

Background Correction Method for DNA Microarray Image Processing

Omar Salem Baans
School of Microelectronic Engineering
University Malaysia Perlis
Perlis, Malaysia
Email: omersalim4901@gmail.com

Asral Bahari Jambek
School of Microelectronic Engineering
University Malaysia Perlis
Perlis, Malaysia
Email: asral@unimap.edu.my

Abstract- Most microarray image scanning approaches provide an estimation of the intensity of the foreground and background for each spot. Background intensity must be corrected in order to remove the effect of non-specific binding or spatial heterogeneity across the array, but when such corrections are applied many problems appear, such as negative intensity for the spot or high variability of low-intensity log ratios. In this paper, many alternative methods for calculating background intensity are discussed and many approaches for background correction are tested and compared. GenePix, ScanAlyze and QuantArray are the strategies that were reviewed for background locations to extract their intensity. Similarly, to GenePix, a new approach for background calculation was proposed and tested. It shows more accurate results and the occurrences of error become lesser.

Keywords—Microarray; image; processing; foreground; background; location; DNA; correction.

I. INTRODUCTION

Nucleic Acid expression regulates the production of proteins which control all cellular processes in the human biological system. The understanding of gene expression and the mechanism of protein production has many applications in terms of diagnosis, staging and finding suitable treatments for diseases. Using the cDNA microarray, it is possible to diagnose rapidly and efficiently the level of gene expression in the sample [1], [2].

There are many commercial and freeware microarray analysis software packages available. Each software program can be separated into three main tasks. The first is gridding or addressing, which is the process of specifying coordinates for every spot on the slide. The second is segmentation, which classifies each pixel as either foreground, corresponding to a spot of interest, or as background, corresponding to error or noise. The third and final task is intensity extraction, which is the step which, for each spot on the array, calculates the green and red foreground fluorescence intensity [3], [4], [5].

The estimation of the background intensity is a very important step to be performed as part of the background correction process. This is because each spot intensity measure includes a contribution to the fluorescence that is not due to hybridization of the mRNA sample to produce spotted DNA. Background intensity can be estimated by more than one method, for example, using the main concentration of pixels located outside the spot mask [6], [7].

The background correction for the spot intensity can usually be performed by subtracting the background intensity from the foreground intensity of the spot, but sometimes negative values appear where the background value is greater than the foreground value for a given spot, which seems illogical. Many studies have discussed this issue and proposed various methods that can be applied to avoid the problem [8] [9].

The main aim of this paper is to propose and examine a new background correction alternative that can be used to calculate the background value pair (Rb, Gb). In this paper, several microarray background correction methods are discussed. Section II presents the existing algorithm systems for calculating background intensity values, while section II.A discusses and compares the different methods used for background correction. Section III proposes the new background correction in this paper. Section IV discusses the methodologies used in this project to correct background intensity. Section V elaborates and compares the results of different methods and Section VI, which is the final part of this paper, presents the conclusions.

II. LITERATURE REVIEW

A. Most Popular Background Intensity Extraction Methods

Most microarray analyses define foreground intensity values for red and green fluorescence (Rf, Gf) as the mean or median value of all the pixel values inside the segmented spot mask as shown by the red circle at the center. However, other variations are available for the calculation of the background intensity values for red and green fluorescence (Rb, Gb) especially with respect to selecting the background region. Among the most popular approaches that utilize the median for the specially-selected region around the spot mask are GenePix, ScanAlyze and QuantArray.

The GenePix package considers the median value of all the pixel values inside the valley region as the foreground intensity value of the spot. The valley region is represented in Figure 1 as four pink rectangular areas surrounding the spot. The background intensity estimation methods implemented in ScanAlyze consider all the pixels that are located outside the spot mask but within the square where the spot lies. This is represented in Figure 1 as the blue square surrounding the red circle. The median of all these pixels is used as the estimated value of the local background intensity of the spot. Finally, the QuantArray method uses the median of every pixel in the area

between the two green concentric circles in Figure 1 as the background intensity value for that spot [11]. While the QuantArray method uses the median of every pixel in the area between the two green concentric circles in Figure 1 as the background intensity value for that spot [12], [13], [14].

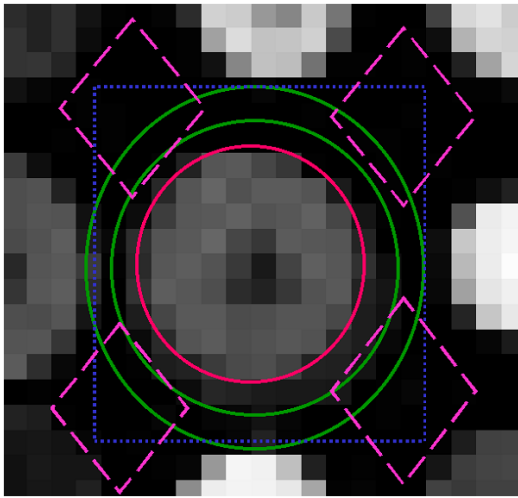


Fig. 1: Different background adjustment method

TABLE 1 COMPARISON BETWEEN DIFFERENT BACKGROUND ESTIMATION ALTERNATIVES

Features	[10]	[11]	[12]
Method	GenePix	ScanAlyze	QuantArray
Background	Four diamonds	Blue square & spot mask	Two concentric circles
Background calculation	Median	Median	Median

From the data collected in Table 1, it can be seen that using the median of the pixels' values in the GenePix region decreases the possibility of calculating an erroneous value, because the region is very distant from the spot and will not include any pixel that belongs to the spot. Thus, miscalculation is avoided for the bright pixel. However, estimation inaccuracy may occur in the other approaches, e.g., in ScanAlyze where there is overlap between the foreground and background areas, and hence it is inevitable that the values of pixels belonging to the spot will be included in the calculation.

B. Background Correction Method

The usual calculation method for finding the corrected value for the intensity of the spot for a two-color microarray is that of subtracting the values of the background intensity of the two colors, red and green (R_b , G_b) from the values of the foreground intensity of the two colors (R_f , G_f). This allows the true spot intensity (R , G) to be more accurately evaluated. This is because the observed foreground intensity value is the sum of the background value and the true intensity value of the spot [15], [16].

A simpler method for avoiding negative intensity values was presented by Edwards using a local median of the background values. This method is similar to the standard method, but when the value after subtraction is less than a specific small threshold value, it undergoes a monotonic function. Whereas when it is larger than the threshold value, it can be considered to be the true intensity of the spot [17], [3].

III. PROPOSED METHOD

From the algorithms in section II, the criteria of DNA microarray image processing can be improved by proposing a new algorithm specially in calculating background correction. It was mentioned before that the best method for background location is GenePix. However, to get more accurate result, the number of background square locations should be more than four to take more locations especially on the four edges of each side as well the red rectangular in Figure 2 below.

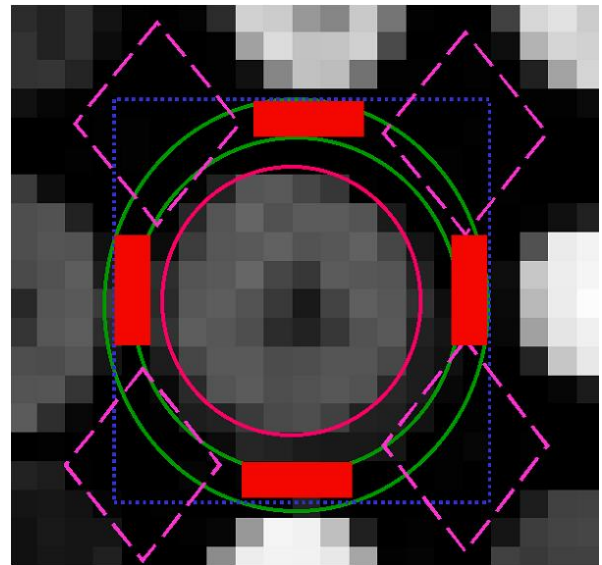


Fig. 2: Background proposed locations

IV. METHODOLOGY

As it is mentioned before in this article, to extract the intensity from DNA microarray slide, the user must go through three main process; addressing, segmentation and intensity extraction. Using Matlab R2016b, we developed a code that can extract the intensity for 100 spots automatically. As it can be seen in the flow chart in Figure 3, matlab code starts by importing the microarray image slide as the image in Figure 4. Then, it precedes gridding process by convert the image to grayscale image, estimate spot spacing by autocorrelation, locate the centers of spots, draw gridding lines to separate each spot. After that, it proceeds by segmentation using logically combine local and global thresholds.

Background intensity is calculated by taking the median of the four squares in the corner of each spot. Similarly, we tested adding another four rectangular on the middle edge of each side of each spot. Background correction performed using

Edward algorithm [17]. Finally, the intensity of each spot was extracted by taking the median of the spot mask.

The mathematical Equations of background corrections starts firstly by calculating the red and green foreground (R_f , G_f) by finding the median value of the pixels inside each spot mask and calculate red and green background intensities (R_b , G_b) by finding the median value of the pixels inside the local background areas of the spot. Then, the spot intensity was

calculated by subtracting the background from the foreground as in Equation (1) and (2):

$$R = R_f - R_b \quad (1)$$

$$G = G_f - G_b \quad (2)$$

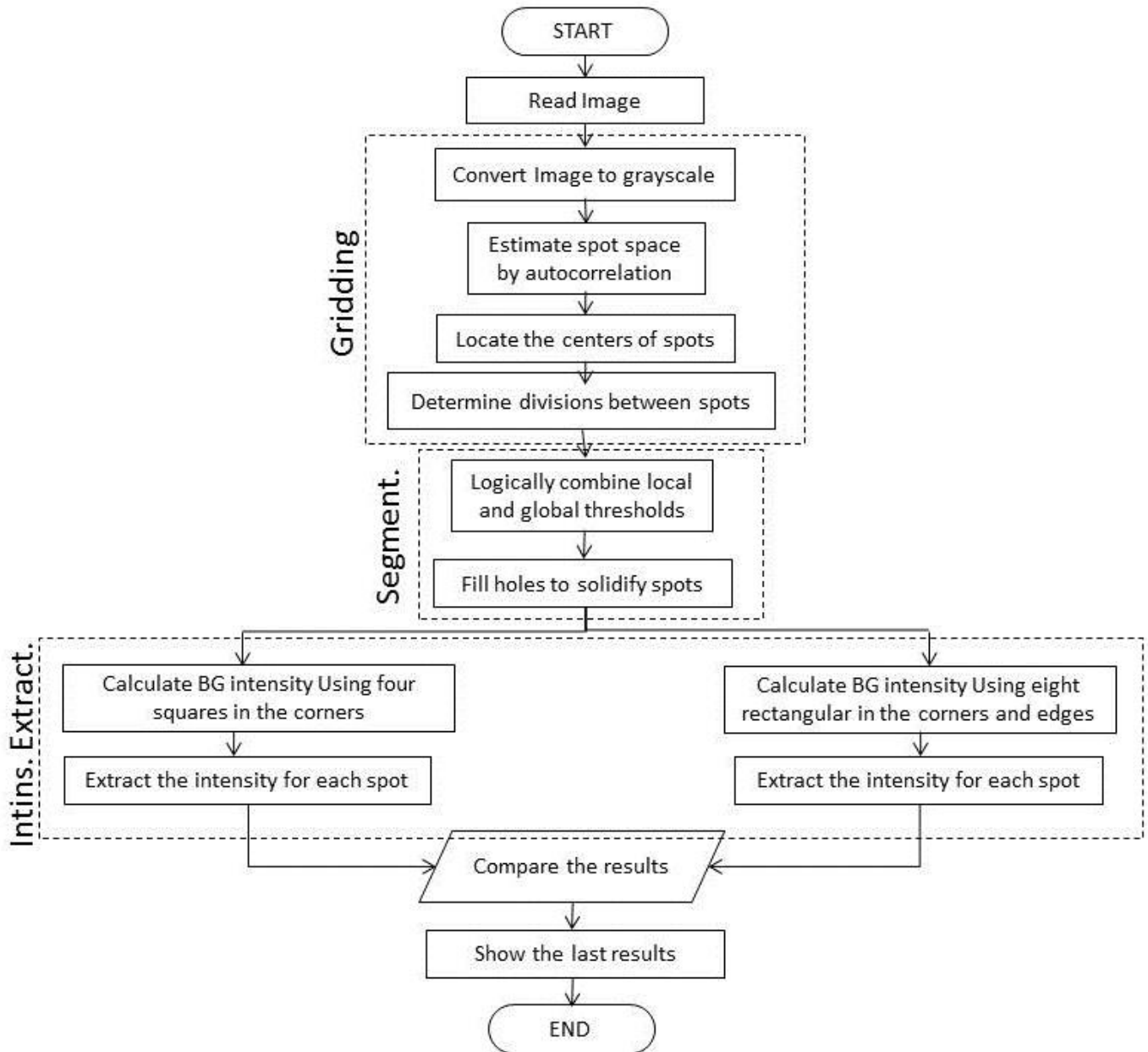


Fig. 3: Methodology Flow chart

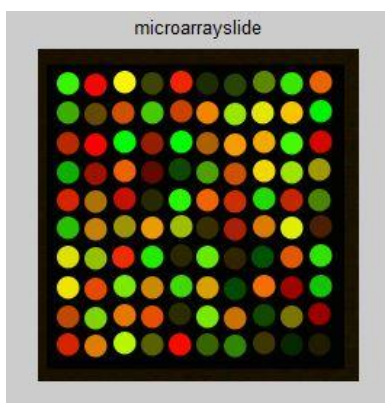


Fig. 4: Microarray ideal background image slide

The results of GenePix and this new algorithm compared and discussed. Another 6 more images _such as image in figure 5 compiled using this code with different background values while the foreground is similar. Thus, the conclusion becomes easier to pick the most accurate method.



Fig. 5: Microarray image slide

In order to validate the results, another three microarray images as in Figure 6, each image consists of one hundred spots. These images utilized in this work are real pictures obtained from a public database of the Princeton University microarray database. Princeton University microarray database provide the measured intensity information for each image [18]. Thus, this information was used as a reference to compare and validate this research results according to four parameters, these parameters are PSNR (Peak Signal to Noise Ratio).

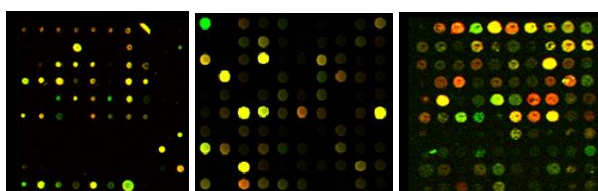


Fig. 6: Princeton DNA Microarray Images to Validate the Results [18]

V. RESULT AND DISSCUSSION

First of all, almost all the methods share a similarity especially when we compare the difference between the green and the red intensity for the same spots. The different usually

fluctuating between zeros to 6 pixels. Matrix 1 below shows the results for Ideal Microarray Image that has a similar background values by 3 pixels and that's why once we tried to show 100% similarities between GenePix BG location and the proposed method.

Red_Intensity =									
51	243	251	58	242	25	37	92	51	234
58	100	208	68	202	241	149	228	246	0
185	251	0	150	0	173	247	239	65	211
11	156	235	96	14	74	204	240	154	158
215	171	195	37	37	239	201	29	198	75
35	193	156	234	156	54	169	219	219	72
220	147	239	33	40	96	47	0	222	44
237	229	122	203	64	213	0	244	156	16
191	124	231	230	42	118	202	15	121	151
215	221	181	86	250	51	47	61	5	29
Green_Intensity =									
245	3	242	66	33	41	66	133	231	97
171	70	77	200	62	130	227	226	193	238
37	0	252	27	252	93	152	167	249	4
173	15	95	6	67	158	78	213	225	151
32	112	13	38	246	97	43	216	34	129
190	127	145	152	189	43	28	119	236	28
221	191	41	231	37	233	33	79	86	225
225	71	231	134	214	158	69	107	0	194
69	211	116	75	41	231	111	69	115	0
32	122	244	94	5	98	131	51	36	25

Matrix 1: Results of background correction method for Ideal Image

Matrix 2 shows the intensity extraction for DNA microarray image in Figure 5. Therefore, from matrix 1 and 2 the reader can know the background intensity for each spot in Figure 5 because it is fix number equals to three for every spot in Figure 4. Thus, red background intensity in the first in Figure 5 is equal to 2. However, green background intensity for the same spot is equal to 18 and so on. From this we can see how important of performing background correction before proceeding to extract the intensity of the spot.

Red_Intensity =									
52	214	239	50	218	22	33	84	50	217
55	89	186	67	184	222	144	216	230	3
167	224	0	134	0	158	225	222	66	189
10	138	214	82	12	70	186	225	150	147
192	156	169	30	36	219	183	24	173	70
36	177	112	217	149	45	151	197	209	61
208	141	212	37	34	95	37	0	201	45
225	203	118	188	60	201	0	228	139	15
173	121	209	204	33	113	181	12	110	133
194	203	177	78	222	49	47	52	5	21

Green_Intensity =

230	7	234	61	36	38	61	126	215	96
163	68	77	189	62	129	216	218	188	221
38	2	236	27	235	92	149	163	234	8
161	19	96	7	61	151	78	207	213	146
34	110	15	34	230	97	45	194	33	122
180	122	109	148	181	41	29	116	228	27
214	181	42	214	33	220	31	72	87	212
219	72	220	131	203	153	62	106	3	181
69	201	113	75	38	220	110	63	110	1
36	121	235	89	11	91	123	47	31	22

Matrix 2: Results intensity extraction for DNA microarray image in Figure 5

Matrix 3 shows the different between using eight rectangular instead of four to calculate background intensity. As it can be seen below, there is for most of the spots and these different is varies between 0 and 6. However, one of the green spots shows a different of 32 pixels. That because when the user increases the locations of background the accuracy will be increased as well and the occurrences of the error will be decreased.

R_bg_diff=

2	-2	-3	0	-4	-1	-1	-1	0	-2
0	-3	-3	1	-1	-3	-1	-2	-2	2
-2	-6	0	-4	0	-3	-6	-3	1	-3
0	-4	-4	-4	0	-1	-6	-2	0	-1
-6	-4	-3	1	3	-3	-2	1	-4	1
1	-4	-4	-1	0	-3	-3	-5	0	-1
0	2	-3	5	1	2	0	0	-3	3
-2	-3	2	-2	2	-3	0	-4	-3	2
-2	3	-3	-4	-3	1	-4	0	-2	-3
-3	-2	1	-1	-2	1	3	0	1	0

G_bg_diff=

-3	0	-3	-1	1	-1	0	-1	-2	0
-2	0	1	-2	0	-1	-3	-1	-2	-3
1	2	-2	0	-2	0	0	-1	-3	1
0	2	1	2	0	0	1	-1	-4	0
1	0	0	-1	-3	1	1	-2	2	-1
-1	-1	-32	-1	-1	1	2	0	0	0
-1	-3	2	-2	-1	-2	-1	-2	3	-2
1	3	0	0	-2	0	-1	2	1	-2
1	-2	1	1	0	-1	1	-2	0	1
3	0	-1	-1	4	-1	-2	-1	0	-1

Matrix 3: The different between using eight rectangular instead of four to calculate background intensity.

Table 2 compares background corrections method applied on three DNA microarray images in Fig 6. The results of GenePix and the proposed method were compared with the Princeton database as a reference. The comparison was done depending the Peak Signal to Nois Ration PSNR as parameters for the comparison.

TABLE2: BACKGROUND CORRECTION COMPARISONS BETWEEN THE PROPOSED METHOD AND GENEPIX METHOD.

Method	Image (2)	Image (2)	Image (2)
GenePix	40.22 dB	43.17 dB	41.39 dB
Proposed	40.35 dB	43.18 dB	41.47 dB

According to Table2, the accuracy of the proposed method is much better than GenePix methods where PSNR for the proposed is greater than PSNR for GenePix method for all the tested three images.

VI. CONCLUSION

In this paper, a number of background correction calculations and algorithms were reviewed. The differences and similarities of the existing systems were studied in order to identify criteria for selecting the best program, i.e., the program which gives results closest to the true intensity values. Three methods for allocating and calculating the background intensity values were discussed and compared. These methods were GenePix, ScanAlyze and QuantArray. To correct the background intensity, a new algorithm was applied on microarray slide image using MATLAB in order to find the most accurate intensity value of the two-color microarray for each spot; and the result for that method and other existing methods were calculated and compared. Based on the findings, it is sufficed to state that choosing more background locations in the corners and the sides is the best way to find the most accurate background intensity.

VII. ACKNOWLEDGMENT

This research was funded by ScienceFund, Ministry of Science, Technology and Innovation (MOSTI), Malaysia (2015).

VIII. REFERENCE

- [1] Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T. Improved background correction for spotted DNA microarrays. *Journal of Computational Biology*. 2002 Jan 1;9(1):55-66.
- [2] Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*. 2000 Aug 29;97(18):9834-9.
- [3] Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. *Journal of computational and graphical statistics*. 2002 Mar 1;11(1):108-36.
- [4] Mabrouk M, Fouad I, Sharawy A. Automated statistical and morphological based gridding methods for noisy microarray image processing. *Journal of Bioinformatics and Intelligent Control*. 2013 Dec 1;2(4):265-75s.
- [5] Ahmad MM, Jambek AB, bin Mashor MY. A study on microarray image gridding techniques for DNA

- analysis. In Electronic Design (ICED), 2014 2nd International Conference on 2014 Aug 19 (pp. 171-175). IEEE
- [6] Sifakis EG, Prentza A, Koutsouris D, Chatziioannou AA. Evaluating the effect of various background correction methods regarding noise reduction, in two-channel microarray data. *Computers in biology and medicine*. 2012 Jan 31;42(1):19-29.
- [7] Smyth GK, Speed T. Normalization of cDNA microarray data. *Methods*. 2003 Dec 31;31(4):265-73
- [8] Geeleher MP, Morris D, Golden A, Hinde J. *BioconductorBuntu Users Manual*.
- [9] Argyropoulos C, Daskalakis A, Nikiforidis GC, Sakellaropoulos GC. Background adjustment of cDNA microarray images by Maximum Entropy distributions. *Journal of biomedical informatics*. 2010 Aug 31;43(4):496-509.
- [10] Fielden MR, Halgren RG, Dere E, Zacharewski TR. GP3: GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics*. 2002 May 1;18(5):771-3.
- [11] Eisen, Michael. "ScanAlyze user manual." Stanford Univ. US (2013).
- [12] Luminomics, G. S. I. "QuantArray Analysis Software, Operator's Manual." GSI Luminomics (2009).
- [13] Bengtsson A, Bengtsson H. Microarray image analysis: background estimation using quantile and morphological filters. *BMC bioinformatics*. 2006 Feb 28;7(1):96.
- [14] Affymetrix Expression Console Software Version 1.0, User Guide. Santa Clara: Affymetrix Inc.; 2006.
- [15] Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*. 2007 Oct 15;23(20):2700-7.
- [16] Yin W, Chen T, Zhou SX, Chakraborty A. Background correction for cDNA microarray images using the TV+L1 model. *Bioinformatics*. 2005 May 15;21(10):2410-6.
- [17] Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*. 2003 May 1;19(7):825-33.
- [18] Ball, Catherine A., et al. "The Stanford Microarray Database accommodates additional microarray platforms and data formats." *Nucleic Acids Research* 33.suppl_1 (2005): D580-D582.