Fully-Automated Analysis of Multi-Resolution Four-Channel Microarray Genotyping Data

Mohsen Abbaspour[a], Rafeef Abugharbieh[a], Mohua Podder[b], and Scott J. Tebbutt[c]

[a]Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada;
[b]Department of Statistics, University of British Columbia, Vancouver, BC, V6T 1Z2, Canada;[*]
[c]iCAPTURE Centre (Department of Medicine), St. Pauls's Hospital, University of British Columbia, Vancouver, BC, V5Z 4E3, Canada

ABSTRACT

We present a fully-automated and robust microarray image analysis system for handling multi-resolution images (down to 3-micron with sizes up to 80 MBs per channel). The system is developed to provide rapid and accurate data extraction for our recently developed microarray analysis and quality control tool (SNP Chart). Currently available commercial microarray image analysis applications are inefficient, due to the considerable user interaction typically required. Four-channel DNA microarray technology is a robust and accurate tool for determining genotypes of multiple genetic markers in individuals. It plays an important role in the state of the art trend where traditional medical treatments are to be replaced by personalized genetic medicine, i.e. individualized therapy based on the patient's genetic heritage. However, fast, robust, and precise image processing tools are required for the prospective practical use of microarray-based genetic testing for predicting disease susceptibilities and drug effects in clinical practice, which require a turn-around timeline compatible with clinical decision-making. In this paper we have developed a fully-automated image analysis platform for the rapid investigation of hundreds of genetic variations across multiple genes. Validation tests indicate very high accuracy levels for genotyping results. Our method achieves a significant reduction in analysis time, from several hours to just a few minutes, and is completely automated requiring no manual interaction or guidance.

Keywords: Segmentation, Mathematical morphology, Four-channel microarrays, Genotyping

1. PURPOSE

After the completion of the Human Genome Project, researchers have begun working to identify the effects of interaction between human genetic variation and the environment on different patients' responses to diseases and medical therapies. The most common type of genetic variation between individuals is single nucleotide polymorphisms (SNPs)[1], which are single base changes at specific DNA sites in the genome. Increasing experimental evidence suggests that different combinations of SNPs, in association with environmental factors, determine an individual's risk for complex genetic disease and responsiveness to medical interventions and drug toxicity[2]. To apply this knowledge to medical conditions such as sepsis[3] requires the rapid determination of multiple SNPs (genotypes) in an individual patient for more accurate prediction of the risk of serious outcomes, including death. Sepsis is a complex genetic disorder that involves an exaggerated inflammatory response throughout the body, following what would normally be benign infections or other environmental triggers. Severe cases of sepsis number over 800,000/year in North America alone and 250,000 of these people will die, often within just a few days of admission to an intensive care unit. Researchers are discovering that certain SNPs within genes involved in inflammatory and blood coagulation pathways are correlated to differences in severity and outcomes between different patients with sepsis[3]. Microarray-based genotyping strategies are potentially well suited to provide rapid multiple SNP genotyping for genotype-based prospective clinical practice, with a diagnostic turn-around timeline compatible with clinical decision-making [4, 5].

[*] Further author information: (Send correspondence to Rafeef Abugharbieh)
Mohsen Abbaspour: E-mail: mohsen a@ece.ubc.ca
Rafeef Abugharbieh: E-mail: rafeef@ece.ubc.ca
Mohua Podder: E-mail: MPodder@mrl.ubc.ca
Scott J. Tebbutt: E-mail: STebbutt@mrl.ubc.ca

Medical Imaging 2006: Image Processing, edited by Joseph M. Reinhardt, Josien P. W. Pluim,
Proc. of SPIE Vol. 6144, 61443M, (2006) · 0277-786X/06/$15 · doi: 10.1117/12.650814

An important stage in every microarray experiment is image processing. However, available microarray image analysis methods are inefficient and inaccurate due to manual interactions required. In this paper, we present an image analysis system for fully-automated gridding, segmentation and data extraction of four channel multi resolution images, based on morphological operations, matched filtering and watershed segmentation. Our results are compared to those of a commercially available microarray image analysis package[6].

2. METHODS

2.1. Data Acquisition and Pre-processing

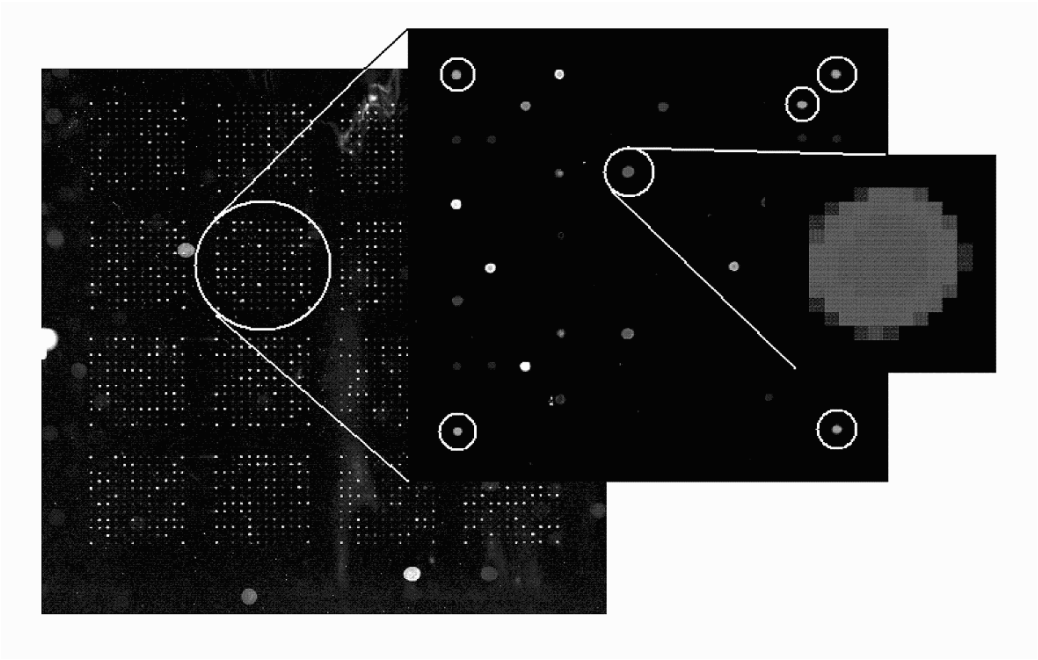
Among several microarray genotyping protocols, we developed our platform based on the Arrayed Primer Extension (APEX) technology[7]. APEX probe oligonucleotides were printed onto specific grid positions on microarray slides. The arrays were imaged with a biochip reader, fitted with filter sets that allow four channels of data to be collected for each sample (one channel for each of the four bases of DNA)[4]. The average theoretical spot diameter d varies for different imaging resolutions from about 10 pixels for 10-micron resolution to 33 pixels for 3-micron resolution scanning, and the size of each channel (TIFF image) varies from 8 MBs for low resolution data, to about 80 MBs for that of high-resolution. Each of the four TIFFs consists of an array of spots arranged in rectangular groups called subgrids, which usually have the same number of rows and columns of spots, and are arranged in relatively equal spacing, forming a "meta-array" image. Multiple positive control spots scattered across the array grid are used for normalization purposes. Figure 1.a shows the general features of a microarray image, obtained by adding the four channels. Positive control spots of a subgrid are indicated by circles around them.

Among two approaches of using the sum, or union of microarray channel images for gridding and segmentation, the former has a higher SNR. This is because although many of the noisy pixels are saturated in each channel, the sum of most signal pixels is smaller than the saturation limit. In other words, summation increases signal level, without making any changes in noise level. Since the channels contain different background levels, caused by the varying response of fluorescent dyes to different excitations, before adding the four TIFF images, we apply a simple morphological background subtraction8 to each of them, using an opening filter by a disk-shaped operating element with a diameter of *2d*, where *d*, the spot diameter, is the only input parameter of the system. Unless otherwise stated, in the rest of this section, we work with the sum of the background subtracted images.

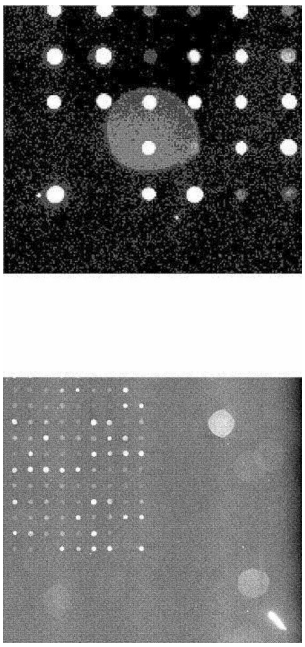
2.2. Separating Sub-Grids

Most of the currently used gridding algorithms are based on manual interaction[9]. We designed a fully-automated gridding algorithm based on the use of horizontal and vertical projection profiles (defined below) of the image. Our algorithm divides the grid into subgrids, and the subgrids into individual spots. We then apply our segmentation technique (described later) to each cell separately, making the automated analysis very suited for parallel processing and real-time operation, which is our ultimate goal. Other works were reported that utilize projection profiles[10], however, most of them need manual adjustment of many parameters, in addition to being sensitive to high noise levels, which we overcome by applying robust noise removal steps. Several noise types may affect the process described. Two types of noise are of special concern: *blobs* and *scratches*, illustrated in Figure 1.b. The background subtraction method applied removes the blobs to some extent, but not completely. In order to remove the noise effects, two noise removal strategies are implemented. We remove the effect of blobs by applying a "closing followed by opening" morphological filter, with a disk operator of size 2d. The scratch effects are removed with the $1 \times d$ morphological "closing followed by opening" filter applied in four directions. Before computing the projection profiles, a contrast adjustment is applied to the image.

The horizontal projection profile of an arbitrary image is defined as $P_{h,m} = \sum_{n=1}^{N} f(m,n), \quad (m = 1,...,M)$, in which *M*, *N* are the length and width of the image in pixels, and *f(m,n)* is the pixel intensity at the coordinates m and n (similarly for the vertical projection profile, $P_{v,n}$). The first step in finding the gridding information is to fill the valleys of the image, i.e. the low-intensity pixels, so that we can obtain good projection profiles. We do this task by boosting low-intensity pixels to a certain threshold. A proper threshold was found to be the mean value of e entire image pixel $h = \overline{f(x,y)}$. Figure 2.a shows the horizontal projection profile obtained from the image in Figure 1.a without blob removal. The blob effect is observable in the figure. The low values of this profile correspond to the non-data areas, and

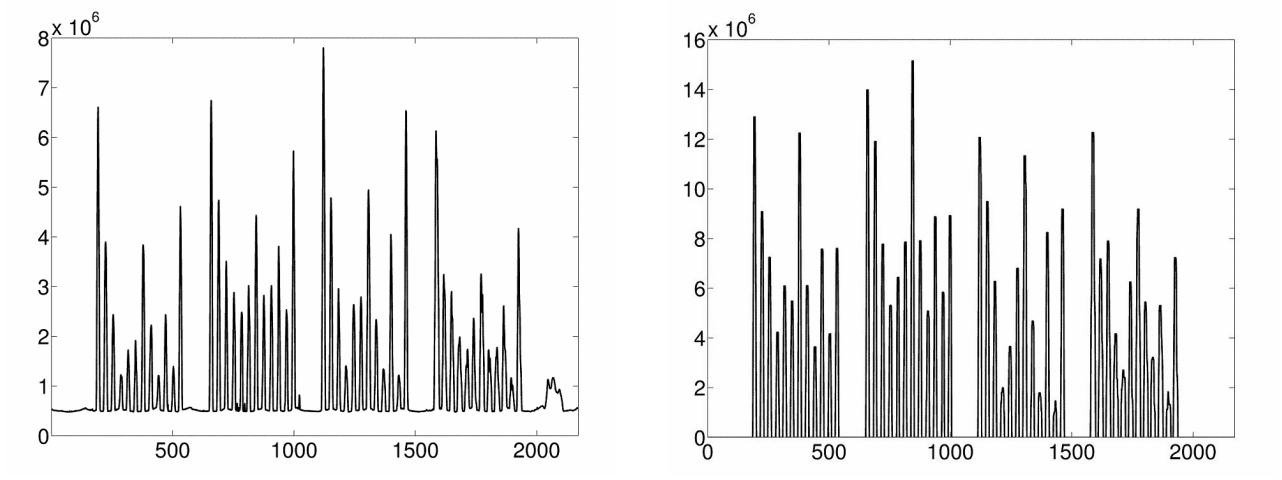


(a)



(b)

Figure 1. (a) Illustration of the sum of the channels for a typical microarray. One of the subgrids is presented in larger scale. (b) Top: A blob is visible as a big circle in the middle of the image. Bottom: A scratch is visible on the lower-right corner. The image is contrast adjusted for clarity.



(a) (b)

Figure 2. (a) Horizontal projection profile of Figure 1.a. Horizontal axis shows image row number, while the vertical axis represents the sum of N 16-bit intensity values of the row. The blob effect is visible as the peak to the right of the 2000 mark. (b) The profile obtained in section 2.2. Zero crossings are used for gridding.

the high level points to the subgrids. In order to detect the data area points, we set the low values of this profile to zero, for which we define a threshold equal to half of the profile average, $h = \overline{P_{h,m}} / 2$. Although the morphological filters reduce the effect of scratches considerably, there may still remain some narrow undesired peaks in the projection profiles, which are removed by a one dimensional median filter of length d. The resulting profile is presented in Figure 2.b. We find the gridding information from the zero-crossings function of this profile, which contains big steps at subgrid edges.

2.3. Spot Gridding and Watershed

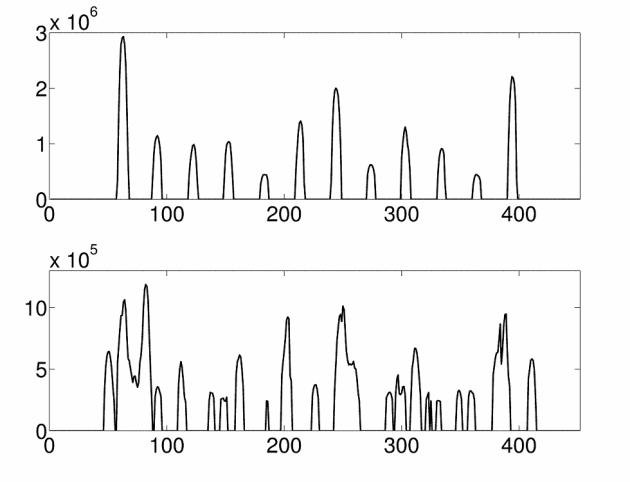
In microarray image analysis literature, seeded region growing is commonly used for spot detection[8]. However, there are many problems with the flow, i.e. how to stop the growing, which makes this method very non-robust and sensitive to input parameters. Other reported approaches include the use of Active Contour Models[11] and the Wavelet Transform[12], however, the reported results were still prone to errors due to noise. The segmentation method we use here is the watershed[13]. We incorporate a number of morphological operations and matched filtering to build a robust method for finding the watershed lines as well as the starting points. The background subtraction and noise removal steps of section 2.2 are necessary for accurate segmentation as well.

Our watershed-based algorithm receives as input two images. The first input image is the morphological gradient $\nabla_b = \delta_b - \varepsilon_b$ of the image, in which b is a morphological operator (of size 3×3 in our case), and δ_b, ε_b are the morphological *dilation* and *erosion* operations respectively. This gradient operator performs enhancement of the edges[14]. The second input is a marker image composed of the centers of the spots and the watershed lines. We used a morphological method for finding the watershed lines, obtained from the spot gridding information, and matched filtering to find the points at which the watershed starts, instead of using marker cell midpoints, because of microarray printing inaccuracies. We implement matched filtering on the marker cells instead of the entire image to ease and automate the process of finding a proper matched filter, as well as making it more suitable for real-time processing.

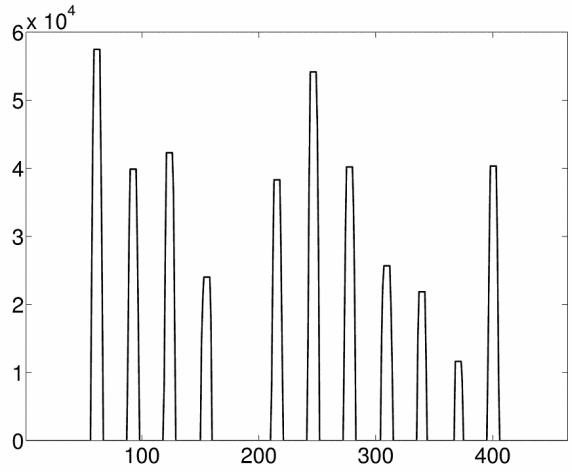
In most microarray image analysis systems, a manual rotation estimation step is performed before finding the grid information. We developed an automated rotation estimation algorithm based on the fact that rotation of a subgrid expands its projection profiles (Figure 3.a). However, no rotation was observed in our data samples.

The projection profiles are again used for extracting subgrid information. The profiles are obtained from the contrast adjusted version of the image. We then apply a one dimensional morphological opening of size d[8] to the profiles, to remove the profile background. Now all low values of the profiles are suppressed. The threshold used here is the mean value of the profile. Finally, a one dimensional median filter is used to remove any remaining noise effects. The resulting

profile is shown in Figure 3.b. Gridding points are obtained as the midpoints of the zero crossings in this profile. If we call this zero crossing function $I = \{i_1, i_2, i_3, ..., i_X\}$, the difference function of I defined by $DI = \{d_1, d_2, ..., d_{X-1}\}$, $d_1 = (i_2 - i_1)$, ..., $d_{X-1} = (i_X - i_{X-1})$ shows the length of intervals corresponding to spot and non-spot areas. $d_k \in DI$ are supposed to be almost equal, because of the almost-equal spacing arrangement of the spots. Hence, we should apply a compensation algorithm if there exists any k for which d_k is much different than the median of DI, as the case in Figure 3.b. In other words, we detect that compensation is required if $\exists\ k \in DI\ dk > \alpha \times median(DI)$, in which α must be chosen between *1* and *2*, since d_k either corresponds to correct values, which are close to the median, or to areas with missed information, which are at least about twice the value of the median. We set α to *1.5* in our implementation. If compensation is required, we define the new gridding points as follows: we first obtain the corners of the sub-grid by simply thresholding the profiles by a threshold equal to their average value (the resulting profile is similar to figure 3.a, which gives the corner coordinates easily). We then define a new set $DI_m = \{d_k \in DI\ \&\ d_k \leq \alpha \times median(DI)\}$, and divide the distance between the corners of profile to equal intervals of size $mean(DI_m)$. For high-resolution data, the entire process is done on sub-sampled data for efficient computation in terms of time and memory. Robustness of the gridding algorithm was successfully tested on several images of different spacing values and different resolutions. Results were also consistent with the variation of parameters and thresholds used.

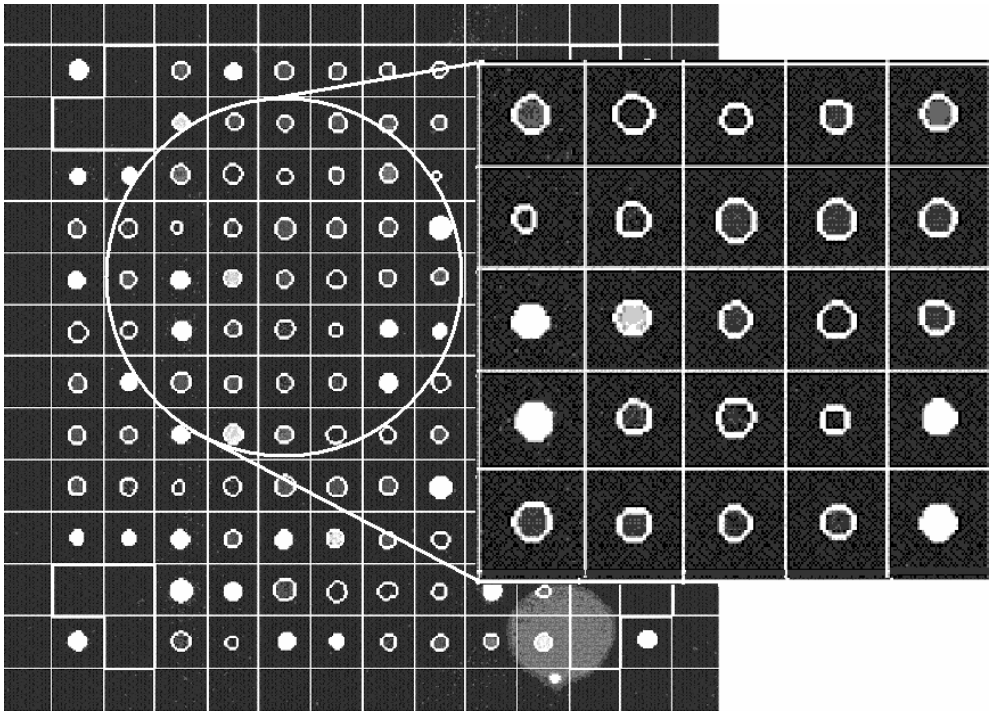


(a)



(b)

Figure 3. (a) Top: profile of a non-rotated subgrid. Bottom: profile of a rotated subgrid (expanded). (b) Profile of a subgrid after applying filters and other operators. Zero crossings are used for gridding purposes. A missing area (wide zero intervals) is visible before the 200 mark on the horizontal axis, which needs gridding compensation.



(a)

(b)

Figure 4. (a) Result of watershed segmentation on a subgrid. (b) Top: Effect of noise on watershed. Bottom: Noise effect removal by opening filter.

After finding the watershed lines (which correspond to the grid lines), the next task is to identify a proper matched filter (used to initialize the watershed in each spot), which could be one of the bright smooth control spots of the subgrid. However, we use the median of the five control spots to avoid the effect of outliers.

Watershed works on the images spot by spot, making it proper for parallel processing. Figure 4.a shows the result of segmentation together with the marker image for one subgrid. Some filtering precedes feeding of an image to the watershed segmentation algorithm, since there exist small bright noisy spots within the image, which sometimes trap the watershed. To solve the problem, we apply an opening filter to remove the noise before finding the gradient. Since the spot shaped noise diameters appearing in our images are typically about d/5, we set the opening filter's size to this value. Figure 4.b shows the advantage of this noise compensation. The results of segmentation accuracy are summarized in the next section.

3. RESULTS

The medians of pixel intensities in each spot is used for quantification. The use of median instead of mean prevents the outliers (i.e., saturated noisy pixels) to affect decision-making. Following spot segmentation, a background subtraction and normalization step is required before quantitative analysis can be performed. We used the region method for background subtraction[15], however, we define the background as the minimum value of the background area (spot surrounding pixels) instead of using the median, to remove negative estimates that appear for low-intensity spots. The results are finally normalized by either of two methods: Method 1, according to the entire control spot sets in the image; Method 2, according to the control spots in each subgrid. Results of the latter are slightly better.

Twenty different human DNA samples were genotyped on separate but identical arrays. Each slide was imaged at 10 micron resolution. Several weeks later, the arrays were re-imaged at 3.25, 6.5, and 9.75 micron resolution. The low-resolution TIFF files were imported into either the commercial software6 our system was previously using, or our own image analysis system. The gene calling algorithms (unpublished) produce either genotypes or non-calls. Validated genotypes were available in public databases for 109 SNPs across the majority of the 20 human DNA samples. The scored genotypes were compared with the validated "true" genotypes.

Both call rate and accuracy results confirm that our fully-automated method is comparable with commercial software, however our approach is much faster and is completely automated. The results of multi-resolution scanning are not compared with those of the original 10-micron analysis, because the multi-resolution images were scanned from the microarray sample slides several weeks after the original images, and hence suffered somewhat from fluorescent dye degradation over time. Although the call rates and accuracy rates are equivalent for the different resolution scans, we believe that the 3 micron images will prove to be more accurate when future iterations of image analysis algorithms are developed, simply because greater numbers of pixels are available for analysis.

Table 1. Genotyping results. First column compares the commercial software with our algorithm. Second column shows the multi-resolution analysis results. (Call rate: Proportion of scored genotypes compared to the total number of validated genotypes. Accuracy: Concordance rate of scored genotypes and the "true" genotypes.)

	10 μ analysis			Multi-resolution analysis (our system only)		
	Software	Method 1	Method 2	9 μ	6 μ	3 μ
Call Rate	87.6%	86.3%	86.0%	82.9%	82.2%	82.2%
Accuracy	98.4%	98.8%	98.9%	98.0%	98.0%	98.4%

4. CONCLUSIONS

We presented an accurate and robust fully-automated microarray image analysis system, which improves the speed and accuracy of current analysis frameworks. Our analysis pipeline also enables us to handle mlti-resolution data of four-channel APEX microarrays. The presented image analysis methods can be used for the quick determination of genetic variation for prospective personalized genetic medicine applications. Our analysis system is generalized to work on both

two- and four-channel microarrays, in contrast to most currently available methods that are designed to analyze two-channel slides or individual channels of a slide only. In contrast with our method, most commercial analysis packages and state of the art methods need manual interaction or at least manual assignment of a number of parameters. Processing time is also reduced in our analysis pipeline. Further work in this area includes increasing the robustness of algorithm for severe image background changes and for low-intensity spots. Other quantifications of spot pixel information should be considered to seek improvements in gene calling, and to elucidate the value of high-resolution scanning. We are also currently working on parallel processing implementations for spot segmentation and quantification for the ultimate goal of achieving real-time genotyping.

REFERENCES

1. D. G. Wang, J. B. Fan, C. J. Siao, et al., "Large-scale identification, mapping, and genotyping of snps in the human genome," *Science* **280**, pp. 1077–1082, 1998.
2. N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science* **273**, pp. 1516–1517, 1996.
3. C. L. Holmes, J. A. Russell, and K. R. Walley, "Genetic polymorphisms in sepsis and septic shock: role in prognosis and potential for therapy," *Chest* **124**, pp. 1103–1115, 2003.
4. S. J. Tebbutt, J. Q. He, K. M. Burkett, J. Ruan, et al., "Microarray genotyping resource to determine population stratification in genetic association studies of complex disease," *Biotechniques* **37**, pp. 977–985, 2004.
5. S. J. Tebbutt, I. V. Opushnyev, B. W. Tripp, et al., "Snp chart: An integrated platform for visualization and interpretation of microarray genotyping data," *Bioinformatics* **21**, pp. 124–127, 2005.
6. M. Inc., "http://www.softworx.com/," 2002.
7. A. Kurg, N. Tonisson, I. Georgiou, et al., "Arrayed primer extension: solid-phase four-color dna resequencing and mutation detection technology," *Genet Test* **4**, pp. 1–7, 2000.
8. Y. H. Yang, S. Dudoit, et al., "Comparison of methods for image analysis on cdna microarray data," *Technical Report, Department of Statistics, University of California at Berkeley* **584**, 2000.
9. A. Petrov and S. Shams, "Microarray image processing and quality control," *Journal of VLSI Signal Processing* **38**, pp. 211–226, 2004.
10. R. J. Hirata, J. Barrera, R. F. Hashimoto, et al., "Segmentation of microarray images by mathematical morphology," *Real-Time Imaging* **8**, pp. 491–505, 2002.
11. M. Katzer and F. Kummert, "Methods for automatic microarray image segmentation," *IEEE Trans. Nanobioscience* **2**, pp. 202–214, 2003.
12. X. H. Wang and S. H. Istepanian, "Application of wavelet modulus maxima in microarray spot recognition," IEEE Trans. *Nanobioscience* **2**, pp. 190–192, 2003.
13. R. C. Gonzalez and R. E. Woods, Digital Image Processing, Prentice Hall, Boston, 2002.
14. J. Serra, Image Analysis and Mathematical Morphology, Academic Press, London, 1982.
15. A. Bengtsson, "Microarray image analysis: background estimation using region and filtering techniques," *Master's theses, Lund University* **E40**, 2003.