interpret each image. Or if the analysis involves the recognition of highly complex features or patterns, as in the inspection of tissue by a medical pathologist to render a diagnosis, manual interpretation may be required. But if the user requires precise and objective quantification, or analysis of signals that are difficult to locate by eye, or the analysis of many data sets, automated interpretation

*Correspondence to: Brian B. Haab, PhD, Van Andel Research Institute, 333 Bostwick NE, Grand Rapids, MI 49503, brian.haab@vai.org, 616-234-5268.

would be preferable.[1–2] With the ever-improving quality, content, and volume of image data, the demands upon the software tools for image analysis are increasing.[1]

Among the many applications of automated image analysis, an important area is medical practice and research. In clinical practice, where results from images could be used to inform treatment decisions, a significant goal is to remove the subjectivity and inter-operator variability that sometimes influence results. Scientists are developing new tools for the analysis of images from X-rays,[3] MRI, PET, ultrasound, CT, cytology,[4–5] and immunohistochemistry, [2, 6–8] among others. In biomedical research, automated image analysis is important for high-throughput methods such as tissue microarrays,[9] blood cell analysis,[5] high-content screening of cellular features or behavior,[10–11] cell-based drug screening,[11–12] or imaging of animal models such a *C. elegans*.[7, 13–14] Many such studies would not be possible without some level of automation in the image analysis.

The development of robust algorithms for image analysis continues to be a challenge. A common difficulty in automating the analysis of images is to account for the diverse and unpredictable nature of image data; a broad range of signal levels, amounts, and morphologies is common within any given data type.[15] Most algorithms perform well when the image has predictable characteristics or conforms to certain assumptions, but not well if the image has other qualities. A widespread strategy is to use histograms of pixel intensities to model the signal and background distributions and to find thresholds.[2, 16–18] The use of histograms requires sufficient representation of signal and background to properly find the distributions, and it can have difficulty handling images with noise spikes in the background. Other strategies rely on edge detection to locate signal regions, typically by finding steep intensity gradients or high spatial frequencies.[19–21] Such approaches may not be reliable where steep edges are not present in the signal regions, or where shapes are irregular. In some images, portions of true signals have sharp edges and others do not, making a single threshold in gradient or spatial frequency inaccurate in certain places. The Watershed Transformation looks for contiguous regions that are higher than surrounding regions, thus distinguishing cohesive “hills” from neighboring “valleys.”[22] Several variants on this approach have appeared that function well in particular applications such as the identification of atypical cells in cytological images.[5] But similar to the above methods, the optimal threshold may be significantly different between images. Furthermore, methods using a single threshold to distinguish signal from background have the problem of allowing spikes—random, sharp elevations from noise—to be counted as signal. Filtering can reduce spikes but also blur or alter true signals. Alternatively, one could model the shapes and sizes of true signals,[23] and the user can train a system to recognize specific features,[14] but such methods function well only where true signals are predictable.[24] In practice, one could combine manual and quantitative interpretation, for example by having the software perform a primary analysis and having the user adjust settings for unusual cases,[7] but a manual review of the data limits objectivity and throughput.

We explored a new approach to image analysis, called Segment and Fit Thresholding (SFT). SFT is based on statistical characteristics of signal and background signal that hold true in segmented portions of an image. The method defines thresholds for identifying background and signal by fitting the relationships among the segmented data. By eliminating the need

for assumptions about the amount, location, intensities, or sizes of the signals, SFT can properly locate signals in images with highly diverse characteristics, without user intervention. Here we present the development and optimization of the method and the testing of its performance for analyzing microarray and multi-color immunofluorescence data.

Methods

Antibody Microarray Data Collection and Analysis

EDTA-plasma samples were collected at the University of Pittsburgh School of Medicine from patients scheduled to undergo examination by endoscopic ultrasound. All samples were collected under approved human-subjects protocols. We analyzed the plasma samples on antibody arrays using methods previously described.[25–28] The Supplementary Information provides details of the experiments used here.

Immunofluorescence Experiments

The tissue microarrays were produced using resected pancreatic tumors collected at the Memorial Sloan Kettering Cancer Center, as described earlier.[29] In addition, the VARI Biospecimen facility provided formalin-fixed, paraffin-embedded tissue from patients who underwent pancreatic resections at a regional hospital affiliate in Grand Rapids, MI. All samples were collected under approved human-subjects protocols. We performed the immunofluorescence experiments according to standard protocols[30] (see Supplementary Information).

For analysis using a fixed threshold and Otsu's method, we used Inform 2.1.1 from PerkinElmer. The colocalization and single-channel analyses by Otsu's method were performed individually on each image.

Software Development and Data Preparation

We developed and tested the software to implement SFT using MATLAB version R2014a, supplemented with the image processing and curve fitting toolboxes. We used Microsoft Excel for analyzing numerical output, GraphPad Prism for the preparation of graphs, and Canvas XIV for the preparation of figures.

Results

Segment and Fit Thresholding

The core feature of the SFT method is dividing the image into component segments and calculating statistics for each segment (Fig. 1). The program calculates various statistics among the pixels in each segment, such as the mean and coefficient of variation (CV). The program plots the values for particular statistics over all segments and fits a quadratic equation to the relationship. We use a quadratic fit because the relationships are expected to be relatively regular, rather than showing up-and-down behavior as allowed by higher-order fits.

Analyzing an Image to Find Signals

SFT makes use of the fact that background regions generally have different statistical features from signal regions. For example, signal regions generally have higher intensities and higher CVs than the background regions. By fitting relationships over multiple segments, we remove the noise associated with any individual segment and get a better estimate of the true relationship between statistics, such as between signal and CV.

The first step in analyzing an image is to locate the background regions (Fig. 2A). The user can empirically determine for a given data type the maximum CV that is typical for the background regions (we give more details below on this choice). The program then finds the mean that best corresponds to the CV threshold, based on the quadratic fit of the relationships among the segments. The resulting mean is used as a threshold for finding background regions. The program marks the segments with a mean less than the threshold, and for each pixel, it counts the number of times it is included in a marked segment. (Using a 3x3 segment, each pixel would be included in 9 segments, except for pixels near the edge of the image.) The pixels for which greater than a certain percentage of the segments (we used 50% for the microarray data) are marked are defined as background pixels.

The next step is to analyze the background to determine a threshold for finding signal (Fig. 2B). Any segment for which greater than 50% of its pixels are background pixels is counted as a background segment. For all background segments, the standard deviations and medians are calculated, plotted, and fitted. The program then calculates the median of all background segments and plugs the resulting median into the quadratic fit to find the corresponding standard deviation.

The next step is to locate the signals in the image (Fig. 2C). In the example shown here, the threshold for each segment is the background median plus three times the background standard deviation. The program segments the image and marks the segments that have a median above the threshold. To be classified as signal, a pixel must be included in a minimum percentage of segments with medians exceeding the segment threshold. In addition, a pixel must have an intensity that exceeds a pixel threshold. The threshold for the intensity of the pixel is based on the median and standard deviation of all background pixels identified previously.

In the application of SFT to microarray and immunofluorescence data, we set the parameter values empirically based on detailed examinations of results (Fig. S1). We found that the optimal settings differed between the data types (Table S1), but that within each data type, the optimized settings performed well over all images, as described below. Also, the final results were not highly sensitive to changes in the settings of a parameter; changing the background CV threshold from 0.05 to 0.4 in immunofluorescence data resulted in minor differences in pixels selected (Fig. S2).

Application to Antibody Microarray Data

We tested the method on data from the analysis of plasma glycoproteins using antibody-lectin sandwich arrays, as presented earlier[25–27]. The implementation of SFT for microarray

analysis involved the additional capabilities of finding the locations of the spots and matching the identities of the antibodies to the spots.[31] (See Supplementary Information.)

SFT gave nearly identical values to a manual analysis (using GenePix 5.0) over all spots for an array with good spot intensities and morphologies (Fig. 3A). For an array with generally weak signals, the correlation between methods was lower because of noise in the signals, but the methods arrived at a similar value for the bright, positive-control spot (Fig. 3B). Using an array with some defective spots, the manual method with semiautomated spot finding arrived at higher values than the SFT method (Fig. 3C). An examination of the spots showed the reason: the manual method using spot-finding features zoomed in on the highest portion of the disjointed spots and ignored the remainder of the true signal, thus giving an abnormally high value by not averaging over all signal pixels. The SFT method, in contrast, found all true signal pixels, whether in an orderly pattern or not, and thus gave lower average values.

Application to Immunofluorescence Data

To implement the method for immunofluorescence (IF), we had to account for regions with no tissue. Regions with no tissue are neither background nor signal; they simply do not contain data. We used the concept that regions with no tissue should generate very little fluorescence and thus have lower CVs than the regions with tissue. Using a maximum CV threshold of 0.1, as determined by empirical testing (see Table S1 for the other settings for this step), we found that the method correctly identified non-tissue regions, including both the corners of the image and the hollow regions of the core (Fig. 4A). The identification of non-tissue areas using the same parameter settings was robust over all images tested (Fig. S3). After eliminating the non-tissue regions, the method identified with good accuracy both the background and signal regions in each channel of 3-color immunofluorescence data (Fig. 4A).

An important use of IF experiments is to detect colocalization of fluorescent signals from distinct probes.[32] The program scans the analyzed results from each color, and if a minimum percentage of the pixels within a segment in each channel are signal pixels, then all the signal pixels of either color within the segment are counted as colocalized (Fig. 4B). The parameters of the size of the segment and the percentage of pixels with signal could be adjusted according to the needs of the experiment; in the examples of Figure 5, we used a 15 x 15 pixel segment, which was the average size of a cell in the tissue images, and required 25% of the pixels in the segment for all colors to be signal pixels.

We applied the method to images for which the subcellular localization and expected colocalization of individual signals was known. The DAPI dye stains DNA and thus localizes to the nuclei; the small nuclear U6 RNA is specific to nuclei and thus should colocalize with DAPI; and the microRNA miR-21 is cytoplasmic and should not colocalize much with DAPI (some overlap would occur due to the 3D nature of cells). An automated analysis of 12 cores from a tissue microarray (TMA) showed higher levels of U6/nuclear colocalization than miR-21/nuclear colocalization (Fig. 4C). In contrast, cores from the same TMA stained for two cytoplasmic proteins—vimentin and CK19—showed no differences between the proteins in colocalization with DAPI (Fig. 4C).

We further tested the automation of SFT by analyzing multiple images from a TMA stained for 2 different proteins, MRC2 and epcam. We processed 14 images using the pre-defined settings (given in Table S2) without user intervention. The images had huge differences between them in the signals from the three colors, yet in each case the relative amounts of signal and colocalization detected by SFT matched a visual inspection of the images (Fig. S4).

Comparison to Fixed Thresholds and Otsu's Method

We compared two other methods to SFT: a fixed threshold that was optimized by visual inspection of several images, and the widely-used Otsu method, implemented in a commercial package. Otsu's method fits theoretical signal and background distributions to a histogram of pixel intensities and chooses a threshold to minimize overlap between the distributions.[33] Using images described above (Fig. 4C), a comparison of miR-21/DAPI to U6/DAPI colocalization confirmed the expected higher levels of U6/DAPI colocalization when analyzed by SFT but showed a few cores with divergent behavior when analyzed by the fixed threshold or Otsu's method (Fig. 5A). Three cores (cores 3, 4 and 5) had elevated miR-21/DAPI by the fixed threshold, and two cores (cores 1 and 5) were higher by Otsu's method.

We speculated that the elevated miR-21/DAPI colocalization was due to low thresholds in the miR-21 channel, given that DAPI staining is typically regular and distinct. We thus examined the thresholds of each method relative to the histograms of pixel intensities in the miR-21 channel (Fig. 5B). For a core that had similar miR-21/DAPI values between the methods (core 10), all the thresholds were in the broad tail of the histogram, which is the region of expected signal, and the maps of identified pixels were similar. The three cores with elevated miR-21/DAPI by the fixed threshold had histograms shifted to higher intensities, resulting in a threshold that selected pixels from nearly the entire cellular regions of cores 4 and 5. Core 11 had a low-intensity distribution, resulting in a fixed threshold that was off the tail of the histogram and that selected few pixels. Otsu's method performed well for most cores but selected thresholds in the background distributions for cores 1 and 5, resulting in widespread signal detection.

In contrast, the SFT box and pixel thresholds were in the tail of the histogram in each case, and the maps of detected pixels showed identifications of the regions that were clearly brighter in the raw image. The advantage of using both a box and pixel threshold appeared in cases where the thresholds were similar between the methods, such as cores 3 and 10. By rejecting noise spikes through the use of a box threshold, SFT produced cleaner selections of the bright regions than Otsu's method or the fixed threshold. A detailed view of the selected pixels at precisely-optimized, fixed thresholds confirmed this difference between the methods (Fig. S5).

Discussion

The goal of this work was to create a method that can find signals in image data without the need for manual adjustments based on visual inspection. This ability is important for a wide range of research and technological applications, such as in the analysis of

immunofluorescence signals from cohorts of patients. [34] The characteristics of images rarely are consistent across all instances; signal levels, relative proportions of signal and background, and shapes and sizes of signals can vary immensely, especially among images of complex material such as tissue. SFT introduces two features that can help with this challenge. The first is the use of statistical properties of background and signal regions that are more consistent between images than the properties mentioned above, and the second is the scanning of segments to determine best-fit relationships between the properties and to locate regions meeting certain criteria. We demonstrated that SFT was robust for the analysis of microarray and immunofluorescence data. The immunofluorescence images were particularly challenging application because they were acquired using several different antibodies and RNA probes and displayed large variation in signal intensities and distributions.

The method relies on certain foundational assumptions. First, it assumes that background regions have lower means, standard deviations, and CVs than signal regions. This assumption is reasonable because in background regions the source of variation between pixels is due mainly to random noise. Random variation in signals follows a Poisson distribution, so the great majority of pixels intensities will be within 2 standard deviations of the mean. True signal, on the other hand, will have additional sources of variation between pixels, such as the amount of analyte present in a region. This relationship, however, is not true for every segment, due to noise and signal variation, which is why we examine overall trends among segments. By fitting the relationships, we can find the best approximation of the value in one statistic that corresponds to a value in another statistic.

The method also assumes that the segments are small enough to properly sample the background; that is, that some areas of the image are mostly background, so that many of the segments are entirely in background regions. This assumption is fulfilled in nearly all data, because if measurements do not include background and are almost entirely signal, one has no basis for distinguishing signal from background. Furthermore, the user can adjust the segment size according to the resolution of the image.

The comparison with a fixed threshold and with Otsu’s method revealed the challenges in the analyses. A fixed threshold was not sufficient for data such as immunofluorescence, because background and signals levels varied so much between cases. Otsu’s method performed well in many cases but not where the histograms had irregular shapes, such as a low-signal peak from regions where there was no tissue. Potentially one could modify the method to account for no-tissue regions, but because such regions are not always present and have widely varying amounts and levels, predicting the characteristics would be difficult. Another problem with a single threshold, whether determined by Otsu’s method or set manually, is the allowance of “spikes” of single, high-intensity pixels. The SFT method requires the median of the whole segment to be above a segment threshold, thus eliminating spikes. SFT can locate fine features through the use of small segments and by requiring only one segment to be above the threshold.

We anticipate that this approach could be used in conjunction with additional statistics. For example, one could examine relationships between intra-segment correlations and median

signals. One might expect correlations between neighboring rows in a segment to be higher in true signal regions than in background regions. One also could examine relationships with auto-correlations—calculated as the correlation between segments shifted by a given number of pixels—as a means of finding signals, because one would expect regions containing no signal to have no auto-correlation but regions containing rising or falling signal to have higher auto-correlations.

While colocalization analysis was not the main focus of this work, the SFT approach could add to the colocalization toolkit by complementing or being used in conjunction with previous methods. The choice of colocalization method would depend on the goals and needs of the particular analysis, as reviewed in detail elsewhere[32]. In some cases, complete pixel overlap might be best, in other cases some flexibility in overlap would be good, and in other cases the identification of cellular components such as membranes may be necessary. The method presented here is a valid option when markers would be expected to be near one another but not necessarily overlapping, and not necessarily linked to subcellular localization. Examples would be two extracellular markers, or a membranous and extracellular marker, or two markers in cells with unpredictable shapes.

The SFT approach could be used for other data types beyond those presented here, such as linear data collected over time, three-dimensional images, or mass-spectrometry data. For each application, we predict that statistical relationships in background and signal regions exist that will allow locating signals without making assumptions about the amount, strength, size or shape of true signals. Only by eliminating the dependence on such assumptions will it be possible to reliably find signals across the diverse range of behaviors encountered in most experimental situations. The method presented here could meet such a requirement and contribute to the full automation of image analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge support of this work by the National Cancer Institute (Early Detection Research Network, U01CA152653; Alliance of Glycobiologists for Cancer Detection, U01CA168896) and the Van Andel Research Institute. We thank Katie Partyka for assistance preparing the microarrays; the VARI Pathology and Biorepository Core for assistance with tissue preparation; and the VARI Confocal Microscopy and Quantitative Image Core for assistance with immunofluorescence image acquisition.

References

1. Eliceiri KW, Berthold MR, Goldberg IG, Ibanez L, Manjunath BS, Martone ME, Murphy RF, Peng H, Plant AL, Roysam B, Stuurman N, Swedlow JR, Tomancak P, Carpenter AE. Nat Methods. 2012; 7:697–710. [PubMed: 22743775]
2. Niederlein A, Meyenhofer F, White D, Bickle M. Comb Chem High Throughput Screen. 2009; 9:899–907. [PubMed: 19531001]
3. Shamir L, Ling SM, Scott W, Hochberg M, Ferrucci L, Goldberg IG. Osteoarthritis Cartilage. 2009; 10:1307–1312. [PubMed: 19426848]
4. Cornelisse CJ, van Driel-Kulker AM, Meyer F, Ploem JS. J Microsc. 1985; (Pt 1):101–110. [PubMed: 3973916]

5. Alferez S, Merino A, Bigorra L, Mujica L, Ruiz M, Rodellar J. Am J Clin Pathol. 2015; 2:168–176. quiz 305. [PubMed: 25596242]
6. Varghese F, Bukhari AB, Malhotra R, De A. PLoS One. 2014; 5:e96801. [PubMed: 24802416]
7. Mulrane L, Rexhepaj E, Penney S, Callanan JJ, Gallagher WM. Expert Rev Mol Diagn. 2008; 6:707–725. [PubMed: 18999923]
8. Riber-Hansen R, Vainer B, Steiniche T. APMIS. 2012; 4:276–289. [PubMed: 22429210]
9. Newberg JY, Li J, Rao A, Ponten F, Uhlen M, Lundberg E, Murphy RF. Proc IEEE Int Symp Biomed Imaging. 2009:1023–1026. [PubMed: 20628548]
10. Shamir L, Delaney JD, Orlov N, Eckley DM, Goldberg IG. PLoS Comput Biol. 2010; 11:e1000974. [PubMed: 21124870]
11. Celli JP, Rizvi I, Blanden AR, Massodi I, Glidden MD, Pogue BW, Hasan T. Sci Rep. 2014:3751. [PubMed: 24435043]
12. Ljosa V, Caie PD, Ter Horst R, Sokolnicki KL, Jenkins EL, Daya S, Roberts ME, Jones TR, Singh S, Genovesio A, Clemons PA, Carragher NO, Carpenter AE. J Biomol Screen. 2013; 10:1321–1329. [PubMed: 24045582]
13. Murray JI, Bao Z, Boyle TJ, Boeck ME, Mericle BL, Nicholas TJ, Zhao Z, Sandel MJ, Waterston RH. Nat Methods. 2008; 8:703–709. [PubMed: 18587405]
14. Zhan M, Crane MM, Entchev EV, Caballero A, Fernandes de Abreu DA, Ch'ng Q, Lu H. PLoS Comput Biol. 2015; 4:e1004194. [PubMed: 25910032]
15. Choudhury KR, Yagle KJ, Swanson PE, Krohn KA, Rajendran JG. J Histochem Cytochem. 2010; 2:95–107. [PubMed: 19687472]
16. Yoshitomi H, Togawa A, Kimura F, Ito H, Shimizu H, Yoshidome H, Otsuka M, Kato A, Nozawa S, Furukawa K, Miyazaki M. Cancer. 2008; 9:2448–2456. [PubMed: 18823024]
17. Tang C, Lee AS, Volkmer JP, Sahoo D, Nag D, Mosley AR, Inlay MA, Ardehali R, Chavez SL, Pera RR, Behr B, Wu JC, Weissman IL, Drukker M. Nat Biotechnol. 2011; 9:829–834. [PubMed: 21841799]
18. Liu Y, Liang G, Saha PK. Med Phys. 2012; 1:514–532. [PubMed: 22225322]
19. Torre V, Poggio TA. IEEE Trans Pattern Anal Mach Intell. 1986; 2:147–163. [PubMed: 21869334]
20. Kim JH, Kim HY, Lee YS. Exp Mol Med. 2001; 2:83–88. [PubMed: 11460886]
21. Barba J, Jeanty H, Fenster P, Gil J. J Microsc. 1989; Pt 1:125–134. [PubMed: 2685317]
22. Gomez W, Leija L, Alvarenga AV, Infantosi AF, Pereira WC. Med Phys. 2010; 1:82–95. [PubMed: 20175469]
23. van Tonder GJ, Ejima Y. Neural Netw. 2000; 3:291–303. [PubMed: 10937963]
24. Kalaidzidis Y. Eur J Cell Biol. 2007; 9:569–578. [PubMed: 17646017]
25. Chen S, LaRoche T, Hamelinck D, Bergsma D, Brenner D, Simeone D, Brand RE, Haab BB. Nat Methods. 2007; 5:437–444. [PubMed: 17417647]
26. Yue T, Goldstein IJ, Hollingsworth MA, Kaul K, Brand RE, Haab BB. Mol Cell Proteomics. 2009; 7:1697–1707. [PubMed: 19377061]
27. Yue T, Maupin KA, Fallon B, Li L, Partyka K, Anderson MA, Brenner DE, Kaul K, Zeh H, Moser AJ, Simeone DM, Feng Z, Brand RE, Haab BB. PLoS ONE. 2011; 12:e29180. [PubMed: 22220206]
28. Tang H, Singh S, Partyka K, Kletter D, Hsueh P, Yadav J, Ensink E, Bern M, Hostetter G, Hartman D, Huang Y, Brand RE, Haab BB. Mol Cell Proteomics. 2015
29. Winter JM, Tang LH, Klimstra DS, Brennan MF, Brody JR, Rocha FG, Jia X, Qin LX, D'Angelica MI, DeMatteo RP, Fong Y, Jarnagin WR, O'Reilly EM, Allen PJ. PLoS ONE. 2012; 7:e40157. [PubMed: 22792233]
30. Sempere LF. Methods Mol Biol. 2014:151–170. [PubMed: 25218384]
31. Qin L, Rueda L, Ali A, Ngom A. Appl Bioinformatics. 2005; 1:1–11. [PubMed: 16000008]
32. Bolte S, Cordelieres FP. J Microsc. 2006; Pt 3:213–232. [PubMed: 17210054]
33. Otsu N. IEEE Trans Sys, Man, Cyber. 1979; 1:62–66.
34. Kallioniemi OP, Wagner U, Kononen J, Sauter G. Hum Mol Genet. 2001; 7:657–662. [PubMed: 11257096]

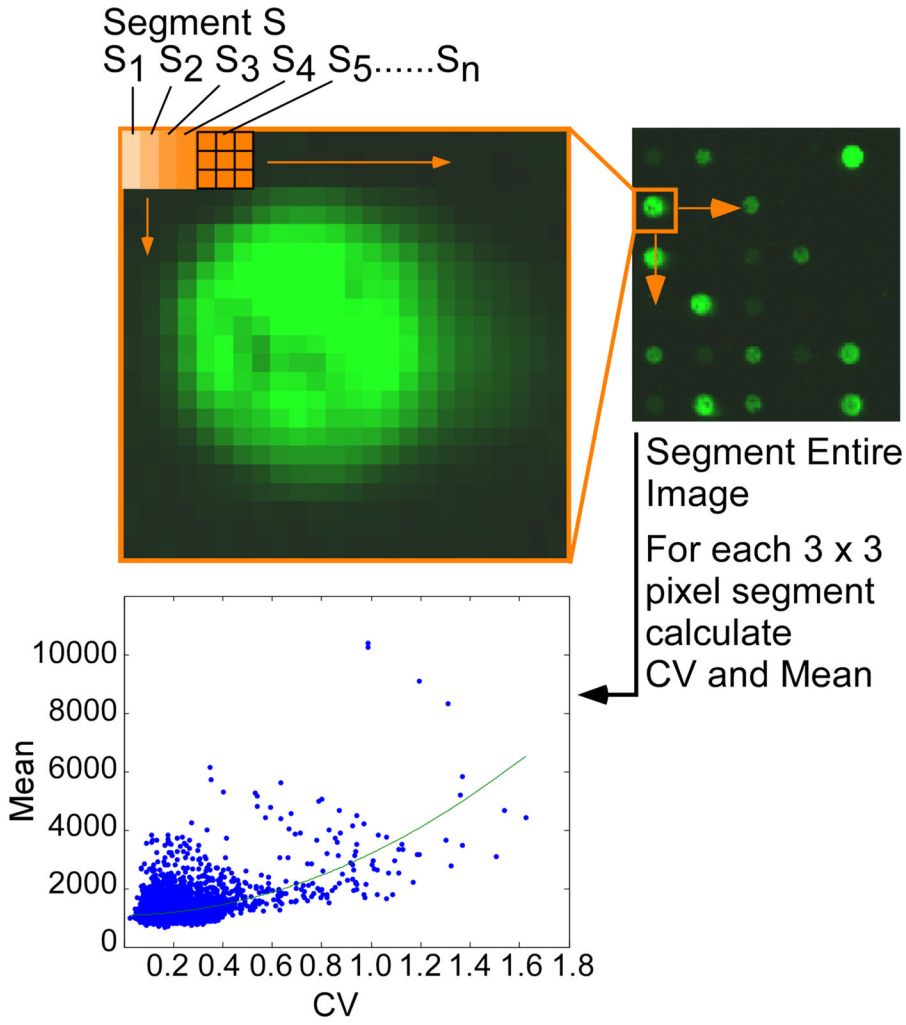

Figure 1. Segmenting and fitting image data

The program uses a sliding box to segment the entire image. For each segment, the program calculates statistics such as mean, median, standard deviation, and coefficient of variation (CV). The results from selected statistics are plotted against each other and fit with a quadratic line.

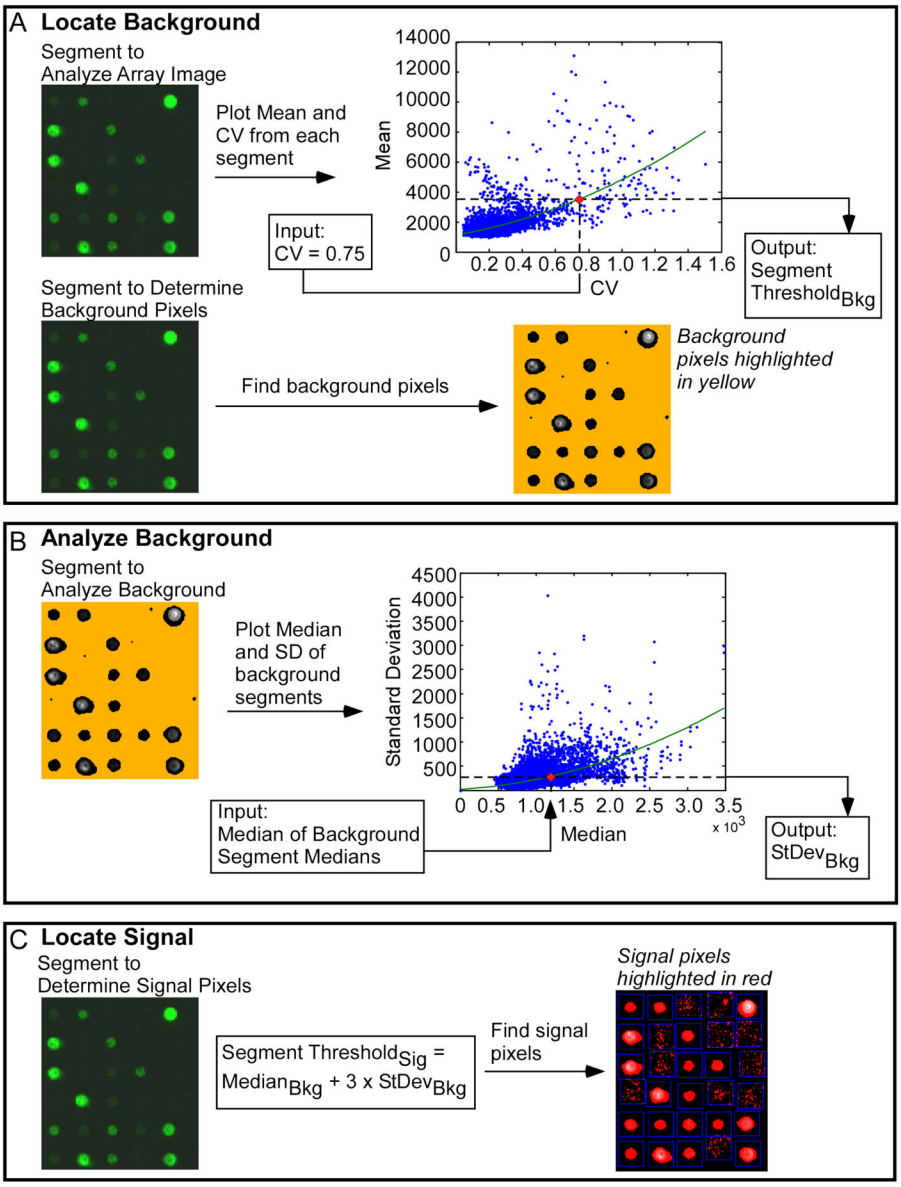

Figure 2. Using SFT to locate signals in images

A) Background location. The program segments the image, plots the means of the segments with respect to their CVs, and finds the best quadratic fit. A CV value, determined empirically as a typical maximum CV in background regions, is input into the line of best fit to obtain the corresponding mean as a background threshold. Each segment is compared to the background threshold to identify the background pixels. B) Background analysis to determine signal thresholds. The program computes the medians and standard deviations of segments that contain mostly background pixels. The median of all segment medians is entered into the equation from the quadratic fit to calculate the corresponding standard deviation. C) Signal location. A threshold for each segment is calculated using the median and standard deviation of the background. Pixels are counted as signal based on the

percentage of segments in which they are included that are above the threshold, and on whether they exceed a pixel threshold.

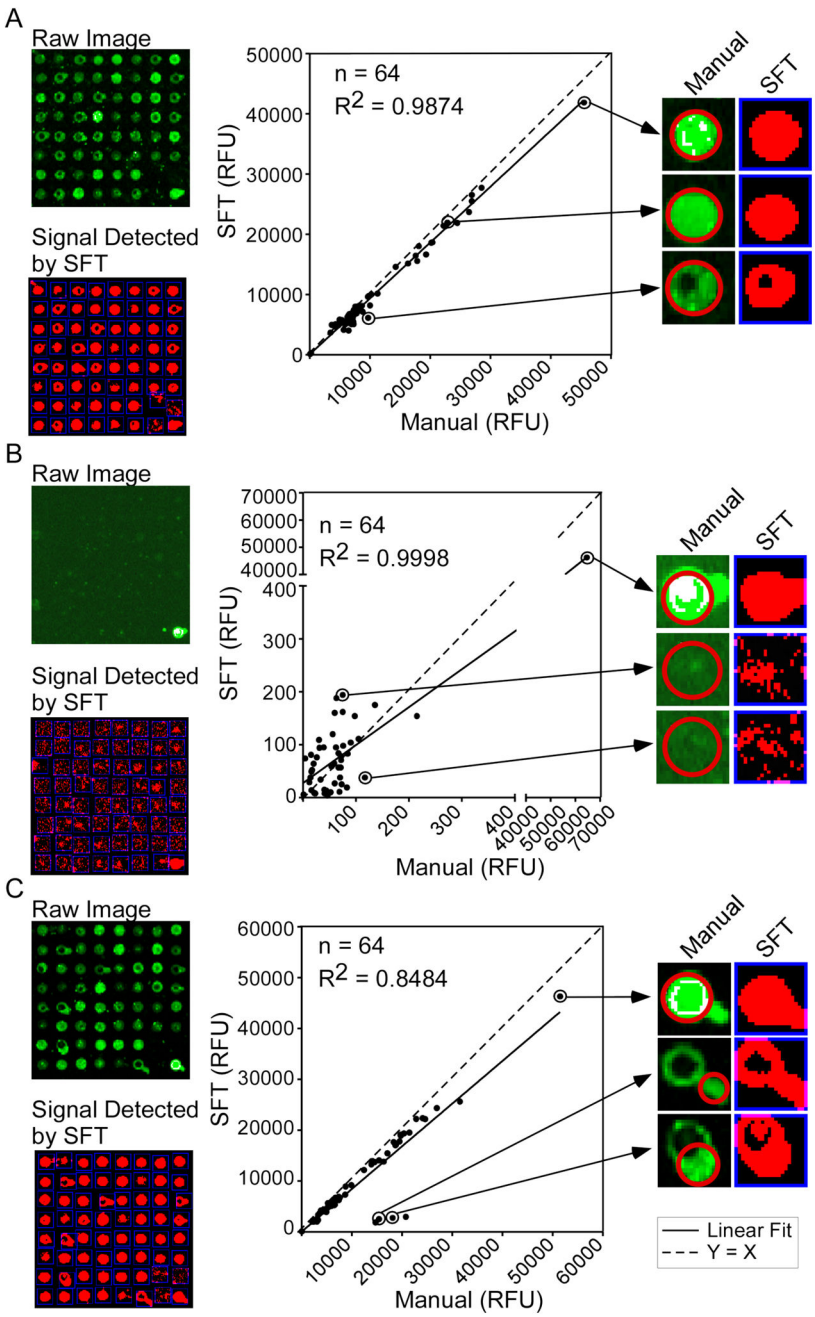

Figure 3. Application of SFT to microarray data and comparison with manual analysis

We quantified the median signal in each spot both by SFT and by manual analysis and plotted the resulting values with respect to each other. In the zoomed images of individual spots, the pixels within the red ring were counted as signal pixels for the manual method. A) Using an array with many spots containing signals, the correlation between methods over all spots was very good. B) In arrays with low signals, the correlation was less for weak spots but still apparent. SFT properly found very weak signals. (C) In an array with defects to spot morphologies, outliers to the overall correlation were typically caused by bias in selecting signal regions in the manual method.

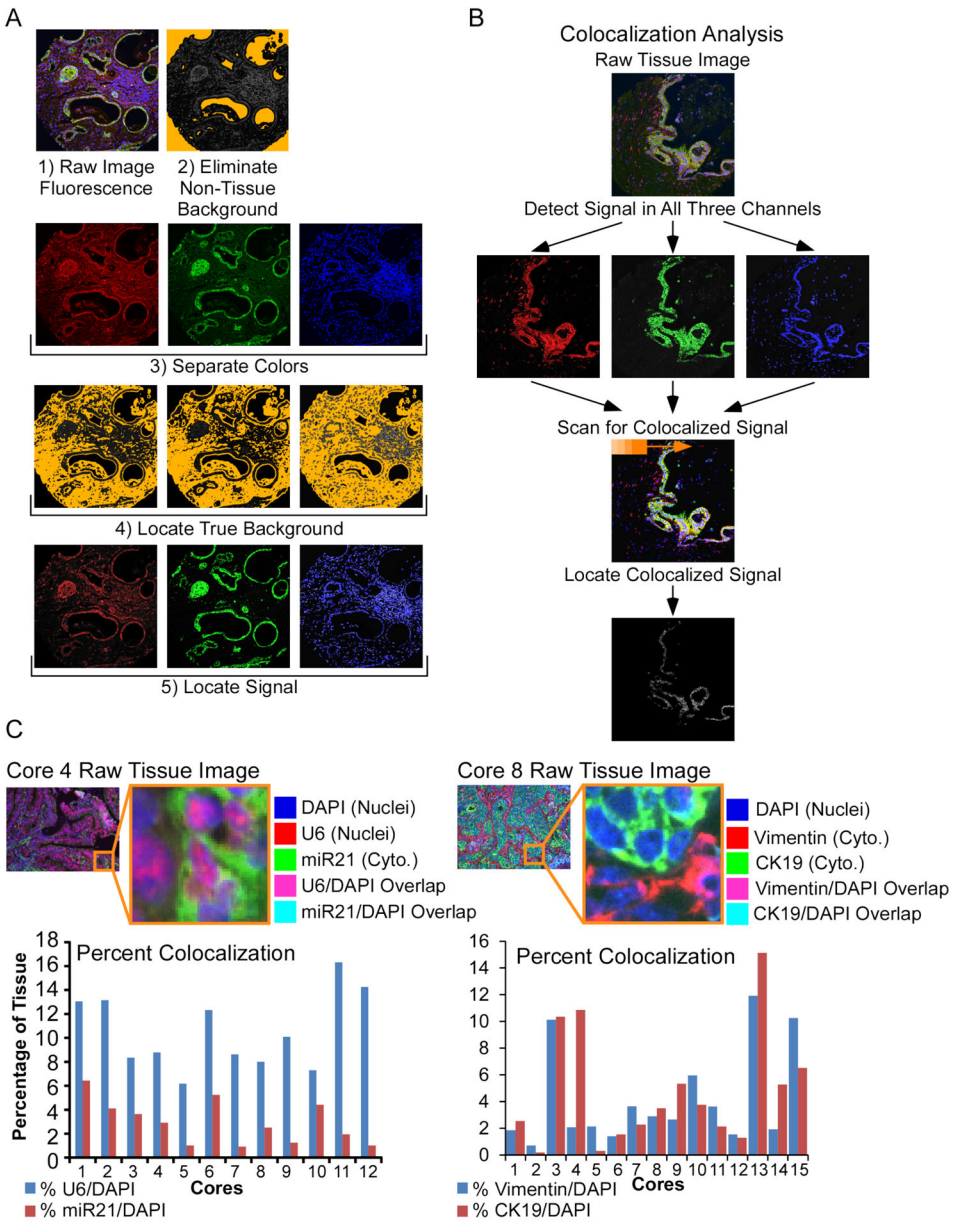

Figure 4. Application of SFT to mulit-color immunofluorescence data

A) Processing steps in 3-color immunofluorescence data. The SFT method determines a threshold to locate and eliminate the regions of the raw image (panel 1) that contain no tissue. The non-tissue pixels are highlighted gold in panel 2. The program separates the channels (panel 3); identifies the background pixels (highlighted gold in panel 4); and locates the signal pixels (highlighted in their respective colors in panel 5). B) Colocalization analysis. The program scans segments to find those with a minimum number of signal pixels in each segment. C) Application to markers with known subcellular locations. SFT properly quantified relative levels in colocalization between DAPI and the indicated markers.

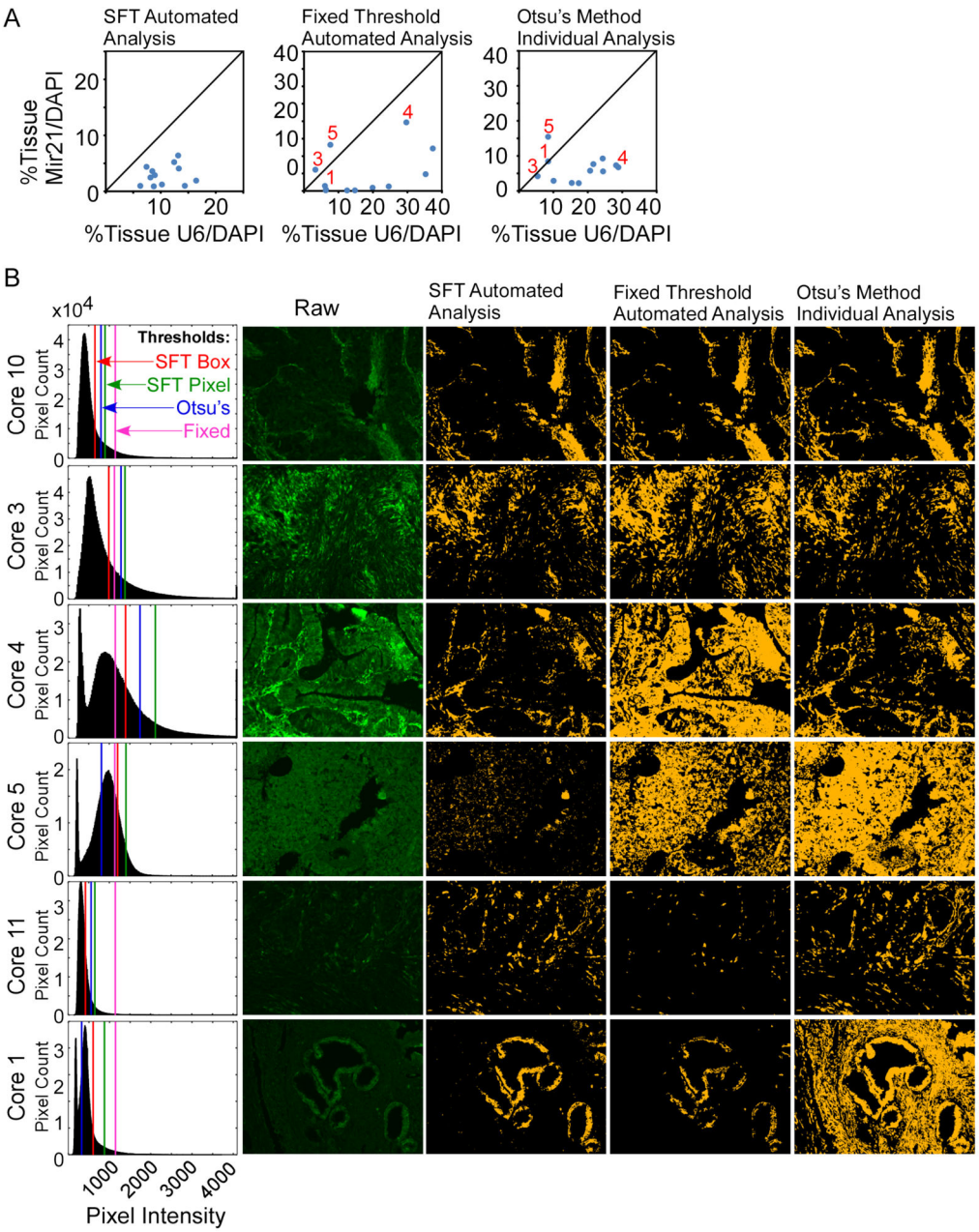

Figure 5. Comparisons between methods

A) miR-21/DAPI versus U6/DAPI colocalization. SFT found a higher level of U6/DAPI colocalization in each case, but the other methods found higher levels of miR-21/DAPI colocalization in certain cases. B) Distributions and identified signals in the miR-21 channel. The thresholds were in the proper location of the tail of the pixel-intensity histogram in some cases (core 2) but not always. An improper threshold resulted either in detected signal from entire cellular region (cores 3, 5, and 6) or missed signal (core 4). In each case, the SFT box and pixel thresholds were in the tail of the histogram and resulted in pixel detection that matched the signal regions in the raw image.