Journal of proteome research

Article

pubs.acs.org/jpr

Crossword: A Fully Automated Algorithm for the Segmentation and Quality Control of Protein Microarray Images

Todd M. Gierahn,[‖] Denis Loginov,[§] and J. Christopher Love*[,†,‡,‖]

[†]Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

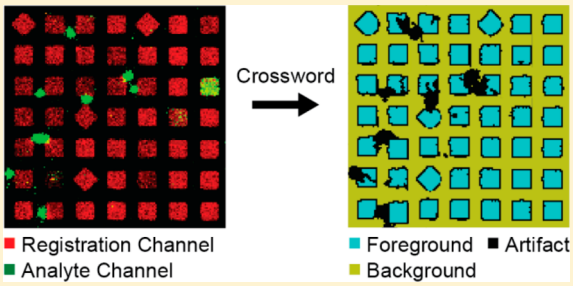
[‡]The Ragon Institute of MGH, MIT, and Harvard, Charlestown Navy Yard, Boston, Massachusetts 02129, United States

[§]Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

[‖]The David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Supporting Information

ABSTRACT: Biological assays formatted as microarrays have become a critical tool for the generation of the comprehensive data sets required for systems-level understanding of biological processes. Manual annotation of data extracted from images of microarrays, however, remains a significant bottleneck, particularly for protein microarrays due to the sensitivity of this technology to weak artifact signal. In order to automate the extraction and curation of data from protein microarrays, we describe an algorithm called Crossword that logically combines information from multiple approaches to fully automate microarray segmentation. Automated artifact removal is also accomplished by segregating structured pixels from the background noise using iterative clustering and pixel connectivity. Correlation of the location of structured pixels across image channels is used to identify and remove artifact pixels from the image prior to data extraction. This component improves the accuracy of data sets while reducing the requirement for time-consuming visual inspection of the data. Crossword enables a fully automated protocol that is robust to significant spatial and intensity aberrations. Overall, the average amount of user intervention is reduced by an order of magnitude and the data quality is increased through artifact removal and reduced user variability. The increase in throughput should aid the further implementation of microarray technologies in clinical studies.



KEYWORDS: *image segmentation, artifact removal, image quality control, automated image processing, pixel clustering, pixel connectivity*

INTRODUCTION

Multiplexing biological assays on microarrays has been instrumental for generating the data required to understand biology at a systems-level. Although microarrays have contributed to the exponential growth of biological data over the past decade, there remain bottlenecks in the use of this technology. A primary challenge is the accurate and expedient analysis of the data contained within the images of the microarrays. There are multiple commercial and open source software packages commonly used to analyze images of arrays,[1−5] but these packages do not automatically handle the types of distortions commonly observed in protein microarrays (e.g., loss of signal, comet tails, grid distortions, etc.). Often, considerable manual adjustments of an overlaid nominal grid are needed to align it properly to individual features of interest. Most software packages also provide few, if any, tools to identify artifacts in the images, thereby requiring a manual review of features after data analysis to ensure the quality of the data. This is a particular concern for protein microarrays, since the data are more sensitive to low amounts of spurious signal than DNA microarrays. The majority of features in analyte channels typically lack signal, enabling small amounts of spurious signal to increase the measured intensity of a feature above the positive threshold. Overall, manual segmentation and annotation not only cost hours of trained labor per array but also introduce significant variability in data analysis among users. With the motivation to create ever larger data sets for systems-level analyses of clinical samples, these inefficiencies represent a significant barrier in efforts to understand human disease and mechanisms of action for interventions.

Numerous approaches have been proposed to extract information from images of microarrays by both semi-automated and fully automated methods.[6] Initial algorithms used the user-defined specifications of the array and template matching to find the optimal alignment with the expected

Received: March 7, 2013
Published: January 13, 2014

   dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362−371

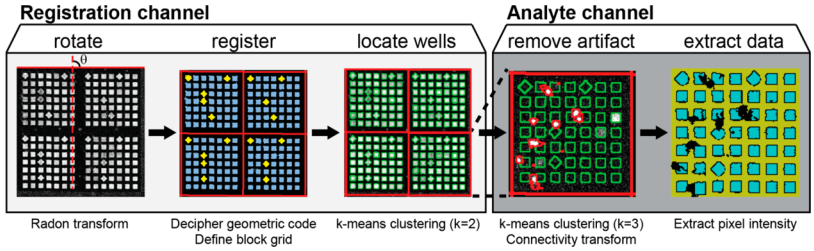


Figure 1. Schematic outline of the Crossword algorithm. The image is initially rotated to make the major axes of the array parallel to the edges of the image. The geometric code defining the identity of each block is then deciphered to register the array. An adaptive gridding algorithm draws grid lines between the blocks of the array. The location of each feature is then identified using iterative *k*-means clustering of the registration channel image. Next, artifactual pixels in each analyte channel are removed using a combination of *k*-means clustering and a custom connectivity transform algorithm to identify signals in the analyte channels and a geometric comparison of the identified signal across the array channels. Finally, the intensities of nonartifactual pixels are extracted for analysis.

geometries.[3] This approach was later expanded to allow adaptive deformation of the templates.[1,2] Approaches using 1-D image projections with and without image filtering were also implemented for aligning grids to arrays.[7,8] Clustering algorithms used to segregate foreground and background pixels have proven to be some of the most successful algorithms for identifying features.[9−11] Other mathematical approaches have also been proposed, including seeded region growing,[5] SVM,[12] Markov random fields,[13] and mixture models,[4,8] among others. With a few notable exceptions,[11,14,15] most implemented approaches rely heavily on a single approach. In our experience, however, no single approach or algorithm can automatically handle the variety of aberrations that occur on experimentally produced arrays, especially protein-based arrays. A better approach therefore would combine the strengths of different methodologies in a logical way to create an algorithm for hierarchical gridding that is more robust to a variety of aberrations common in protein microarrays. The availability of cheap computational power and parallelized processing enables the implementation of multiple approaches to registering and aligning features on the same array within a reasonable processing time.

The increase in computing power also allows a more in-depth analysis of the signal in the image, enabling the deconvolution of true signal and artifact prior to the extraction of data. Multiple approaches have been proposed to qualify data extracted from microarrays, but most rate the reliability of each data point using metrics assigned after the initial extraction of data (e.g., covariance (CV), signal-to-noise ratio (SNR)).[16−18] Far more information is encoded in the image itself, particularly at the granularity of the location and intensity of individual pixels. Robust analysis of the image prior to data extraction should yield a more accurate description of the artifacts present and, thus, enable their removal, yielding more accurate and reliable data than qualification of data after extraction. Previously implemented techniques for identifying artifacts in images of microarrays include clustering pixels into more than two groups to identify and remove outlier pixels,[4,8] and a geometric comparison of bright pixels to a feature template.[9,10,19] These approaches, however, can miss artifacts that overlap the foreground or that have intensities similar to the true signal, and most have problems with overclustering of images containing only "salt-and-pepper" noise. New methods for identifying artifacts in the image prior to data extraction would enhance the accuracy of data sets generated from microarrays, particularly for protein microarrays.

Microengraving is a technique developed to produce protein microarrays wherein each spatial element represents proteins captured from individual, or a small number of, cells.[20] It has been used to discover antigen-specific antibodies from hybridomas and primary B cells,[21] to select mammalian[22,23] and microbial[24] hosts producing biologic drugs, and to profile the functional responses of activated T cells.[25,26] In this method, a slide uniformly coated with multiple capture antibodies seals against an elastomeric array of subnanoliter wells (nanowells) containing cells. The slide captures specific analytes secreted from the cells along with a known protein present in the media (e.g., IgG) to label the location of each well on the slide. After incubation, the slide is removed from the device and stained using fluorescent antibodies in a manner similar to conventional protein microarrays. This process can be repeated multiple times on the same cells to temporally monitor secretion from the same cells over time or measure other analytes using slides coated with different capture antibodies. Microengraving can generate tens of microarrays per biological sample. This large number of arrays makes automated analysis critical for the timely evaluation of data and the widespread implementation of this technology for use in clinical trials with large numbers of biological samples.

Here we describe a new, fully automated algorithm that can robustly and accurately analyze microengraved arrays containing significant aberrations. Microengraved arrays have several characteristics that make them particularly difficult to analyze automatically using previously published approaches and, therefore, a useful model for devising accurate and efficient means to extract data and overcome image artifacts. The features are small (30−50 μm) and densely packed (60−100 μm pitch). The typical signal intensity (~1.5- to 3-fold above background) is considerably lower than that of a typical DNA microarray, requiring sensitive methods for distinguishing the signal. Although the elastomeric arrays are manufactured with high precision, the process of microengraving often distorts the lateral structure of the array and individual features in a nonuniform manner. The arrays are also very large, covering most of the surface area of a typical microscope slide, with more than 80,000 features. Nonlinear distortions over large distances make approaches relying on placing straight lines to delineate the array difficult to implement. The features are arranged in small blocks (50−100 features/block), and it is not uncommon for a row or two of blocks to be missing from an edge of the array. This loss of features makes automatic registration of blocks impossible with any published technique. Finally, each feature in a microengraved array represents a unique measure-

dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362−371

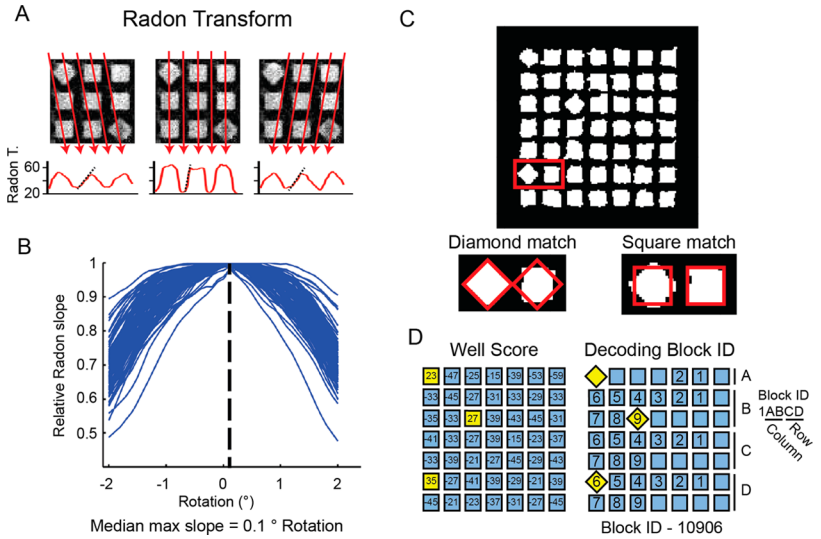


Figure 2. Rotation and registration of the array. (A) Example images of radon integrating vectors at three discrete angles (−10, 0, and 10°) and the resulting radon transform of the image at those angles. The slope of each radon transform curve is marked with a dotted line. (B) Plot of the sums of the absolute slope of the radon curve at each degree of rotation normalized to the highest value for that sample image as a function of the degree of rotation for 500 subimages sampled from the same array. The degree of rotation most frequently yielding the highest radon slope (dotted line) defines the rotation of the array. (C) The algorithm used to decipher the geometric code in each block. A block is transformed into a binary image using a threshold defined by the expected number of positive pixels and the rank order of pixel intensity. The center of each feature is matched to the center of both a diamond and a square template. The well score for each feature is calculated from the number of positive and negative pixels within each template. (D) The well score for each feature in the block (left array) and the cipher used to decode the block location (right panel). In the left panel, a positive score indicates a diamond feature and is highlighted in yellow; a negative score indicates a square and is highlighted in blue. The location of the diamonds defines the row and column number for the block.

ment—a fact that negates any approach that relies on replicate spots for quality control. Here we describe a hybrid algorithm called Crossword that combines a multifaceted grid-aligning pipeline with a versatile framework for identifying artifactual pixels within the image. This method should greatly aid the application of microengraving technology to clinical samples as well as improve the quality of data extracted from standard spotted microarrays.

MATERIALS AND METHODS

Overview of Crossword

Crossword was designed to be an integrated software package that receives as input the filename and basic array specifications of an array (Supporting Information Tables 1 and 2) and returns measurements of the signal quality and intensity in the foreground and background of each feature with no human intervention (Figure 1, Supporting Information Table 3). The output of a microengraving experiment is a TIFF image that contains a fluorescent image of the array for each analyte (analyte channels) and one image for the molecule included in the media (e.g., IgG or a specific cytokine) to define the locations of each well (registration channel). The algorithm initially estimates the global rotation of the major axes of the array (*x*- and *y*-axes) and then straightens the image with respect to each axis. The nominal location of each block is next registered to the image using a novel code that distinguishes each block geometrically. Lines delineating a grid on the array are then adaptively drawn between blocks. The subimage for each feature in the delineated blocks is iteratively clustered to define the foreground area. A novel algorithm for quality control is then applied to the image of each feature to identify pixels containing artifacts. Finally, the signal intensity, variance, and quality of the background and foreground pixels in each feature is extracted and returned in a tabular format along with images of each feature. Batched processes for multiple files can be implemented to allow the analysis of tens of arrays overnight with no human intervention.

Registration of Array

The initial module in the algorithm uses linear Radon transformations to estimate the rotation of the major axes of the array, similar to the method reported by Rueda et al.[15] A Radon transformation integrates the signal that a parallel set of vectors encounters crossing an image as the angle between the image and the vectors varies. Figure 2A depicts representative Radon integrating vectors at three discrete angles, and the resulting Radon transforms of the image at those angles. When the Radon transform angle matches the rotation of the array, the slope of the transform between areas of high and low signal is maximized. For microengraved arrays, long-range, nonlinear distortions in the structure of the array and an uneven signal sometimes make the Radon transform less accurate at determining image rotation when applied to the whole image. The algorithm, therefore, applies the transformation to many subimages containing two vertical blocks sampled from across the array. Each image is checked for the appropriate frequency of features to ensure the block images are of good quality. The rotation of each subimage is then determined (Figure 2B). Nearest neighbor interpolation is used to rotate the entire image by the median rotation angle of the subimages to yield an array whose axes are aligned with the edges of the image.

After straightening the image, the location of each block in the array is registered to assign an identity to each block. Microengraved arrays often fade near the edges due to imperfect sealing. This effect can lead to misalignment of the entire grid when the identity of each block is defined using only the visible edges of the array. To overcome this issue, the arrays

dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362−371

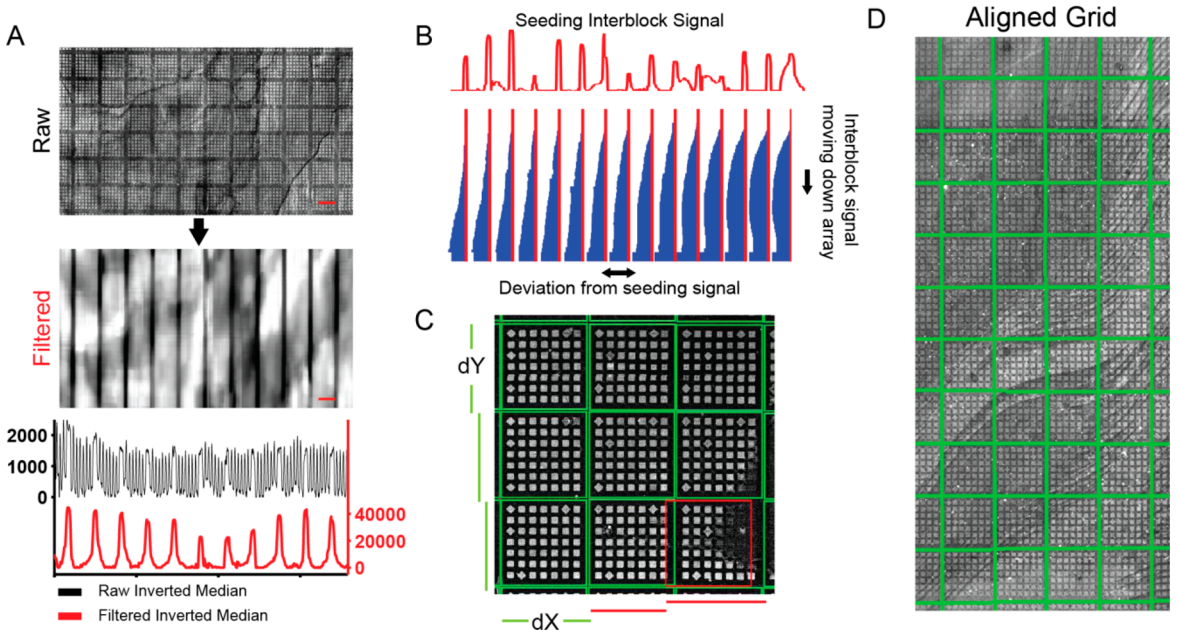


Figure 3. Assignment of interblock grid lines. (A) Raw image of an array and its transformation after passing through a series of low-pass filters designed to emphasize the interblock vertical frequencies (red bar = 500 μm). The 1-D projections of the raw (black) and transformed (red) images are displayed below the images. (B) The deviation from the initial interblock signal used for seeding (top signal trace) as a function of the *y* position in the array. The *x* position of the initial seeding signal is drawn in red. The *x* position of the actual signal at each *y* location in the array is noted by a horizontal blue bar. (C) The final error correction in the block gridding procedure. The top left corner of each block is compared to the neighboring blocks. Blocks that are not correctly positioned (red) are relocated using their neighbors' locations (green). (D) An example of a final overlay for the blocks in an array.

of nanowells employ an internal geometric code using diamond and square wells within each block; this code allows nonambiguous identification of any isolated block.[20] Crossword robustly deciphers these geometric codes. The algorithm breaks the image of the array into overlapping, block-sized subimages. Each subimage is then tested for the appropriate frequency of features using 1-D projections along the *x*- and *y*-axes to identify subimages that contain an entire block. These subimages are transformed into binary images by rank ordering the pixels based on intensity and selecting the expected number of pixels with high intensities as positive. The expected number of pixels derives from the array specifications passed to Crossword by the user for the number and size of features in each block (Figure 2C). Each individual feature is then aligned with a square and diamond template sized to fit the average feature size of the block (Figure 2C). The number of negative pixels within each template is subtracted from the number of positive pixels in each template. The total from the diamond template is then subtracted from the total from the square template to yield the final feature score (Figure 2D). A negative score for a feature identifies a square feature while a positive score identifies a diamond-shaped one.

Once the feature scores are calculated for each subimage, they are used to identify the blocks present in each subimage. All blocks contain one diamond in the upper left-hand corner to establish the orientation of the array. The algorithm initially identifies the corner feature that was most frequently identified as a diamond and then automatically flips the entire array (and all subimages) to ensure proper orientation of the geometric code. The code is then deciphered using a lookup table (Figure 2D). In this way, Crossword assigns a putative unique identifier to each block. The identity of each block with a valid block code is then corroborated using its spatial (*x*,*y*) location within the array to estimate the location of the top left block. All blocks that yield a location within a distance of one-half the block pitch from each other are deemed correct, and the locations of those blocks are used to register the remainder of the blocks in the array. The rates at which blocks were correctly identified by this algorithm for several representative arrays are given in Supporting Information Table 4.

Fine Adjustments of Aligned Grid

In principle, block registration gives sufficient information to locate all blocks in an ideal array by evaluating the interblock distances. We found, however, that these linear relationships often fail to correctly identify the precise location of blocks due to the nonlinear distortions in experimentally derived arrays. We, therefore, developed a method to adaptively assign the positions of blocks that monitors these distortions and ignores spurious and missing signals in parts of the array.

The process begins by applying a series of bandpass filters to the image to emphasize either the vertical or horizontal interblock signals (Figure 3A, Supporting Information Figure 1) similar to the approach recently suggested by Wu et al.[14] The interblock signal is then progressively monitored every 100 pixels across the array using 1-D projections along the axis perpendicular to the emphasized signal (Supporting Information Figure 2). A 1-D projection based on the locations of blocks generated during registration establishes an initial template. This model projection is then shifted in relation to the actual 1-D projection of the first 100 pixels to find the location where the largest number of signals in the projection match the locations specified by the template. The movement is limited to 33% of the block pitch to ensure the grid remains properly registered. If half of all signals or a consecutive run of 20% of the signals in the template match a signal in the projection, the location of each interblock signal is seeded at

dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362−371

the location of the template regardless of whether a signal was detected. If the two projections cannot be matched, the position of the template is used as the location on the array, and then the next 100-pixel region is examined. Once the interblock signal is seeded, signals identified in further projections are aligned to the closest seed signal. If multiple signals align to the same seed signal, they are all discarded. If a signal is more than 10 pixels away from any seed signal, it is also discarded as a spurious signal, since distortions of this magnitude over a 100-pixel distance have not been observed empirically in micro-engraved arrays. If no signal matches a seed signal, the seed signal is carried forward to the next projection. Matched seed signals are replaced by the new signal location. This approach enables long-range, nonlinear distortions in the array to be effectively monitored even if the distortions are not correlated across the array (Figure 3B).

To further increase the accuracy of the block positions, we implemented a metric for quality control that measures the reliability of each grid line at each point in the array. As the grid lines are defined, a corresponding matrix keeps track of whether a signal for each line was detected in a given projection (Supporting Information Figure 2). When the location of each interblock signal is initially seeded by the template, the quality metric for each line is set to zero. In subsequent alignments of the projection, if a signal is not detected, the metric for that line is increased by one. If a signal is detected, the total score decreases by two until zero is reached.

The grid metric is used in two ways to increase the accuracy of the grid. First, neighboring grid lines are compared to each other. If the distance between the lines differs from the expected block distance by more than the feature pitch, the grid metric is used to adjust the less reliable grid line. The gridding process is also performed in both directions across the array (e.g., left-to-right and right-to-left) to address issues of poor signal on the edge of the array during grid seeding. The second use of the grid metric is to choose the best grid location when the grids drawn in opposite directions do not match (Supporting Information Figure 3).

As a final control for the quality of the aligned grid, the top left corner of each block is estimated using 1-D projections of the feature signals in the block areas defined by the grid lines. The corner of each block is then compared to neighboring blocks to make sure the block-to-block distance is within appropriate bounds (<5%) (Figure 3C). Any block that is misaligned with neighboring blocks is realigned, creating the final grid of blocks (Figure 3D). At the end of the process, each grid point has been assessed five different times, yielding a robust protocol for aligning a grid that automatically handles gross distortions in the array, large areas of missing signal, uneven backgrounds, and other artifacts that can occur in experimentally produced arrays.

Identification of Features

Once the location of each block is defined, features are precisely located using a hybrid clustering/template method. A feature grid is created for each block using 1-D projections on each axis (Supporting Information Figure 4a). Next, *k*-means clustering is used to identify pixels in the foreground and background of each feature. An initial analysis using data from a manually curated array was performed using different approaches in the literature to determine the optimal attributes of the data to use for clustering.[11,27] We found that the raw pixel intensities combined with neighboring pixel intensities generated foreground and background clusters that yielded signal-to-noise ratios that best matched the manual curation (Supporting Information Figure 5). Including a preprocessing transformation of the data, location of the pixel in the feature image or a third cluster did not improve the resulting data.

As suggested by Rahnenführer et al.,[9] we implemented an iterative approach to clustering because very bright artifact pixels disrupted the identification of the foreground area when a single instance of clustering was used (Supporting Information Figure 4b). The size of both clusters is compared to the nominal expected sizes of the foreground and background for a feature. If either is less than a third of the expected size, those pixels are discarded as either very bright or very dim artifacts that biased the clustering of the other pixels. The remaining pixels are reclustered until the size criteria are met. In order to more accurately locate the feature, the final cluster with the higher mean pixel intensity is convoluted with a template kernel (Supporting Information Figure 4c). The shape of the kernel is defined by the block code and known feature location; the size is defined by the average area of the feature calculated by the algorithm while registering the array. The design of the kernel emphasizes the edges of the feature. The optimal location for the foreground is centered at the pixel that has the highest value in the convolution of the cluster and the kernel (Supporting Information Figure 4d). To account for features that differ in size from the average feature size, pixels in the top cluster that are within 2 pixels of the foreground area are excluded from further analysis to avoid potential foreground pixels contributing to background measurements, and groups of negative pixels larger than 4 within the template are also excluded to account for smaller feature sizes (Supporting Information Figure 4d).

Removal of Artifact

To improve the quality of the data extracted from each feature in the array, we implemented a new approach for automated identification of artifacts. We found previously reported methods based on iterative clustering[9] or mixture models[4] were effective in most cases at separating artifact from signal in the same image. When analyzing images containing only "salt-and-pepper" noise, however, we found the methods either overclustered the pixels or, in the case of the published iterative clustering method, failed entirely, as the method requires the accumulation of signal in the foreground to trigger clustering termination, which cannot happen if there is no signal in the image. Efficiently handling images containing only noise is essential in the context of protein microarrays, as the vast majority of features are negative in the analyte channels. We initially attempted to identify and exclude images containing only noise from the cluster analysis using measures of the skewedness of the pixel intensity distribution or image entropy but could not effectively recognize all blank images. We therefore developed a novel approach to enable artifact removal using iterative clustering that is robust to the presence of images containing only noise.

Our iterative clustering approach is based on pixel intensity, location, and correlation between image channels. The module for removing artifacts begins by identifying structured pixels in each feature subimage using iterative *k*-means clustering (where $k = 3$) and an image transformation that we call a connectivity transformation. The connectivity transformation changes the values of positive pixels in a binary image to the pixel's connectivity value—that is, the number of positive pixels any

dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362−371

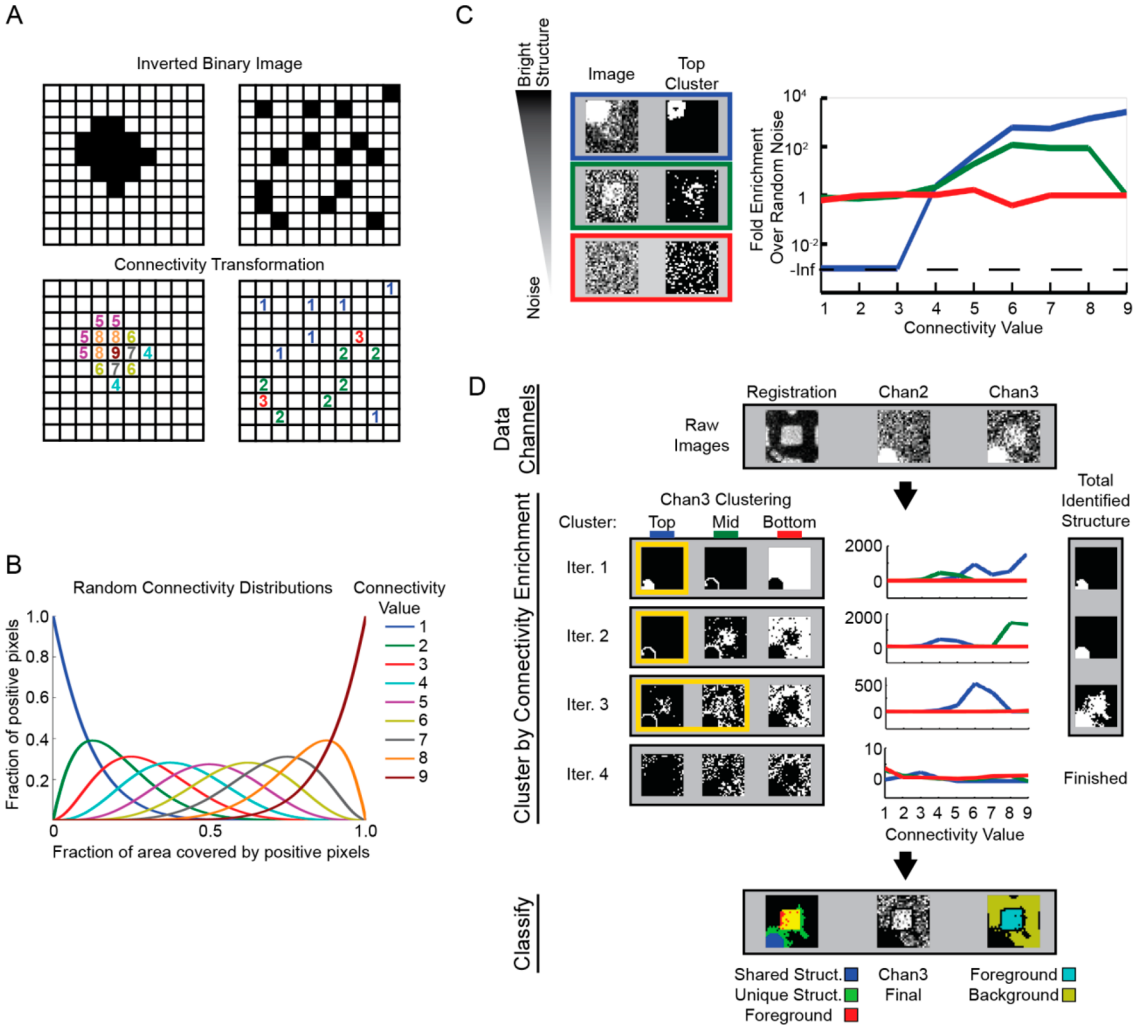


Figure 4. Identifying artifacts using connectivity transform. (A) Examples of binary images and their connectivity transforms. Each positive pixel receives a value that is the sum of the number of adjacent positive pixels and itself. (B) Plots of the expected fraction of positive pixels in a binary image with a particular connectivity value based on random noise as a function of the fraction of positive pixels in the image. (C) Examples of three types of images, their top ranked clusters determined by *k*-means clustering, and the enrichment of connectivity values in each cluster relative to random noise. The plot of the enrichment in the connectivity values is color coded according to the color of the box surrounding each image. (D) Schematic for the process used to remove artifacts. The registration channel and two analyte channels for the same feature are displayed in the top box. Beneath, the iterative clustering of Channel 3 is depicted. The binary images of the three clusters created by *k*-means clustering of the pixels based on their intensities and their neighbors' intensities are displayed for each iteration. The enrichment in connectivity values over random noise for each cluster is plotted next to the binary images. The structure removed in each iteration combined with the previous iterations is displayed next to the enrichment plots. The binary image/images of the clusters defined as structure in each iteration is/are boxed in yellow. Clustering is terminated when no cluster is significantly enriched in connectivity values relative to noise. The first image in the bottom box overlays pixels that were identified as structure in both analyte channels (blue), as structure only identified in Channel 3 (green), and as defined foreground area (red). The second image depicts Channel 3 with the pixels designated as artifact removed. The third image displays the final foreground and background of this feature.

given pixel is touching, including itself in our implementation (Figure 4A). The frequency of each connectivity value in a given cluster is then determined, thereby providing a simplified description of the proximity of the pixels to each other in the cluster. The expected frequencies of connectivity values for random noise depend on the fraction of the available space occupied by positive pixels (Figure 4B)—that is, the more space taken up by positive pixels, the more pixels touch each other. To determine whether the pixels in a cluster are part of a defined structure, the frequency of each connectivity value is compared to the frequency expected from random noise (given the same fraction of space covered by positive pixels). The presence of structured pixels in the image leads to an enrichment of high connectivity values (6–9) compared to noise (Figure 4C). Clusters derived from images with bright structures are enriched in high connectivity values and depleted of low connectivity values relative to random noise. Weak or diffuse structures show a modest enrichment in high connectivity values, and clusters from apparent noise show little to no enrichment in any values over expected values. Thus, the connectivity curves have characteristic traits for different structures and can indicate the type of structure present in the cluster (bright, dim, noise).

Crossword uses the geometric location of structured pixels within each image to identify artifact (Figure 4D). Each image channel is iteratively clustered for three groups of pixels using

dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362–371

the pixel intensity and those of its neighbors. The enrichment of the connectivity value for each cluster and the combination of the top and middle cluster is then determined. The algorithm progresses from the brightest cluster to the dimmest cluster, classifying each as noise, diffuse structure, or punctate structure based on the connectivity enrichment (Supporting Information Figure 6). If a structure is found in a cluster, the pixels are labeled as the appropriate type of structure and removed from the analysis. The remaining pixels are reclustered until no cluster has enrichment values significantly different from noise. Once all channels have been analyzed, the structure in each channel is compared to the previously identified foreground (Supporting Information Figure 7). Punctate structure that occurs in multiple analyte channels and does not match the foreground area is labeled as shared punctate artifact and removed from further analyses. Several geometric criteria are then used to compare the remaining structure to the shape of the foreground area. If the structure covers a defined amount of the foreground pixels and does not cover more than a defined amount of background pixels, the structure covering the foreground area is retained in the image and the structured background pixels are defined as artifact and excluded from further analysis. If the number of structured background pixels surpasses a set threshold, all structured pixels are defined as artifact. The exact levels for these cutoffs that are implemented in Crossword are given in Supporting Information Figure 7; these thresholds can be adjusted, however, depending on typical array performance. After removing structured artifacts, a ring of two pixels around the foreground area is also discarded to ensure no transition pixels are included in the background pixels. Once the foreground and background pixels are set, pixel intensities and other relevant metrics are extracted for each feature and exported in tabular form (Supporting Information Table 3) along with a file containing each feature subimage for downstream analysis.

We compared the utility of analyzing by connectivity value enrichment to previous clustering methods for artifact identification. We initially examined the effectiveness of the algorithm at analyzing 20 images extracted from real array images containing only salt-and-pepper noise as determined by manual review. Analysis by connectivity enrichment identified all 20 images as noise, preventing the separation of the pixels into separate clusters (Figure 5A, data not shown). In contrast, using minimization of the Bayesian Information Criterion (BIC) to select the number of models for Gaussian mixture-model (GMM)-based clustering separated all 20 images into 3 distinct clusters (Figure 5A). We also examined the efficiency of GMM and connectivity enrichment analysis in removing all structured pixels from images. While both approaches successfully extracted all structured pixels from images containing features with moderately strong signals (data not shown), connectivity enrichment displayed an improved ability in removing signal spread from bright features (Figure 5B). The connectivity analysis lowers both the mean and standard deviation of the background. Since signal-to-noise ratios are decreased both by higher background means and higher background standard deviations (eq 1), improvements in artifact removal can have significant effects on the calculated SNR value.

$$\mathrm{SNR} = \frac{\mu\mathrm{F} - \mu\mathrm{B}}{\sigma\mathrm{B}} \quad (1)$$

where μF is the mean of the foreground pixel intensity, μB is the mean of the background pixel intensity, and σB is the standard deviation of the background pixel intensity.

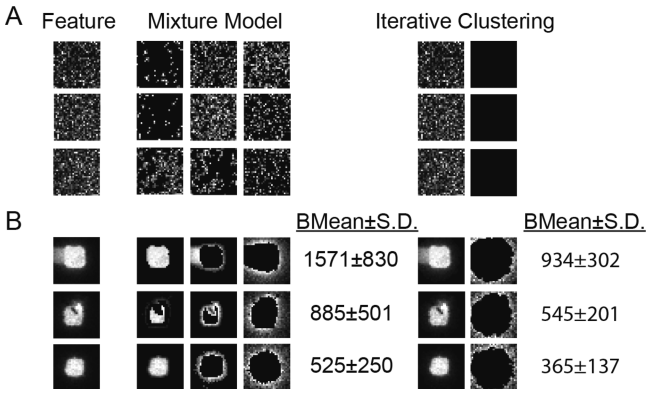


Figure 5. Comparison of Gaussian mixture models and connectivity enrichment clustering. (A) Twenty feature images containing only background noise were randomly selected from an array image and subjected to clustering using BIC minimization to select the optimal number of GMMs or connectivity enrichment to define the optimal number of clustering cycles. The resulting clusters for three representative examples are displayed. (B) Feature images containing bright foreground area were extracted from a microarray image and clustered using GMMs or connectivity enrichment. The mean and standard deviation of the pixels in the cluster with the lowest mean intensity are displayed for each image.

RESULTS

In order to analyze the quality of the data produced by our algorithm, Crossword and the commercial software GenePix Pro were used to analyze the same images of three protein microarrays containing different levels of artifact. Manual review of each feature for a positive signal in each analyte channel was used as the "gold standard" for identifying positive features. To analyze the images with Crossword, each image file path was passed to the algorithm and the data for all features was automatically exported with no user intervention. Data was extracted from the same images with GenePix Pro after manual adjustment of the locations of blocks. To examine the quality of the data produced by the two extraction methods, we compared the SNR and median-background values for each feature. Lack of correlation between mean and median intensity values of features has previously been used to identify features containing artifacts.[18] SNR values demonstrate even greater disparity from median values than mean values if artifact is present in the background due to the inverse relationship between SNR and the standard deviation of the background (eq 1), thereby making a lack of correlation between SNR and median values a sensitive indicator of the presence of artifact. Indeed, the SNR values and the background-subtracted median values extracted by Crossword demonstrate a much stronger correlation than the data extracted by GenePix Pro (Figure 6A, Supporting Information Figure 8). The increase in data quality yielded a significant improvement in the true positive and false positive rates of SNR values extracted by Crossword, particularly for images of arrays that contain medium to high amounts of artifact (Figure 6B).

To further analyze the rate of false positives called by Crossword, the SNR values extracted from the high artifact array were grouped according to whether the corresponding nanowell was occupied during microengraving. Features

dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362−371

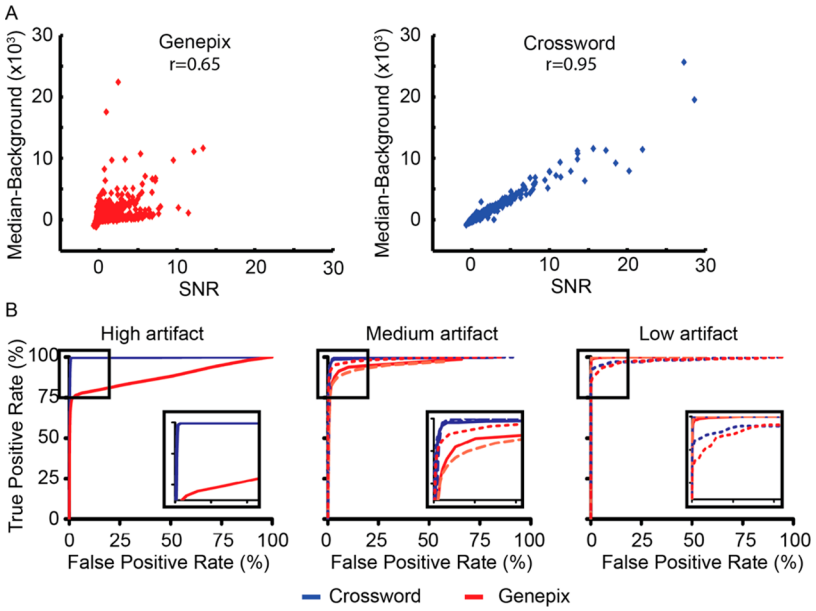


Figure 6. Comparison of extracted data generated by Crossword or GenePix Pro. Three images of microengraved arrays containing different levels of artifact were analyzed with Crossword and commercial software GenePix Pro. (A) Plots of the feature SNR and background-subtracted median intensity generated by GenePix Pro (left panel) or Crossword (right panel) from the high artifact array. (B) ROC curves generated by increasing the SNR cutoff using values generated by GenePix Pro (red) or Crossword (blue) from the three array images measuring CXCL8 (line, first panel), IFNγ (line, second and third panel), IL-2 (dashed line, second and third panel), and MIP-1β (dotted line, second and third panel).

corresponding to unoccupied nanowells represent the distribution of negative events. When the SNR values from nanowells containing cells were compared to the unoccupied wells, there was a significant enrichment in features with high SNR values, corresponding to measurable events of secreted analytes (Supporting Information Figure 9). Interestingly, the features defined as false positive events by the receiver operating characteristic (ROC) curve in the first panel of Figure 6B (according to the manually curated data set) are also enriched in occupied wells, suggesting many of these events classified as false positive may in fact be true positive events not readily recognized by visual inspection due to low signal. Therefore, the rate of false positives may be significantly less than the estimate based on manual annotation. These results suggest Crossword may increase sensitivity to low signal events on microengraved arrays compared to manual annotation of the images.

DISCUSSION

One motivation for developing Crossword was to create a fully automated software package that could extract accurate data from images of microengraved protein arrays irrespective of spatial distortions in the array and aberrations in signal common in experimentally produced arrays. The major approach used to improve the robustness of the algorithm was to create multiple redundancies in the gridding process. The location of each block is interrogated five times by the algorithm, and each feature relies on three separate techniques to assign the most accurate location. The ability to effectively combine information provided by multiple approaches through appropriate hierarchies, and numerical metrics describing the quality of the aligned grid yields accurate registration and definition of features. Overall, Crossword reduces a process for extracting data that often takes 1−2 h of manual adjustment and annotation per array using existing software such as GenePix Pro to 5 min of user time, thereby greatly expanding the number of arrays that can be feasibly analyzed by a single user.

We have used Crossword to analyze hundreds of microarrays over the past year, with a failure rate of roughly 5−10% of microengraved arrays. Typically, Crossword fails when registering blocks using the geometric code due to degraded resolution of the individual elements in the registration channel image that prevents distinguishing diamonds from squares, even with the human eye (Supporting Information Figure 10). If Crossword deciphers the codes for ten or fewer blocks correctly, the software returns a message to the command line stating that it was unable to properly register the array and then moves on to the next array in the processing list in order to continue batch processes. (If a user wants to extract information from these degraded arrays, the software also encodes a module for manual gridding. After manual alignment, artifact removal proceeds as in the automated process using the manually applied grid.)

Crossword was developed, optimized, and implemented for microengraved protein microarrays, but the algorithms described here should also be applicable to protein or DNA microarrays produced using conventional technologies. Although the geometric decoding of block identity relies on a unique feature arrangement of the microengraving method, spotted arrays could employ a very similar approach by encoding block identity by spotting fluorescently labeled probes or blank features in defined locations. Also, the automated analysis of microengraved arrays needs block decoding because the small block sizes make it possible to lose whole rows of blocks and the absolute location of an individual block on the glass slide varies significantly due to imprecise alignment of the slide with the array of nanowells during microengraving. For most spotted arrays, however, the absolute location of a block on the glass slide, particularly for arrays with larger block structures, should enable the identification of individual blocks without the need for block decoding. The remaining aspects of

dx.doi.org/10.1021/pr401167h | J. Proteome Res. 2014, 13, 362−371

Crossword could be used to identify the features within blocks and effectively remove artifact from measurements.

One disadvantage of Crossword for automated grid alignment is that it is not currently conducive to the use of hexagonally close-packed arrays due to the use of orthogonal 1-D projections. Crossword's algorithms are also not ideal for handling donut-shaped artifacts produced by some array platforms, as the center will likely still be considered structured. If these artifacts are commonly produced by a platform, a hybrid approach would likely be ideal, using connectivity enrichment to identify all structured pixels and one of several other described approaches[4,8,13] to remove the center of the donut from the foreground measure.

The availability of cheap computing power has enabled us to contemplate ways to improve the quality of the data extracted from the arrays. Fully automating the image analysis in itself improves data quality by removing user variability from the analysis. The establishment of a flexible framework for identifying artifactual pixels within the image should also substantially improve the quality of data. One powerful approach to identify certain artifacts, particularly in protein arrays, is the comparison of signal across different channels of data. Dust and other debris often cause bright fluorescent signals that are present in most, if not all, channels of a multichannel array. This class of artifact in microengraved arrays can be readily identified by correlated connectivity enrichment patterns present in multiple channels, allowing the algorithm to confidently remove these pixels from the analysis. Inclusion of a scanned channel with no corresponding analyte would make this approach even more robust, as it would eliminate any discrepancies when the artifact closely matches the foreground shape. This modification would also make this metric for quality control useful for arrays containing a significant signal in all image channels (such as DNA microarrays). The segregation of the remaining structured pixels into signals arising from artifacts and analytes is accomplished by a set of criteria for geometric comparisons established predominately by empirical testing. These criteria could be further refined using machine learning algorithms to systematically optimize them. The simple yet accurate description of the image structure by nine connectivity enrichment values combined with pixel location information should enable easy adaptation of computer learning algorithms to improve enumeration of artifacts.

In its current implementation, Crossword requires considerable computer processing time, requiring roughly 40 min to complete processing of a 12,000 × 4,000 pixel image containing four channels. Although this time is considerably longer than the time required for programs such as GenePix Pro (~2−5 min), it is still less than the total time required for processing images of microengraved arrays (1−2 h per array if manual alignment is required). The removal of artifact is the most computationally intensive portion. This module is encoded to allow massive parallel processing if sufficient numbers of compute cores are available (up to 10× more cores than we currently use). To take full advantage of this capability, we are currently refining the software implementing Crossword to use graphical processing units (GPUs), which have thousands of smaller cores designed for massively parallel processing. We are also examining whether iterative clustering can operate effectively at the block level instead of the feature level, thereby greatly reducing the necessary time for computation. Increasing the processing efficiency of the algorithm should enable even more detailed analysis of pixel structure as well as integration of methods for computer learning to improve artifact removal further. Intensive analysis of raw images prior to data extraction through the use of newly developed capabilities for parallel computational processing, such as implemented here in Crossword, represents a major opportunity for improving the quality of data culled from images of microarrays or other image-intensive biomolecular assays such as next-generation sequencing.

■ ASSOCIATED CONTENT

Supporting Information

Supplemental figures and supplemental methods. This material is available free of charge via the Internet at http://pubs.acs.org. The software implementing Crossword is available upon request from the MIT Technology Licensing Office for academic research.

■ AUTHOR INFORMATION

Corresponding Author

*J. Christopher Love, Ph.D., Department of Chemical Engineering, Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 77 Massachusetts Ave., Bldg. 76-253, Cambridge, MA 02139. Phone: 617-324-2300. E-mail: clove@mit.edu.

Author Contributions

T.M.G. conceived and wrote most of the computer code for both the segmentation and artifact removal modules. D.L. established protocols for distributing the software to multiple computing cores, enabling the timely completion of the algorithm. J.C.L. provided expert advice on the analysis of microengraved arrays. All authors wrote and edited the manuscript.

Funding

This work was funded in part by grants from the Ragon Institute of MGH, MIT, and Harvard, the National Institute of Allergy and Infectious Diseases (Grants 1R56AI104274, 1U19AI089992, and 5U01AI068618), and the W.M. Keck Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy And Infectious Diseases or the National Institutes of Health.

Notes

The authors declare the following competing financial interest(s): Dr. J. Christopher Love is a founder, shareholder, and consultant for Enumeral Biomedical. Dr. Todd Gierahn and Denis Loginov declare no competing interests.

■ ACKNOWLEDGMENTS

We thank Viktor Adalsteinsson for assistance producing the microengraved array and helpful discussion and Rita Contento and Yvonne Yamanaka for assistance testing the algorithm on a large set of microengraved arrays. J.C.L. is a Camille Dreyfus Teacher-Scholar. The authors dedicate this paper to the memory of Officer Sean Collier, for his caring service to the MIT community and for his sacrifice.

■ ABBREVIATIONS

SVM, support vector machine; SNR, signal-to-noise ratio; CV, coefficient of variance; IgG, immunoglobulin G; CXCL1,

chemokine (C-X-C motif) ligand 1; CXCL8, chemokine (C-X-C motif) ligand 8; IFNγ, interferon-gamma; IL-2, interleukin-2; Mip1β, macrophage inflammatory protein-1beta; ROC, receiver operating characteristic; GMM, Gaussian mixture model

REFERENCES

(1) Molecular Devices. *GenePix Pro*; http://www.moleculardevices.com/Products/Software/GenePix-Pro.html.

(2) Biodiscovery, I. *ImaGene*; http://www.biodiscovery.com/software/imagene/.

(3) Eisen, M. B. *Scanalyze*; http://rana.lbl.gov/EisenSoftware.htm; 1999.

(4) Li, Q.; Fraley, C.; Bumgarner, R. E.; Yeung, K. Y.; Raftery, A. E. Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics* **2005**, *21* (12), 2875−82.

(5) Yang, Y. H.; Buckley, M. J.; Dudoit, S.; Speed, T. P. Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graphical Stat.* **2002**, *11* (1), 108−136.

(6) Bajcsy, P. An Overview of DNA Microarray Grid Alignment and Foreground Separation Approaches. *EURASIP J. Appl. Signal Process.* **2006**, *2006*, 1−13.

(7) Angulo, J.; Serra, J. Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics* **2003**, *19* (5), 553−62.

(8) Blekas, K.; Galatsanos, N. P.; Likas, A.; Lagaris, I. E. Mixture model analysis of DNA microarray images. *IEEE Trans. Med. Imaging* **2005**, *24* (7), 901−9.

(9) Rahnenführer, J.; Bozinov, D. Hybrid clustering for microarray image analysis combining intensity and shape features. *BMC Bioinformatics* **2004**, *5*, 47.

(10) Wu, S.; Yan, H. Microarray Image Processing Based on Clustering and Morphological Analysis. *Proc. First Asia Pacific Bioinf. Conf.* **2003**, 111−118.

(11) Giannakeas, N.; Fotiadis, D. I. An automated method for gridding and clustering-based segmentation of cDNA microarray images. *Comput. Med. Imaging Graph.* **2009**, *33* (1), 40−9.

(12) Bariamis, D.; Maroulis, D.; Iakovidis, D. K. Unsupervised SVM-based gridding for DNA microarray images. *Comput. Med. Imaging Graph.* **2010**, *34* (6), 418−25.

(13) Gottardo, R.; Besag, J.; Stephens, M.; Murua, A. Probabilistic segmentation and intensity estimation for microarray images. *Biostatistics* **2006**, *7* (1), 85−99.

(14) Wu, E.; Su, Y. A.; Billings, E.; Brooks, B. R.; Wu, X. Automatic Spot Identification for High Throughput Microarray Analysis. *J. Bioeng. Biomed. Sci.* **2012**, *S3*, 002.

(15) Rueda, L.; Rezaeian, I. A fully automatic gridding method for cDNA microarray images. *BMC Bioinf.* **2011**, *12*, 113.

(16) Wang, X.; Ghosh, S.; Guo, S. W. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.* **2001**, *29* (15), E75−5.

(17) Sauer, U.; Preininger, C.; Hany-Schmatzberger, R. Quick and simple: quality control of microarray data. *Bioinformatics* **2005**, *21* (8), 1572−8.

(18) Tran, P. H.; Peiffer, D. A.; Shin, Y.; Meek, L. M.; Brody, J. P.; Cho, K. W. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res.* **2002**, *30* (12), e54.

(19) Petrov, A.; Shams, S. Microarray Image Processing and Quality Control. *J. VLSI Signal Process.* **2004**, *38*, 211−226.

(20) Ogunniyi, A. O.; Story, C. M.; Papa, E.; Guillen, E.; Love, J. C. Screening individual hybridomas by microengraving to discover monoclonal antibodies. *Nat. Protoc.* **2009**, *4* (5), 767−82.

(21) Sendra, V. G.; Lie, A.; Romain, G.; Agarwal, S. K.; Varadarajan, N. Detection and isolation of auto-reactive human antibodies from primary B cells. *Methods* **2013**, *64* (2), 153−9.

(22) Park, S.; Han, J.; Kim, W.; Lee, G. M.; Kim, H. S. Rapid selection of single cells with high antibody production rates by microwell array. *J. Biotechnol.* **2011**, *156* (3), 197−202.

(23) Park, S.; Kim, W.; Kim, Y.; Son, Y. D.; Lee, S. C.; Kim, E.; Kim, S. H.; Kim, J. H.; Kim, H. S. Array-based analysis of secreted glycoproteins for rapid selection of a single cell producing a glycoprotein with desired glycosylation. *Anal. Chem.* **2010**, *82* (13), 5830−7.

(24) Panagiotou, V.; Love, K. R.; Jiang, B.; Nett, J.; Stadheim, T.; Love, J. C. Generation and screening of Pichia pastoris strains with enhanced protein production by use of microengraving. *Appl. Environ. Microbiol.* **2011**, *77* (9), 3154−6.

(25) Han, Q.; Bagheri, N.; Bradshaw, E. M.; Hafler, D. A.; Lauffenburger, D. A.; Love, J. C. Polyfunctional responses by human T cells result from sequential release of cytokines. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (5), 1607−12.

(26) Varadarajan, N.; Kwon, D. S.; Law, K. M.; Ogunniyi, A. O.; Anahtar, M. N.; Richter, J. M.; Walker, B. D.; Love, J. C. Rapid, efficient functional characterization and recovery of HIV-specific human CD8+ T cells using microengraving. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (10), 3885−90.

(27) Nagesh, A. S.; Varma, G. P. S.; Govardhan, A.; Babu, B. R. An Enhanced Fuzzy C-Means Clustering (ECFMC) Algorithm for Spot Segmentation. *Commun. Comput. Inf. Sci.* **2011**, *260*, 320−327.