

# Mathematical model for saturation and quantization reduction in microarray experiments

## SUPPLEMENTARY MATERIAL

Jorge García de la Nava<sup>1</sup>, Sacha A.F.T. van Hijum<sup>2</sup>, and Oswaldo Trelles<sup>1\*</sup>

<sup>1</sup> Computer Architecture Department, University of Malaga, Spain

Complejo Politécnico, Campus de Teatinos, Apdo. 4114, E-29080 Málaga, Spain

<sup>2</sup> Department of Molecular Genetics, Groningen Biomolecular Sciences and

Biotechnology Institute, University of Groningen, P.O. Box 14, NL-9750 AA Haren, the  
Netherlands.

---

\* *To whom all correspondence should be addressed*

1.- Data Representation .....	3
2.- Regression .....	3
3.- Threshold algorithm .....	4
4.- Maximum likelihood algorithm .....	5
5.- Results .....	6
6.- Gamma curve justification .....	7
7.- Clipping curve justification.....	8
8.- Detailed description of data sets .....	9
8.1.- RC data set.....	9
8.2.- D, IE and IR data sets .....	11
8.3.- RL data set.....	11
9.- Relaxing the linear assumption .....	11
10.- Detailed model of quantification.....	12
11.- Effect of quantization and saturation in image and measure histograms.....	14
12.- Effects of quantization in low intensity spots .....	16
13.- Effects of saturation in high intensity spots .....	17
14.- References .....	18

## 1.- Data Representation

The data set must contain the following data fields for each of the two channels (green and red or control and target) and for both measures (High- and Low-sensitivity): luminescence (raw intensities) and background intensities. Algorithms will be applied independently to each channel, one for the red channel and other for the green channel. Two independent copies of any parameter used (gamma, multiplicative coefficient, etc.) will be kept for each channel. Background will not be subtracted, keeping the spot intensity as well as the background intensity during the procedure. Adjusted spot and background intensities in both channels will be the result.

Due to the independence on both channels, the description will be supplied for a generic channel.

**Definition:** for a generic channel, we name:

- $L_i$  Intensity of spot  $i$  in the low-sensitivity measure.
- $H_i$  Intensity of spot  $i$  in the high-sensitivity measure.
- $S_i$  Intensity of spot  $i$  in results.

## 2.- Regression

The regression must fit the gamma curve to the data and give an estimation of both parameters  $\gamma$  and  $b$ . This can be done in the logarithmic space as a linear regression which is a very well known technique. First, the means ( $\mu_L$ ,  $\mu_H$ ), the deviations ( $\sigma_L$ ,  $\sigma_H$ ) and covariance ( $\sigma_{LH}$ ) are calculated

$$\mu_L = \frac{1}{N} \sum_i \log_2(L_i) \quad (1)$$

$$\mu_H = \frac{1}{N} \sum_i \log_2(H_i) \quad (2)$$

$$\sigma_L = \sqrt{\frac{1}{N} \sum_i (\log_2(L_i) - \mu_L)^2} \quad (3)$$

$$\sigma_H = \sqrt{\frac{1}{N} \sum_i (\log_2(H_i) - \mu_H)^2} \quad (4)$$

$$\sigma_{LH} = \frac{1}{N} \sum_i (\log_2(L_i) - \mu_L)(\log_2(H_i) - \mu_H) \quad (5)$$

Then, a regression line is taken (L|H or H|L). The parameters that these lines provide are:

$$\gamma_1 = \frac{\sigma_{LH}}{\sigma_L^2} \quad (6) \quad b_1 = 2^{\mu_H - \gamma_1 \mu_L} \quad (7)$$

$$\gamma_2 = \frac{\sigma_{LH}}{\sigma_H^2} \quad (8) \quad b_2 = 2^{\mu_L - \gamma_2 \mu_H} \quad (9)$$

### 3.- Threshold algorithm

Once the parameters have been estimated, it is straightforward to establish a threshold for this algorithm by means of the equation reproduced below:

$$T = b(b\gamma)^{\frac{\gamma}{1-\gamma}} \quad (10)$$

Finally, it will suffice to iterate for choosing the value of one or the other dataset. The pseudo-code is:

```

FUNCTION threshold(ARRAY OF SPOTS H, ARRAY
OF SPOTS L, INTEGER N, ARRAY OF SPOTS S)
{
  REAL gamma, b, T;
  parameter_estimation(H,L,N,gamma,b);
  T:=threshold_calculation(gamma,b);
  FOR i:=1 TO N DO
    IF H[i]>T THEN
      S[i]:=L[i];
    ELSE
      S[i]:=translate(gamma,b,H[i]);
    END IF;
  END FOR;
}

```

The translation procedure is intended to rescale the high-sensitivity measure to the low-sensitivity measure range. It is done via the following equation:

$$L_i = (H_i / b)^{1/\gamma} \quad (11)$$

#### 4.- Maximum likelihood algorithm

The second algorithm requires the minimization of the equation:

$$(H_i - bS_i^\gamma)^2 + (L_i - S_i)^2 \quad (12)$$

Due to this, and given that the equation is quadratic-like with positive coefficients, we recommend finding the root of the derivative that must be the minimum for sure. This means resolving the following equation for each spot  $(H_i, L_i)$  from the data:

$$b\gamma(H_i - bS_i^\gamma)S_i^{\gamma-1} + (L_i - S_i) = 0 \quad (13)$$

A very efficient and recommended method for solving this equation is Bolzano's. The suggested initial interval for this method is  $[v, w]$  where:

$$v = L_i \quad (14)$$

$$w = (H_i / b)^{1/\gamma} \quad (15)$$

The pseudo-code for this second algorithm is:

```

FUNCTION max_lik(ARRAY OF SPOTS H, ARRAY OF
SPOTS L, INTEGER N, ARRAY OF SPOTS S)
{
  REAL gamma, b;
  parameter_estimation(H,L,N,gamma,b);
  FOR i:=1 TO N DO
    S[i]:=minimize_eq(gamma,b,L[i],H[i]);
  END FOR;
}

```

Where the minimization procedure is described above.

## 5.- Results

The algorithmic complexity will depend on the specific implementation of the auxiliary method of regression and minimization. If the suggestions were used, it can be shown that:

- 1.Linear regression has a complexity  $O(n)$
- 2.Minimization has a complexity  $O(\log(n))$
- 3.The threshold algorithm has a complexity  $O(n)$
- 4.The maximum-likelihood algorithm has a complexity  $O(n\log(m))$

The main improvement of these algorithms are that they take advantage of the information a double scanning with different sensitivity provides which, in turn, allows the reduction of the effects of both saturation and quantization that other procedures do not address.

## 6.- Gamma curve justification

To discover which saturation model should be used for a given dataset, the transfer function of the most widely used devices will be studied.

Photomultiplier tubes (PMTs) have long been recognized as the detector of choice for very weak light detection. The high gain of the PMT, in excess of  $10 \times 10^6$ , makes it a useful detector for applications such as LIDAR, bioluminescence, fluorescence, and chemiluminescence. Loosely defined, PMTs are optical components that convert incident photons into electrons via the photoelectric effect. When an incident photon impinges on the active surface of the PTM (the photocathode), a photoelectron is generated. This electron flows through a series of electron multipliers (dynodes) to the anode. The amount of current that flows from the anode is directly proportional to the amount of incident light at the photocathode; until saturation occurs (Hamamatsu 2002).

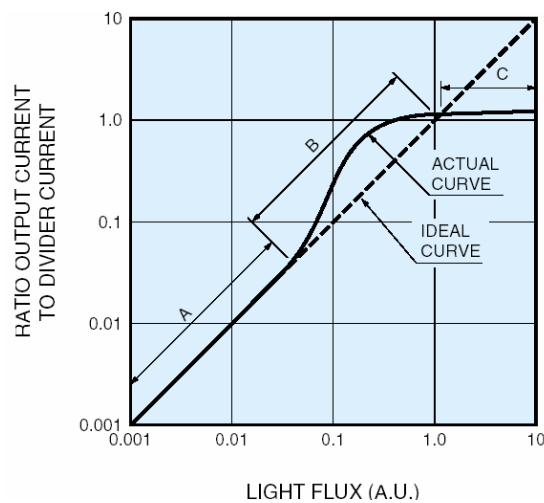
The transfer function of the photo-detectors used in the scanners has the shape depicted in figure a and can be divided into three zones in which the response to true signal ratio are: (i) linear , (ii) over-linear, and (iii) saturation. Approximating each region by a linear segment, the relation between variables can be expressed by the following equation of a line

$$Y = AX + B \quad (16)$$

For non-logarithmic variables this line becomes a gamma curve.

$$y = bx^A \quad (17)$$

This curve also appears in other kinds of electro-optic devices like monitors and televisions [1].



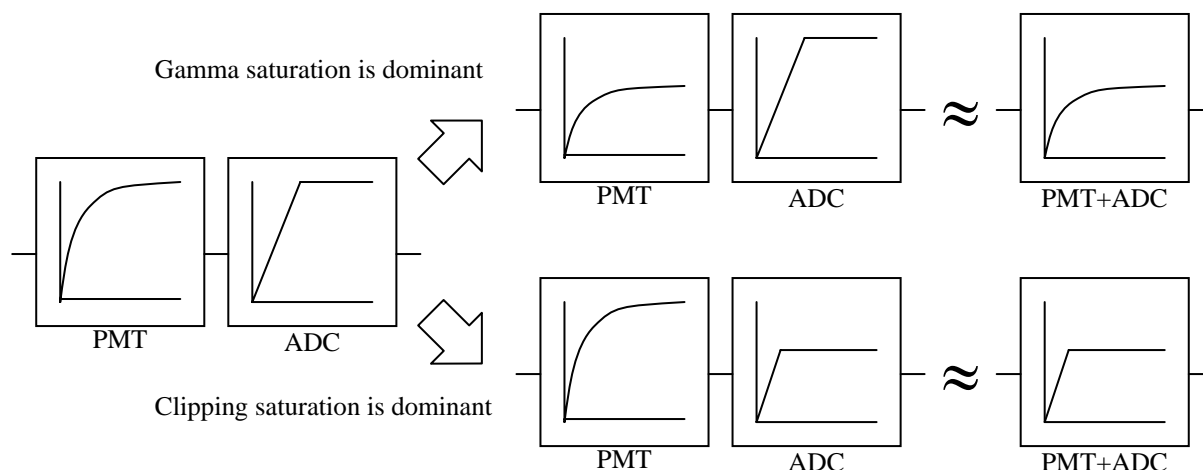
**Figure A.-** Output linearity of photomultiplier tube in which three zones can be discerned: region A maintains linearity, and the region B is the so called over-linearity range in which the output increase is larger than the real level. In region C, the output goes into saturation and becomes lower than the real level. From [2].

## 7.- Clipping curve justification

PMT tubes are only the first step in the image capturing process. After their amplification of the incident photons the signal is sent to an analog to digital converter (ADC) whose saturation is a clipping. This ADC performs a uniform quantization which is the assumption made in the main article.

Cascading two devices with unlike saturation types renders the global saturation scheme difficult to analyse. To overcome this, asymptotic approximations are used and are indeed verified by the experimental data. The first of these approximations is the supposition that the clipping saturation is dominant and, thus, the gamma saturation is neglected. Under this hypothesis, only the clipping saturation is appreciated in the data. The second approximation, the opposite one, makes the gamma saturation obscure the clipping saturation.

Table 1 of the main document shows which scanners provide data under the asymptotic approximation of clipping or gamma saturation. Figure B depicts above approximations.



**Figure B.-** The main devices that compose a scanner are the PMT and the ADC. PMTs can be modeled to have gamma saturation and ADCs a linear saturation. Approximating asymptotically the global saturation model to a dominant saturation in each case leads to the modeling used in the main document (see Table 1).

## 8.- Detailed description of data sets

Since the mathematical procedure claims the quality improvement of data acquisition from gene-expression scanned-images, we are not going to discuss in detail the procedure behind the experimental data.

### 8.1.- RC data set

Images from the different sensitivity scans and quantized data are available as supplementary material for validation purposes. The used data acquisition protocol is described here.

The Lucidea scorecard amplicons (Amersham Pharmacia, Sunnyvale, CA) were diluted 1:1 in Array-It spotting buffer (Array-It, Sunnyvale, CA) and transferred to 384 wells plates (Genetix Ltd, Hampshire, UK) for spotting. The Lucidea scorecard contains reference (R) and Test (T) RNA spike preparations in specific ratios and concentrations that correspond to the scorecard's DNA yeast intergenic regions (YIR). The R and T RNA can be labelled with

either Cy3 or Cy5. The Lucidea scorecard comprises 21 YIR amplicons: (i) 10 calibration controls; with identical R and T RNA concentrations (Luc\_C and cYIR), (ii) 8 ratio controls; with varying R:T RNA ratios (Luc\_R and rYIR), (iii) 3 utility (user) controls, and (iv) 2 negative controls.

Aldehyde coated glass slides were spotted as described [3] with SMP4 stealth pins (Array-It) on a GenPak model GA2 1-1 spotter (Genpak Ltd, Stony Brook, NY). The following geometry was used for spotting: 12 spot pins (4.5 millimeter apart) delivered spots in 4 copies per slide (4 intra-slide replicate spots). Spot centers were at 250  $\mu$ meter distance from each other. A section (grid) spotted by the same spot pin thus consists of  $18 \times 18$  spots. A block consists of the 12 grids spotted by the spot pins. In total, two blocks were spotted. Thus one slide consists of  $4 \times 6$  grids with  $18 \times 18$  spots per grid. On each slide, in total 3444 spots (164 copies of the Lucidea controls) were delivered. Labeling of the Lucidea reference and test RNA was performed using an indirect labelings kit as described previously [3]. Hybridization was performed as described [3].

A GeneTac LS IV (Genomic Solutions Inc., Ann Arbor, MI) microarray scanner device was used for image data acquisition. This device has the flexibility for continuous control of the Photo Multiplier Tube (PMT), allowing two different settings, (i) low-sensitivity (Cy3 at gain 48 and black 50, and Cy5 at gain 55 and black 30) to avoid saturation, and (ii) high sensitivity to recover information on highly expressed signals (Cy3 at gain 55 and black 50, and Cy5 at gain 62 and black 30). The data sets were acquired from the TIFF files (all  $1500 \times 3200$  pixels) by ArrayPro 4.5 (Media Cybernetics Inc., Silver Spring, MD) image analysis software. Net spot intensities were calculated by measuring the whole spot area (raw signal) and subtracting the flat background (calculated from the slide's surface).

The data sets used in this study were obtained from 2 slides (1 and 2) labelled as Cy3 (R) and Cy5 (T) and 2 dye-swap slides (3 and 4) labelled as Cy3 (T) and Cy5 (R). In-house

developed software (PrePreP by S.A.F.T. van Hijum, department of Genetics, Groningen, the Netherlands) was used for removing empty spots (slide positions on which no DNA was spotted) from the data sets and parsing the spot descriptions.

## 8.2.- D, IE and IR data sets

Images from the different sensitivity scans and anonymized quantized data are available as supplementary material for validation purposes. The used data acquisition protocol is not described due to privacy matters, but provided because proposed algorithms are protocol independent.

## 8.3.- RL data set

Although this data set remains private, it is supplied to complete a testing set for the proposed methodology.

## 9.- Relaxing the linear assumption

It is stated in equation 1 of the main document that the relation between the read signal level and the intensity level is linear. This may not be true and the most general approximation to the PMT curve must be used like equations 16 and 17 of this document do. However this approach is equivalent to the linear assumption since all the exponents can be grouped in a single one:

$$L_i = c_L I_i^{\gamma_L} \quad (18)$$

$$H_i = c_H I_i^{\gamma_H} \quad (19)$$

$$H_i = b L_i^{\gamma} \quad (20)$$

Where:

$$\gamma = \gamma_L / \gamma_H \quad (21)$$

$$b = c_L / C_H^\gamma \quad (22)$$

But these parameters are to be estimated and the regression curve to use is the same as the one proposed in the main document (equation 6 in the main document and equation 20 here).

## 10.- Detailed model of quantification

The goal of this section is explaining the noise model. This is done by describing the steps that lead to the mentioned model.

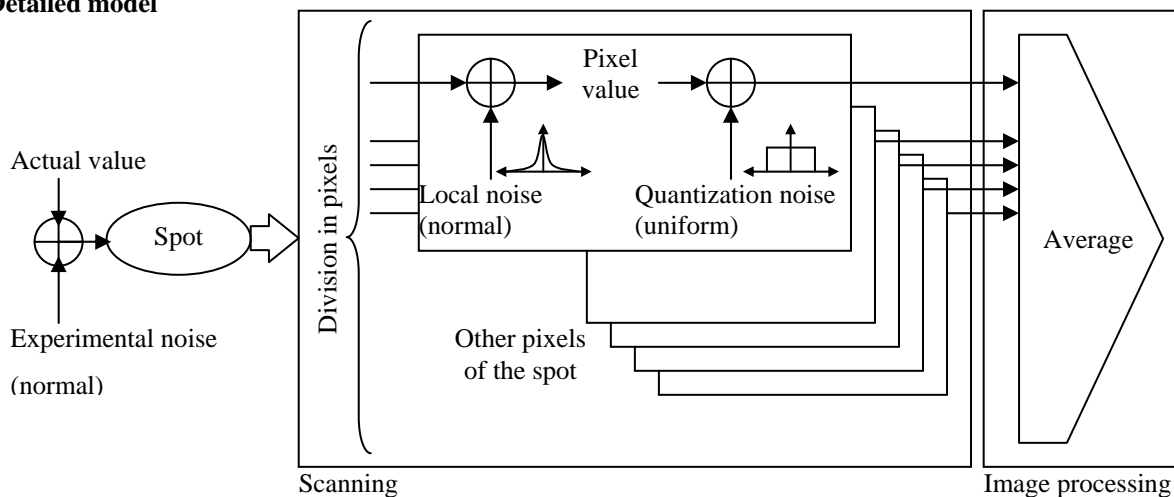
To obtain the noise model, several assumptions are to be done. These are:

- Each spot is formed by many pixels. What we are measuring is the spot intensity in these pixels for one of green or red channels.
- Measures already contain noise. We don't mind the type of this noise, but all the pixels of the spot are affected by the same amount. This noise will be called experimental noise.
- Measures already contain noise in each pixel. This is additive, centred, gaussian and of higher variance than the quantization step. The noise of each pixel will be called local noise and it is supposed to be independent of other pixels of the same spot.
- Image processing software uses an average-like estimator. As such, it will average the values of the pixels of the spot and that will be its estimation of the measure.
- Quantization noise is centred and classical: additive, uniform, dependent of the input of the quantizer, etc.

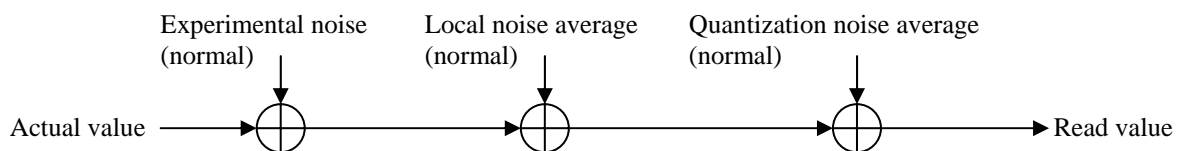
Figure C shows a diagram that relates all the above sources of noise. In it, the following results are used:

1. Since the local noise is supposed to have a much higher variance than the quantization step, quantization noise can be thought as independent.
2. The average of many sources of additive and uniform quantization noise which are independent yields a gaussian noise.
3. The average of local noise keeps the gaussian distribution (but with lower variance).
4. Experimental noise is constant across the pixels and any average procedure will not modify its value.

#### Detailed model



#### Equivalent model



**Figure C:** The noise model diagram. The upper part is a detailed description of the generation of the measure of a single spot. The lower part is the schematic description of a simplified model. The average of the uniform quantization noise of each pixel can be modeled as a single normal noise.

What is really of interest for us is the quantization noise. The characteristics of it in the equivalent model are:

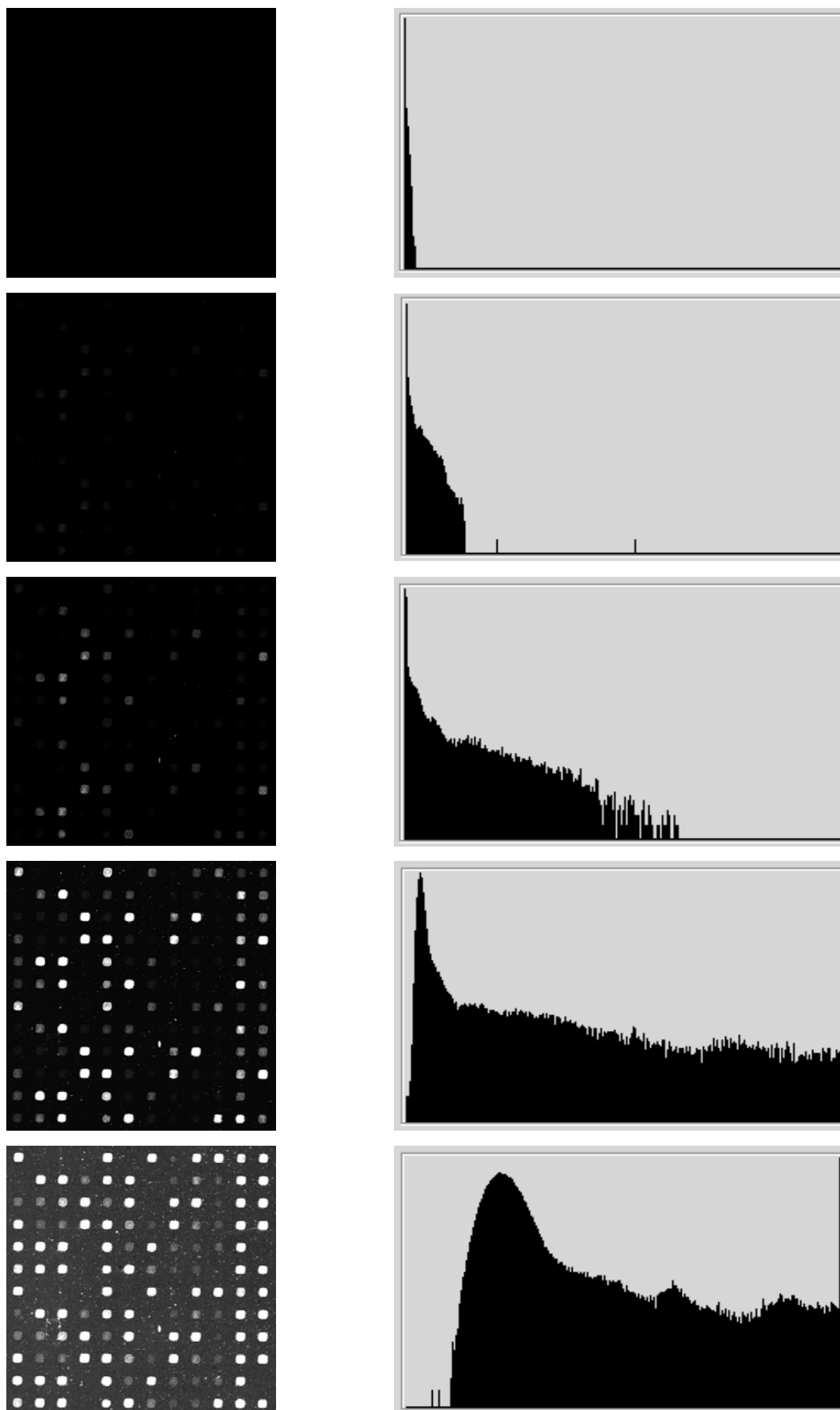
- Centred
- Gaussian
- Independent of measure

### **11.- Effect of quantization and saturation in histograms and comparison with results.**

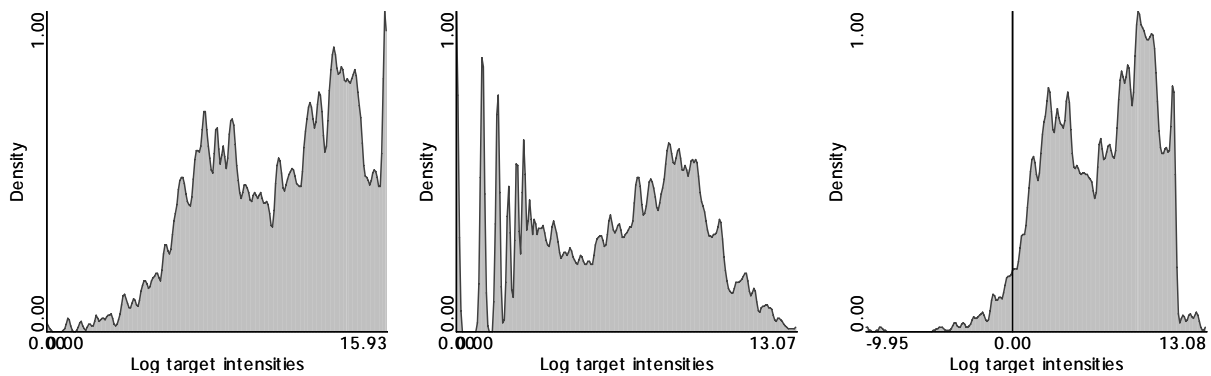
Histograms are a useful tool for analyzing the effect of parameter variations in the acquisition device. They represent the number of pixels in each quantization step (or an interval of them). Darkest values are on the left part of the histogram and brightest on the right. When an image is dark, its histogram shows a tendency to the left and, when it is light, to the right.

In Figure D five images of the same microarray chip but using different scanning sensitivities are shown along with their respective histograms. It is remarkable that a single vertical line appearing on the right of the histogram indicates that saturated spots have appeared. This happens in high sensitivity images (the lower ones). If the same vertical line appears on the left of the histogram, it means that dark values are being measured as black as effect of quantization. Such appearance occurs in low sensitivity images (the upper ones).

Instead of using a pixel-wise histogram, a spot-wise one can be used. Since the number of spots per image is much lower than the number of spots, density estimation was used for data completion. In this case a logarithmic scale is being used in horizontal axes for convenience. The aim of this is that 2Scan algorithms produce a resulting spot data file, not an image, which can be visualized by this kind of histogram, allowing an easier comparison between original and result data. Figure E shows spot-wise histograms for both raw and processed data from D data set. Saturation and quantization effects are evident in raw data and are corrected in processed data.



**Figure D.-** A representative portion of a slide (from the D dataset) scanned at different sensitivity levels. The PMT gain of a GenePix4000b slide scanner was set to 400, 500, 600, 800 and 1000 (from top to bottom). The laser power was fixed at 1000. On the right of each image, its histogram is drawn.



**Figure E.-** Results from the maximum likelihood algorithm for clipping saturation on D data set. The first density graph on the left is the high sensitivity measure. A saturation peak can be observed on the right part of the graph. The middle graph is created from a low sensitivity measure. Diverse peaks appear due to quantization on the left of the graph. On the right graph, the results from the algorithm are drawn. There is neither saturation nor quantization and the dynamic ranges is extended along 23 ( $=13.08 - (-9.95)$ ) logarithmic units ( $2^{23}$ ) instead of the scanner resolution ( $2^{16}$ ).

## 12.- Effects of quantization in low intensity spots and comparison with results.

It was said in the main document that very dark spots lie under the first quantization step and the used scanner reports them as black. Black is the zero intensity. Table A shows the count of spots of zero intensity value in several files of the RL data set. This data set suffer from gamma saturation.

Source	Number of Spots	Zero values	
		Red channel	Green channel
Low (RL)	5760	4300	4736
High (RL)	5760	2028	2338
Threshold	5760	2046	2377
Max. Lik.	5760	2026	2337

**Table A.-** Number of spots of zero value in different files from the RL data set (low and high sensitivity), its results from threshold and maximum likelihood algorithms.

As expected, when the sensitivity is low, the number of zero-valued spots is higher than in high sensitivity count. Both algorithms yield data that corrects this effect

### 13.- Effects of saturation in high intensity spots and comparison with results.

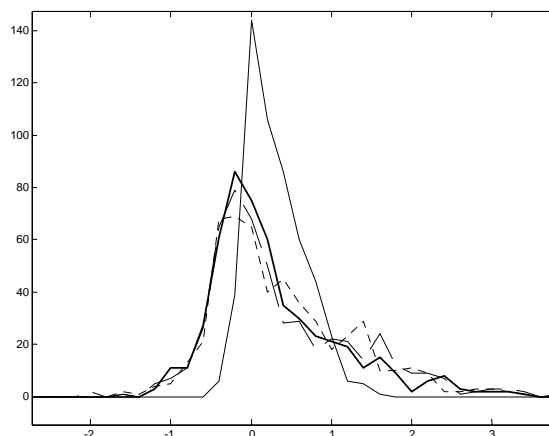
The principal consequence of saturation is ratio flattening. As intensities grow higher, saturation level is reached. Any ratio of saturated values will tend to one (zero in logarithmic space) since the saturation level is the same for any used channel. A way to assess the effect of saturation is observing the ratio behaviour of most bright spots. Table B represents some statistics of these spots.

Source	Maximum intensity (log2).	Number of spots.	Spots within twofold.	Percentage of spots within twofold.	Standard deviation of logratio.
Low (RL)	15.18485	524	428	81.7%	0.8317
High (RL)	15.41565	2235	496	94.7%	0.5185
Threshold	15.18485	439	387	73.9%	1.3569
Max. Lik.	15.17315	553	396	75.6%	1.3005
Yoshida <i>et al.</i> 2001	14.86170	771	308	39.9%	1.5045

**Table B.-** Several statistics from the spots of high intensities. These spots are those whose mean intensity is over a threshold of 1/64 saturation level. The statistics are: maximum value of mean intensity, number of spots, spots whose ratio is within twofold, percentage of them, and standard deviation of the logarithm of the ratio in base 2. Data sets are the RL one (both high and low sensitivity images), its results for threshold and maximum likelihood, and a high quality public domain data set. The used scanner has a dynamic range of 16 bits, image processing software averages spot pixels.

The standard deviation of the ratio logarithm of most intense spots is very small in high sensitivity data, which agree mentioned saturation effects. Algorithms output a higher deviation for these spots, this meaning that the information is expressed in a wider range of values.

If a graph showing the density of ratio logarithms is plotted (see Figure G), it is expected that saturated spots crowd close to zero, while unsaturated ones spread across a broader interval. Again, algorithm results palliate saturation effects.



**Figure G.-** Comparison of distribution of spots with higher intensity in the low-sensitivity scan (thick line), high-sensitivity scan (thin line), result of the threshold method (dashed line) and result of the maximum-likelihood method (dotted line). The effect of the saturation appears when the intensity of both channels is high. At increasing intensities the saturation becomes stronger, pushing the read value toward the maximum possible. Since this happens in both channels, the ratio will tend to one. When representing the ratio of the spots with greater intensity, we expect that in the high-sensitivity scan the ratios will be grouped around the unit, zero in logarithmic scale.

## 14.- References

- [1] Deguchi, T., Katoh, N. and Berns, R.S. (1999) Clarification of “Gamma” and the Accurate Characterization of CRT Monitors, *Proceedings SID International Symposium*.
- [2] Hamamatsu Photonics (2002) Photomultiplier Tubes, Photomultiplier Tubes and Related Devices. *Catalog* June 2002 ([www.hamamatsu.com](http://usa.hamamatsu.com/hcpdf/catsandguides/PMTCAT_accessories.pdf)) ([http://usa.hamamatsu.com/hcpdf/catsandguides/PMTCAT\\_accessories.pdf](http://usa.hamamatsu.com/hcpdf/catsandguides/PMTCAT_accessories.pdf)).
- [3] Kuipers, O.P., De Jong, A., Baerends, R.J.S., Van Hijum, S.A.F.T., Zomer, A.L., Karsens, H.A., Den Hengst, C.D., Kramer, N.E., Buist, G. and Kok, J. (2002) Transcriptome analysis and related databases of *Lactococcus lactis*. *Antonie Van Leeuwenhoek*, **82**, 113-22.