The Impact of Microarray Image Intensity Variations on SNP Genotype Calls

Ching-Yu Huang

***Abstract*—Microarray technology is widely used to determine single nucleotide polymorphism (SNP) genotype calls when studying the association between disease and human subjects. A genotype call is decided relatively with the signal intensity values of spots on the microarray image. Since high-throughput microarray technology can contain millions of SNP spots, it is impossible to manually inspect if the spots are aligned correctly and intensity values are calculated accurately. However, a wrong genotype call will skew the outputted numerical results, which can ultimately cause unsuitable medicine to be prescribed and jeopardize a person's life. Therefore, retrieving correct intensity values is very important. This paper analyzes and shows the effect signals have on the SNP genotype calls when the signals from the same spots are retrieved and represented in different values.**

***Index Terms*—Genotype, image processing, microarray, SNP.**

I. Introduction

A single nucleotide polymorphism (SNP), a variation at a single site of DNA, is the most frequent type of variation in the genome. There are around 50 million SNPs that have been identified in the human genome [1]. An SNP microarray, which consists of many spots, is used to detect polymorphisms within a population. Each spot intensity value represents the hybridization level of a single gene and a specific SNP oligo or probe. The array is a solid substrate (usually a glass slide or silicon thin-film cell) that assays large amounts of biological material for a modern high-throughput genotyping system.

SNP technology is commonly used to the study the association between humans and diseases. Therefore, a wrong genotype call will skew the statistical results and cause patients or doctors to make wrong decisions. Therefore, it is very important to derive accurate spots' intensity values and apply strict quality check rules.

An *N*-multiplex microarray means there are *N* SNPs on the array. The Affymetrix Genome-Wide Human SNP Array 6.0 features 1.8 million genetic markers, including more than 906,600 SNPs and more than 946,000 probes for the detection of copy number variations [2].

Most microarray technology is fluorescence-based and uses the laser to trigger a specific particle. The digital CCD camera then captures the light emitted from the triggered particle as an intensity value. The intensity depth depends on the number of CCD bits. Most of the microarray instruments use a camera with a range between 10 and 16 bits, which will generate intensities from 2^10 (1024) to 2^16 (65536). Lower number of bits will cause problems with detection of the reaction and overall saturation. The intensity value is influenced by a combination of laser power, exposure time, laser wave, hybridization, and environment. Any noise or residue on the array surface will have a great impact on the intensity. Some systems use one laser to generate different waves by rotating different filters, while others use several lasers to generate different waves. Different laser wavelengths will trigger different particles, which will cause differing results and noise. Even if the microarray is captured different times, the results will differ because slight deviations in the laser intensity will produce differing results.

High-throughput microarray technology heavily relies on a fully automatic system for image segmentation and quantification [3]. Most of the spot signal intensity values are retrieved using a square or circular grid to cover each spot [4]-[6]. The system automatically locates both subarray grids [7] and individual spots, requiring no user identification of any image coordinates [8]. Some systems might adjust for signal saturation at the segmentation stage that identifies a pixel as part of the foreground or background [9].

Background correction is also an important preprocess in microarray data analysis [10]. During the image processing, the grid is fixed in size for all spots to calculate the averaged intensity of each spot. Ideally, the spot size should be the same. However, the spot size could vary or be not evenly distributed as a result of different SNP oligo volumes, different concentrations, or hybridization factors of the container's surface. In this paper, we would like to discuss the situations that generate uneven signals and how the genotype calls will be affected.

II. Intensity Computing Method

Fig. 1 is a common gray-scale (range 0 - 65535) image of a Microarray with 6×6 spots. Every spot looks bright, so it is difficult to see the variation in their sizes and intensities. However, if we view it in 3D, as shown in Fig. 2, it is obvious that there is a lot of deviation between spots.

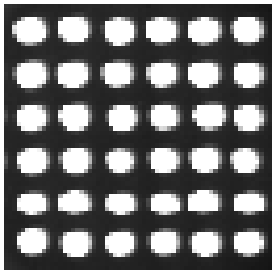

Fig. 1. A microarray image with 6×6 spots.

Manuscript received September 8, 2015; revised March 23, 2016.

Ching-Yu Huang is with the Department of Computer Science, Kean University, USA (e-mail: chuang@kean.edu).

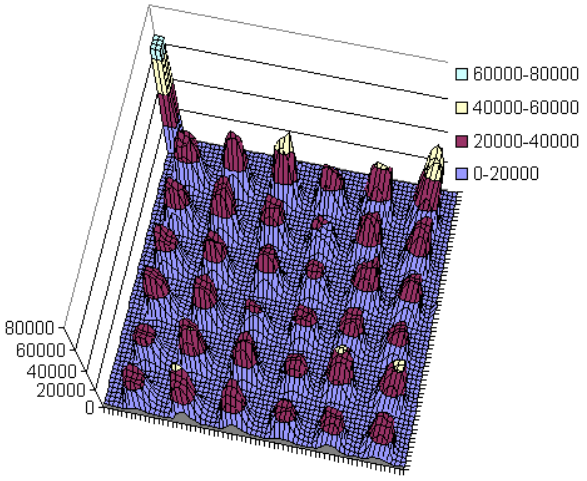

Fig. 2. 3-D view of Fig. 1.

In order to understand how the variations in the intensity affects the averaged value, a 10×10 area in the left upper corner was selected to analyze its intensity distribution. Every pixel's intensity in the 10×10 area is shown in Fig. 3, where the pixel in cell A1 represents the upper left corner of the spot. Table I shows the summary information of these 100 pixels' intensity values.

There are two ways to compute the average intensity value for a spot, which are detailed below.

A. *Fixed Area Method*

This method does not consider if a pixel is part of a "spot", and always uses a fixed area to compute the average once the starting or center position is determined. A 4×4 window is used to move around inside the 10×10 area to compute the average intensity in a raster scan from left to right and top to down. The starting position of the 4×4 window is cell A1, at the left top corner of Fig. 3. The area covered is from row A to D and column 1 to 4. All possible averaged intensities of are shown in Fig. 4.

	1	2	3	4	5	6	7	8	9	10
A	1412	1264	1156	1160	1168	1168	1200	1244	1208	1200
B	1356	1224	1184	1164	1244	1348	1460	1396	1360	1312
C	1280	1224	2232	6204	10420	11828	8636	5148	2820	1504
D	1236	3172	15836	25984	29628	29416	25740	12228	4956	2128
E	1204	7444	26552	32388	30212	28304	28520	19868	7284	2876
F	1220	6984	26804	30460	26560	26692	29344	25148	9752	3308
G	1264	3224	21172	27956	26688	29860	33792	25860	8612	2536
H	1264	1332	7568	20636	23372	23520	20000	11448	3796	1484
I	1256	1232	1352	3376	5496	5420	4208	2484	1508	1324
J	1244	1276	1172	1272	1344	1412	1428	1420	1388	1324

Fig. 3. The intensity values of the 10×10 pixel area surrounding the upper left spot of Fig. 2.

TABLE I: PIXEL INTENSITY SUMMARY OF FIG. 3

Intensity information	Value
Average (Avg)	9,341
Maximum (Max)	33,792
Minimum (Min)	1,156
Standard deviation (Std)	10,887
1st Quartile (Q1)	1,279
2nd Quartile (Q2)	2,848
3rd Quartile (Q3)	19,901

	1	2	3	4	5	6	7
A	4193	6517	8821	9861	8955	6947	4596
B	8105	12257	15872	17031	15337	11895	7952
C	11889	17632	22470	23771	21731	17230	11829
D	14556	21317	27157	28847	26741	21586	15122
E	13592	19960	25547	27394	25574	20738	14602
F	9819	14638	19183	21086	19993	16340	11538
G	6037	9279	12601	14361	13610	11010	7663

Fig. 4. The cell values indicate the average intensity values of 4×4 windows where each cell represents the starting positions on Fig. 3.

	1	2	3	4	5	6	7
A	6905	9307	10732	11252	10440	8915	6470
B	10801	12360	12661	12608	11749	10824	9010
C	12193	11877	9711	8886	8879	10212	10272
D	12374	10388	3956	2269	4910	9614	11380
E	12328	10746	5810	3905	5252	9514	11244
F	11250	11549	10536	10437	10470	11430	11362
G	8823	10511	11417	12244	12036	11545	10116

Fig. 5. The standard deviation of intensities of 4×4 windows at the starting positions on Fig. 3.

We can see that averages range from 4,193 to 28,847, which is a high variance. The standard deviations shown in Fig. 5 range from 2,269 to 12,374, which demonstrates that there is less variation when the computing window covers the center. Since every pixel in the window area contributes to the average, the spot morphology will affect the center or starting position, but it has little impact on the average. Assuming the spot can be narrowed down to the T1 area from C3 to F6, there are still 16 possible averages shown in Table II.

There are two different methods of calculating the spot's intensity value — fixed or dynamic area of intensity calculating area. These will have much different results. A fixed area is possible included background pixels and reduces the averaged intensity. A dynamic area should only count the oversaturated pixels, which is not representative of the reaction and may skew the result. Therefore, determining the threshold to determine a spot is very important. Here, we will analyze both methods.

B. *Dynamic Area Method*

In contrast to the fixed area method, this method computes the average of only the pixels which are considered in the spot. Therefore, the threshold to determine which pixels are within part of the spot will have great impact on the average. Based on the statistical functions, the pixels are classified into different categories as shown in Table II, with different thresholds applied on spot pixel selection. In a binary segmentation, a pixel will be considered as part of the spot if its intensity is greater than BG — the background threshold. Here, 4 different thresholds, *T*0, *T*1, *T*2, and *T*3 are utilized to determine whether a pixel is part of the spot. This method could be more focused on the higher signal pixels which will result in a higher average. However, it is dangerous to exclude the non-spot pixels because the reaction will never be evenly distributed because of the nature of the surface or concentration. A spot could appear perfectly round and bright, but its threshold binary image could be like a donut whose

center has much lower intensities. It is strongly suggested if the deviation within the averaging area is too high, the spot should be considered a failure.

TABLE II: THE AVERAGED INTENSITIES AND NUMBER OF PIXELS IN DIFFERENT THRESHOLDS

Threshold	Condition	value	# of pixels	Averaged Intensity
Background (BG)	<= Q2	<= 2,848	48	1,356
Ambiguous (*T*0)	> BG && < T1	> 2,848 < 6,095	13	16,712
Threshold 1 (*T*1)	>= (Avg+Q2)/2	>= 6,095	7	20,993
Threshold 2 (*T*2)	>= Avg	>= 9,341	8	23,937
Threshold 3 (*T*3)	>= Avg + Std	>= 20,228	22	27,276
Peak (PK)	>= Avg+2*Std	>= 31,116	2	33,090

III. GENOTYPING RESULT ANALYSIS

Most of the microarray genotyping system retrieves the spot intensities from a two color images system and converts the two color intensities into a 2-D plot with the converted intensity of color 1 and 2 as *p* and *q*, respectively. Many of the genotyping call methods use the "angle" method to separate the groups and assign the calls. That means the genotype call of a spot is determined by the angle of location (*p*, *q*) or (*q*, *p*) on the plot. Theoretically, the spots that are located near the *x* and *y* axis will have homozygous genotype calls of XX and YY, respectively. The spots located at 45 degrees will be assigned heterozygous calls XY, as shown in Fig. 6. If the plot is equally divided, each genotype region will cover 30 degrees. As explained in Section I, the intensity value will never be the same even if the same microarray is re-scanned again, because many factors could affect the second outcome. Here, in order to fairly analyze the signals, we assume both colors have the same intensity distribution shown in Tables I and II. Ideally, since it is the same spot on the microarray *p* and *q* should be the same. That means $p/q = 1$, which is 45 degrees. In fact, as shown in previous Sections II.A and II.B, a spot's intensity value could be in a wide range between *p* and *q*. The angle of location could range between *atan*(*p*/*q*) and *atan*(*q*/*p*). The analysis of the two methods is described below.

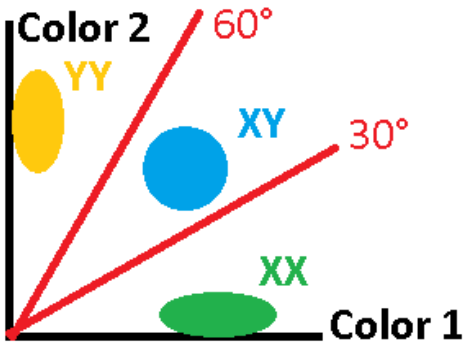

Fig. 6. The genotype call plot in a two-color laser system. The XX is near the *x* axis, the YY near the *y*, and the XY along the diagonal.

A. Analysis on Fixed Area Method

The average intensity values of the fixed area method have a wide range from 4,193 to 28,847, as shown on Fig. 4. It is obviously that many background pixels are included which caused the high variance. If the border pixels are excluded and only the immediate surrounding pixels are utilized, the average intensity values range from 21,731 to 28,847. This covers the cells C3 to E5 (center at D4) shown in Fig. 4. The angle range will be from *atan*(21,731/28,847) to *atan*(28,847/21,731), or between 37 and 53 degrees. If a bigger area is considered, such as from cells B2 to F6, the average intensity values will range between 11,895 and 28,847. The angle range will be from *atan*(11,895/28,847) to *atan*(28,847/11,895), or 22.4 to 67.6 degrees.

B. Analysis on Dynamic Area Method

The average intensity values range from 20,993 to 27,276 if *T*1, *T*2, and *T*3 are applied as shown on Table II. The angle range will be from *atan*(20,993/27,276) to *atan*(27,276/20,993), or 37.6 to 52.4 degrees. If the ambiguous pixels are included (threshold *T*0), the average intensity values range from 16,712 to 27,276. The angle range will be from *atan*(16,712/27,276) to *atan*(27,276/16,712), or 31.5 to 58.5 degrees.

Ideally, spots with an angle below 30 degrees should be definitely assigned different homozygous calls than the spots with angles above 60 degrees. If a spot is located between the main clusters of XX, XY, and YY, it is considered uncertain and its genotype call should be failed in order to avoid assigning a wrong call. From the analysis in Sections III.A and Section III.B, the same spot can generate a wide range of angles if the average value incorporates an uneven distribution intensity including ambiguous pixels or more background area.

IV. CONCLUSION

The much different genotype calls can dramatically change statistical results. The misalignment of spots can change intensities dramatically and cause inaccurate intensity values and genotype calls. Additionally, if the spot contains noise or the spot is like a donut shape, the spot's genotype call will be significantly shifted. Since most microarray technology users do not check if the spots alignments are correct, or if the intensity values are evenly distributed, vendors should always provide a systematic method to perform internal verification and allow users to view the spot alignment and intensity values. It is very necessary for the user to double check the spots' intensity distribution. It is very important to apply very strict rules to fail any suspected spots, as passing them will cause incorrect genotype calls and skew the statistical results.

REFERENCE

[1] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "DbSNP: The NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[2] S. A. McCarroll *et al.*, "Integrated detection and population-genetic analysis of SNPs and copy number variation," *Nature Genetics*, vol. 40, pp. 1166-1174, 2008.

[3] L. Rueda and I. Rezaeian, "A fully automatic gridding method for cDNA microarray images," *BMC Bioinformatics*, vol. 12, p. 113, 2011.

[4] M. Anandhavalli, C. Mishra, and M. K. Ghose, "Analysis of microarray image spots intensity: A comparative study," *International Journal of Computer Theory and Engineering*, vol. 1, no. 5, pp. 1793-8201, December 2009.

[5] R. Nagarajan and C. A. Peterson, "Identifying spots in microarray images," *IEEE Trans Nanobioscience*, vol. 1, no. 2, pp. 78-84, June 2002.
[6] W. T. Yin, T. Chen, X. S. Zhou, and A. Chakraborty, "Background correction for cDNA microarray images using the TV+L model," *Bioinformatics*, vol. 21, no. 10, pp. 2410–2416, 2005.
[7] E. M. Smith, J. Littrell, and M. Olivier, "Automated SNP genotype clustering algorithm to improve data completeness in high-throughput SNP genotyping datasets from custom arrays," *Genomics Proteomics Bioinformatics*, vol. 5, no. 3-4, pp. 256-259, December 2007.
[8] Y. Yang, P. Stafford, and Y. J. Kim, "Segmentation and intensity estimation for microarray images with saturated pixels," *BMC Bioinformatics*, vol. 12, p. 462, November 2011.
[9] A. N. Jain, T. A. Tokuyasu, M. S. Antoine *et al*., "Fully automatic quantification of microarray image data," *Genome Research*, vol. 12, pp. 325-332, February 2002.
[10] J. Deepa and T. Tessamma, "A new gridding technique for high density microarray images using intensity projection profile of best sub image," *Computer Engineering and Intelligent Systems*, vol. 4, no. 1, 2013.

Ching-yu Huang is an assistant professor of the Department of Computer Science at Kean University since September 2014. Dr. Huang was born in 1968, in Taiwan and received a Ph.D. degree in computer & information science from New Jersey Institute of Technology, Newark, New Jersey, USA in January 1998.

Prior to joining Kean University, Dr. Huang had more than 16 years of experience in the industry and academics in software development and R&D in bioinformatics. His research focuses on SNP genotype calling and cluster detection; image processing and pattern recognition, especially in microarray and fingerprint; geotagged images and location information reconstruction; database application development; data processing automation; e-learning, educational multimedia, methodology, and online tools for secondary schools and colleges. Dr. Huang has more than 20 publications in journals and conferences and more than 20 presentations in workshops and invited lectures.