Rough Draft
Not For Distribution

aCGH Data Normalizing Algorithm

Jeffrey C. Miecznikowski[ab1], Daniel P. Gaile[ab], Conroy, Jeffrey, Norma J Nowak[c]

[a]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214-3000, USA
[b]Department of Biostatistics, Roswell Park Cancer Institute, New York 14263
[c]Cancer Genetics, Roswell Park Cancer Institute, New York 14263

Short title:

Normalization for aCGH

Proofs to be sent to:

Jeffrey C. Miecznikowski
Department of Biostatistics
School of Public Health and Health Professions
249 Farber Hall
University at Buffalo
3435 Main Street
Buffalo NY 14214-3000, USA

[1]Corresponding Author. Department of Biostatistics, School of Public Health and Health Professions, 249 Farber Hall, University at Buffalo, 3435 Main Street, Buffalo NY 14214-3000, USA. Tel:+1(716) 829 2754. e-mail:dpgaile@buffalo.edu

In any new technology, studies involving quality control and tolerance should be performed. Once the technology has been benchmarked as suitable, continuing checks should be performed in order to ensure that the technology remains in check. With the burgeoning field of high throughput genomics, specifically microarrays, there have been numerous studies devoted to each facet of the technology, from image processing and data acquisition to end stage processing of gene networks. The success of each data processing step is highly dependent on the steps preceding it. This paper focuses on normalizing the signal by separating and removing the technological signal from the biological signal.

The rest of this paper is organized as follows: an overview of the aCGH microarray technology, the normalization algorithm (called "Smooth2D"), results, and a discussion of the applicability of the methods.

Overview

Pin tip (sometimes called print tip) microarray technology was invented in the early 1990s (ref). The technology has grown tremendously and now there are various flavors of probes and target elements. Target elements can include genes, oligonucleotides, or bacterial artificial chromosomes and new microarrays chips can now contain on the order of a hundred thousand probes. Since the technology is maturing, the cost of analyzing a sample has been steadily decreasing so experiments are now being performed on hundreds of samples, rather than just a handful. Due to the technology the signal obtained is a combination of the biological signal and the signal due to technology. The goal of this paper is to isolate the biological signal by removing the technological signal. The technological signal is composed of three major components: 1) signal due to the intensity of each scanning channel, 2) signal due to spatial location , and 3) signal due to the imaging technology. In our "Smooth2D" process we remove each signal sequentially thus isolating the biological signal. Although our results can be applied to any print tip microarray setting, our examples will focus mainly on Roswell Park Cancer Institute's (RPCI) array based Comparative Genomic Hybridization (aCGH) facility.

Array based Comparative Genomic Hybridization (aCGH) technology is similar to cDNA arrays and is an extension from conventional CGH that is used to identify and quantify DNA copy number changes across the genome in a single experiment. The advantages of aCGH include high-resolution and high-throuput measurement capability, furthermore, more quantitative analysis of the genomic aberrations.

In aCGH technology, the array elements or targets are laid out on a glass slide and are probed with dye labeled samples. In bacterial artificial chromosome (BAC) aCGH technology the target DNA elements are cloned in a bacterial culture and then physically arrayed in a two-dimensional grid on a chemically modified glass slide.

After creation of the chip, differentially labelled total genomic DNA from a "test" and a "reference" cell population are cohybridized to the BAC clones using blocking DNA (Cot-1) to suppress signals from repetitive sequences. After hybridization, a GenePix Axxon scanner generates two images of the chip at the wavelengths of light corresponding to the two dyes. The images are processed to generate a single number corresponding to each sample for each spot on the chip. For the RPCI facilities, Genepix is currently used to perform the image processing. The resulting ratio of the fluorescent intensities at a location on the chromosomes is approximately proportional to the ratio of the copy numbers of the corresponding DNA sequences in the test and reference genomes.

The data we analyzed is contained in a spreadsheet and gives the intensity readings from the Cy3- and Cy5-labeled probes for each spot, as produced by the image processing software. We let P_1 denote the intensity of the Probe 1 signal that was Cy5-labeled for a specific spot. For each spot i, we let $log(\frac{P_{1i}}{P_{2i}})$ denote the differential log expression between the two probes for that spot. For the Nowak array facility the sample labelled with Cy5 represents a collection of mRNA from a pool of normal subjects. Hence, in studying various cancerous tumors the standard $log(\frac{P_{1i}}{P_{2i}}) \equiv M_i$ values can be interpreted as the logarithm of tumor to control values $(log2T_i/C_i)$ for probe i. For each spot we consider the $log_2(\frac{P_{1i}}{P_{2i}}) \equiv M_i$ as the differential log expression between the two probes for each spot i. The M_i value is generally considered as the "signal" for probe i in aCGH experiments. Because of the nature of the technology the M_i value can be considered as a sum of two components, the biological signal which contains the information concerning the two populations under examination, and the technological signal which is present merely due to experimental conditions. The following will describe the "Smooth2D" normalizing process that removes the technological signal present in the M_i values.

The Smooth2D Algorithm

Below is a schematic detailing the Smooth2D process. The Smooth2D process consists of each of the following steps applied sequentially to the log ratios. Let M denote the vector of M_i probe values for a specific chip. Let

Method Number	Normalization Method	Description
1	Global Loess	Fit M according to intensity values A
2	Spatial Kernel Smoother	Fit M according to neighboring M values
3	Spotting Process	Fit M according to its pin, plate, plate row, and plate column

Figure 1: **Table 1**: Table showing each step in the "Smooth2D" normalization process for aCGH microarray experiments. Each step is detailed in the section "Smooth2D Description"

A denote the vector of A_i values, where:

$$A_i \equiv log_2(P_{1i} \times P_{2i}) = log_2(T_i \times C_i).$$

A is the logarithm product for the two channels in an aCGH microarray experiment.

Global Loess Step

The M_i values from the scanner represent the input values to Smooth2D algorithm. Figure 2 represents the input data for Smooth2D. It is important to examine the input data, namely the M values on both a ranked scale and the original scale. This dataset will be used throughout this paper to demonstrate the Smooth2D process. This data represents a normal male versus a normal female hybridization. Since the normal state for human cells is diploid, in aCGH experiments the same copy number exists in different normal samples. When comparing a normal male against a normal female, the only difference should be a 1:2 ratio for X chromosome sites mimicking a single copy deletion. By examining the mean and the standard deviation for the M values located on the X chromosome in this experiment we will have a tool to gauge our success in removing the technological signal.

The first step is to run a loess smoother on the M_i values using the $log2P1 \times P2$ values as the explanatory variables. This is considered a global operation since the entire set of probes from a chip is used in the loess fitting function. The span or α value used for the loess fit is the R default value of .75. Fitting is by (weighted) least squares. The result from this operation is a set of fitted values, $GL(M_i)$ fit according to the $log2P1_i \times P2_i$ values. The fitted values, $GL(M_i)$, represent the bias of the log ratio M according to the log product, A. By

subtracting the fitted values $GL(M_i)$ we account for this technological bias.

The next step in the algorithm is to compute the residual:

$$M_i' = M_i - GL(M_i)\ i = 1, 2, 3, ..., 17,948$$

M_i' represents the signal after removing the technological signal due to the product of the intensities from the two channels. The goal in the next step is to remove the spatial bias present in the aCGH technology. Figures **??** and 4 shows the results from the global loess procedure for a specific data set.

Spatial Kernel Smoother Step

The next step in the algoritm is to remove the technological noise due to the spatial location of the chip. It is reasonable to expect nearby spots to be correlated with each other due to the reagents process and the hybridization process in microarray technology. The goal in this step is to accurately determine the spatial pattern present in the chip and thus remove it. The representation of M values and the ranked M values based on their location on the chip image is a good way to view the spatial bias present on a chip. These plots will be referred to as M-XY plots. Figure 6 (a) shows the M-XY plot for ranked values of M'.

Before the spatial kernel smoothing step, we will run a circular binary segmentation (CBS) algorithm on M' to cluster the M' values into segments of estimated equal copy number according to their location on the genome (ref). After CBS, each probe i is a member of a specific segment where we will denote $CBS(M_i')$ as the $log2\,T/C$ group mean for the segment containing probe i. We compute the residuals from the CBS as, $M' - CBS(M_i')$. From these residuals, we perform a two-dimensional kernel density smoothing on $M' - CBS(M_i')$. Note, by subtracting the CBS group means we are, in a sense, removing the genome/biological signal to ensure that our kernel density estimate of the spatial signal has a minimal amount of biological signal contamination. The two dimensional kernel density smoother is performed using the function "smooth.2d" in the software package R. The smoothing parameter is chosen via a cross validation procedure. Namely, a random subset is removed from the data and the surface is fit. After fitting a surface, the sum of squares for the random subset is computed. The value of λ that yields the smallest sum of squares is used as the optimal value in the kernel smoothing algorithm.

After determining the two dimensional kernel density smoother, we compute:

$$M_i'' = M_i' - KS\left(M_i' - CBS(M_i')\right)$$

M_i'' represents the $log2T/C$ for probe i after the signal due to intensity and signal due to the spatial location have been removed.

Spotting Process Step

At this point, we remove the technological signal due to the spotting of the chip. Sellers,Miecznikowski,Eddy introduce this problem in microarrays and Miecznikowski 2006 demonstrates this problem and the solution on aCGH microarray chips. Similar to determining the kernel smoothed spatial surface, we determine the pin, plate, plate row and plate column effects for $M_i'' - CBS(M_i'')$ where the CBS segment mean is subtracted to isolate the technological signal. Hence, by successively subtracting the median for the spotting pin for probe i, 384 well plate for probe i, plate row and plate column for probe i, we obtain the final M_i''' values representing the remaining signal after removing the estimated effects due to the intensity, the spatial location, and spotting procedure.

Conclusion

Through a series of sequential steps we have developed an algorithm called "Smooth2D" which normalizes log2 T/C values from an aCGH based microarray platform. This normalization occurs according to 3 major sources of technological signal. The technological signal due to the intensity effects is removed first. Secondly the signal due to the spatial location on the microarray chip is accounted for and removed. Lastly the signal due to the spotting process is removed. Each of these sources of signal is a well documented problem in aCGH literature (references). The novelty of this method is that it combines each of these normalization procedure into one stepwise procedure.

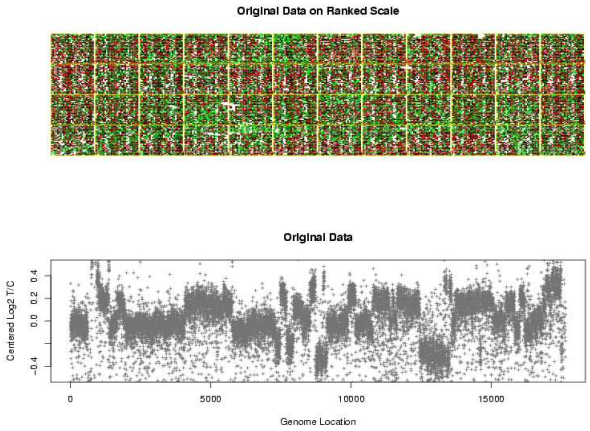


Figure 2: **Original Data**: Image of the chip and genome for the raw original data

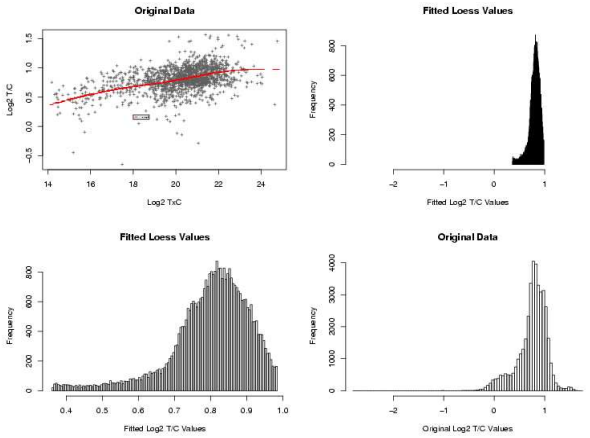


Figure 3: **Global Loess**: Summary images of Global Loess on a test sample

Results

The algorithm is best described in a sequence of figures showing each step in the process to obtaining the final M_i values.

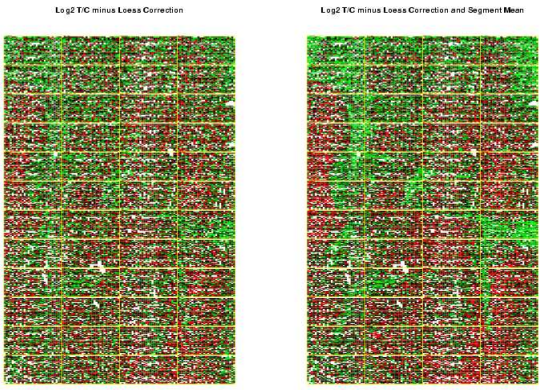
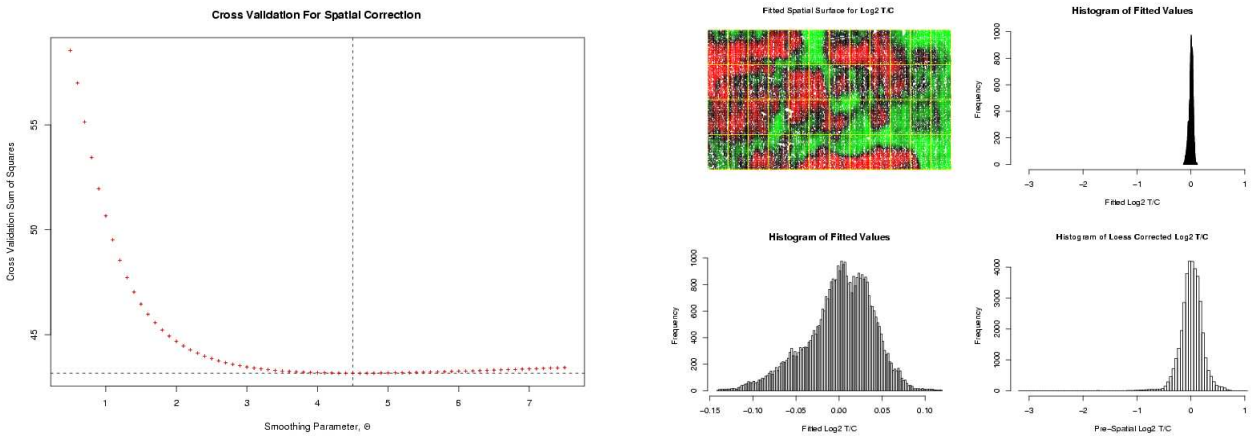
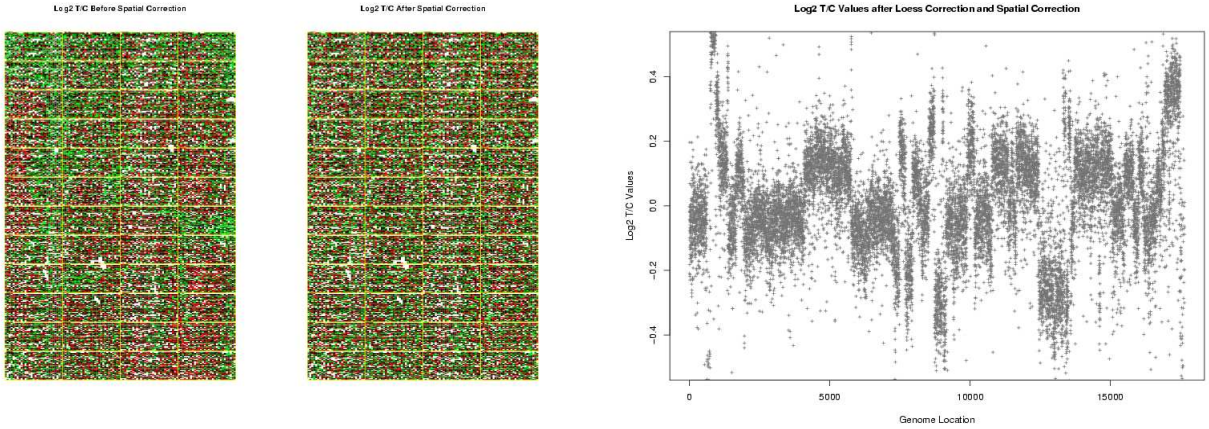


Figure 4: **Global Loess Before and After**: Chip images before and after global loess



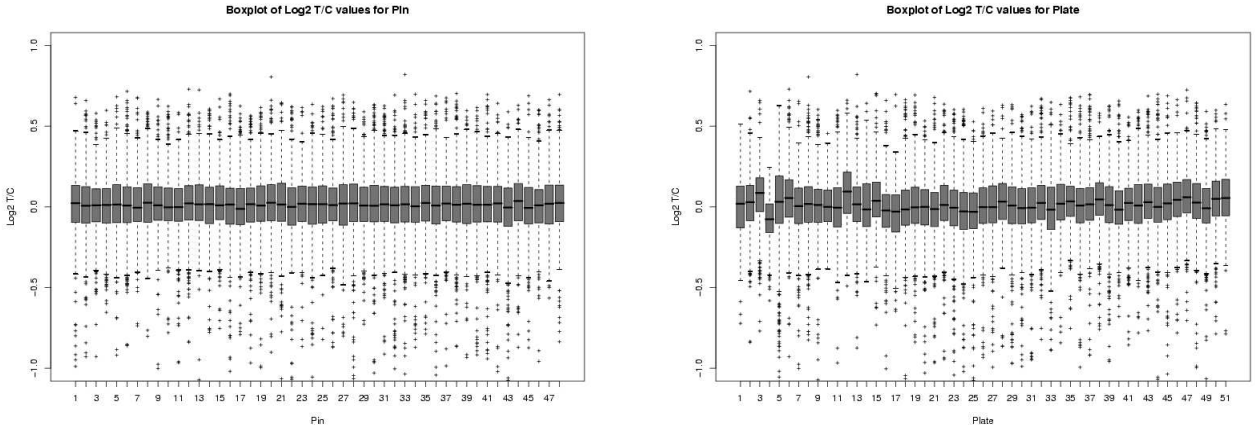
(a) (b)

Figure 5: **Spatial Correction**: Choosing a smoothing parameter that yields the minimum sum of squares error. The fitted spatial surface for log2 ratios.



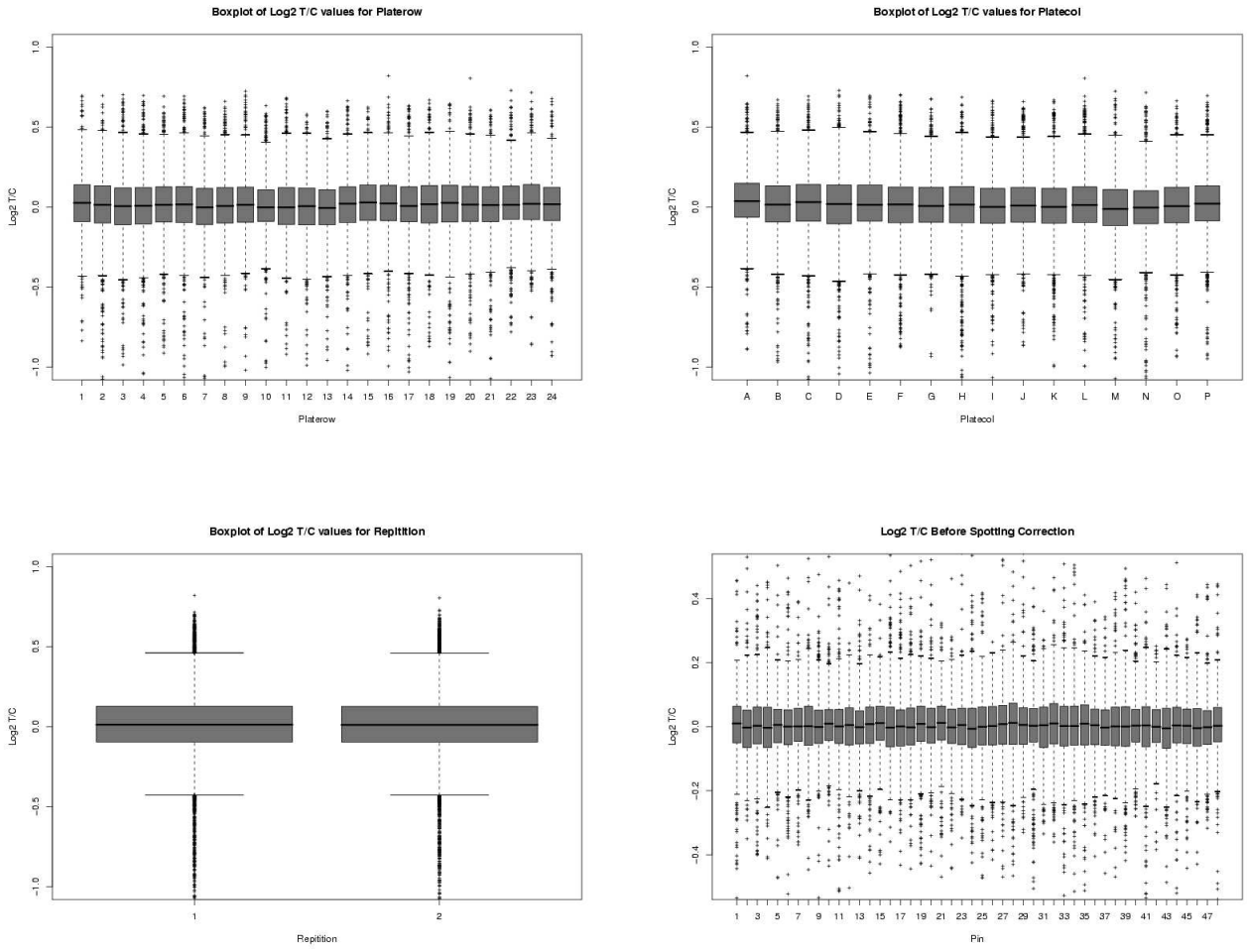
(a) (b)

Figure 6: **Spatial Correction Before and After**: . The chip images before and after spatial correction. The genome plot of Log2 T/C after spatial correction and loess correction.



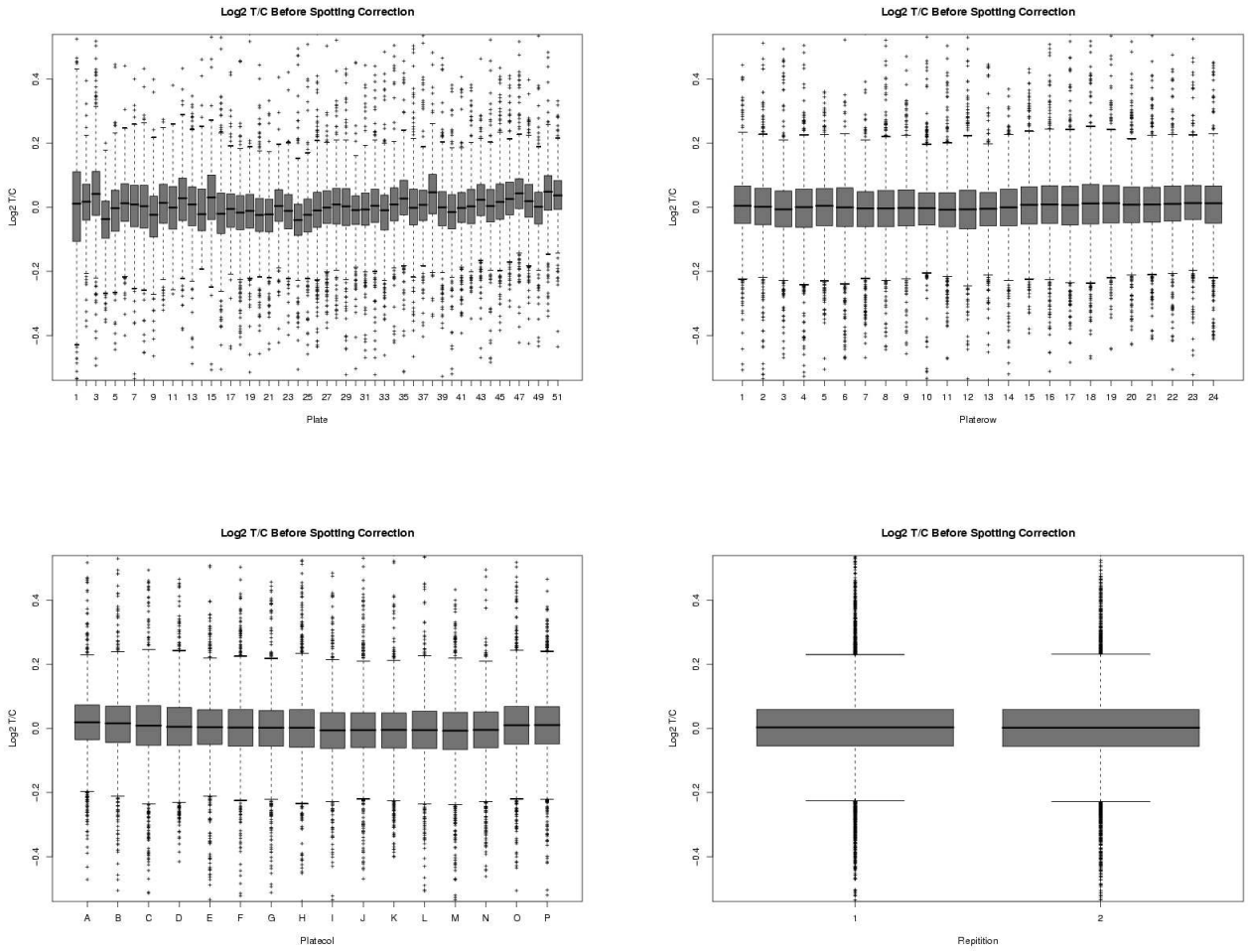
(a) (b)

Figure 7: **Spotting Process**: The boxplot showing the distribution for each pin and for each plate.



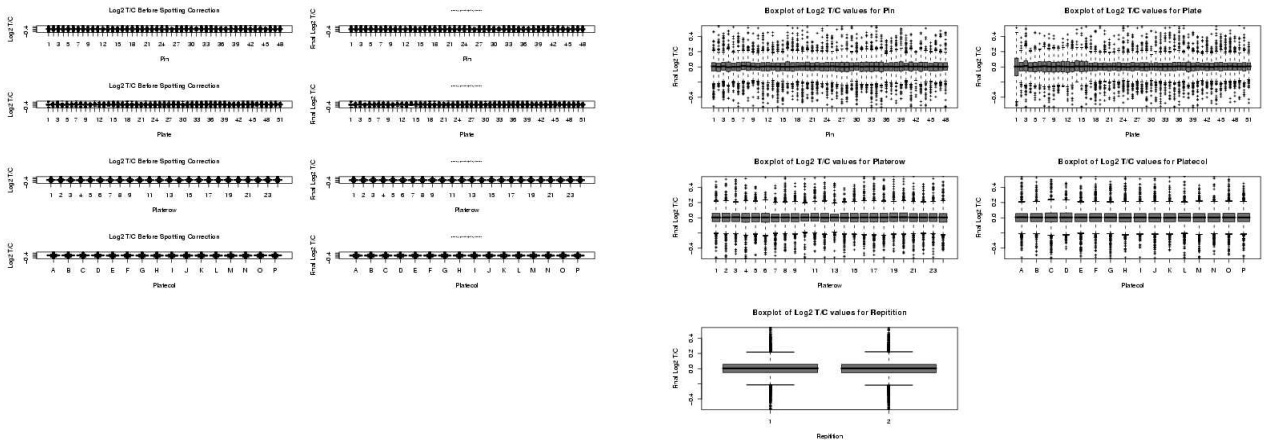
(a) (b)

Figure 8: **Spotting Process**: The boxplot showing the distribution for each plate row and plate column and for each repitition.



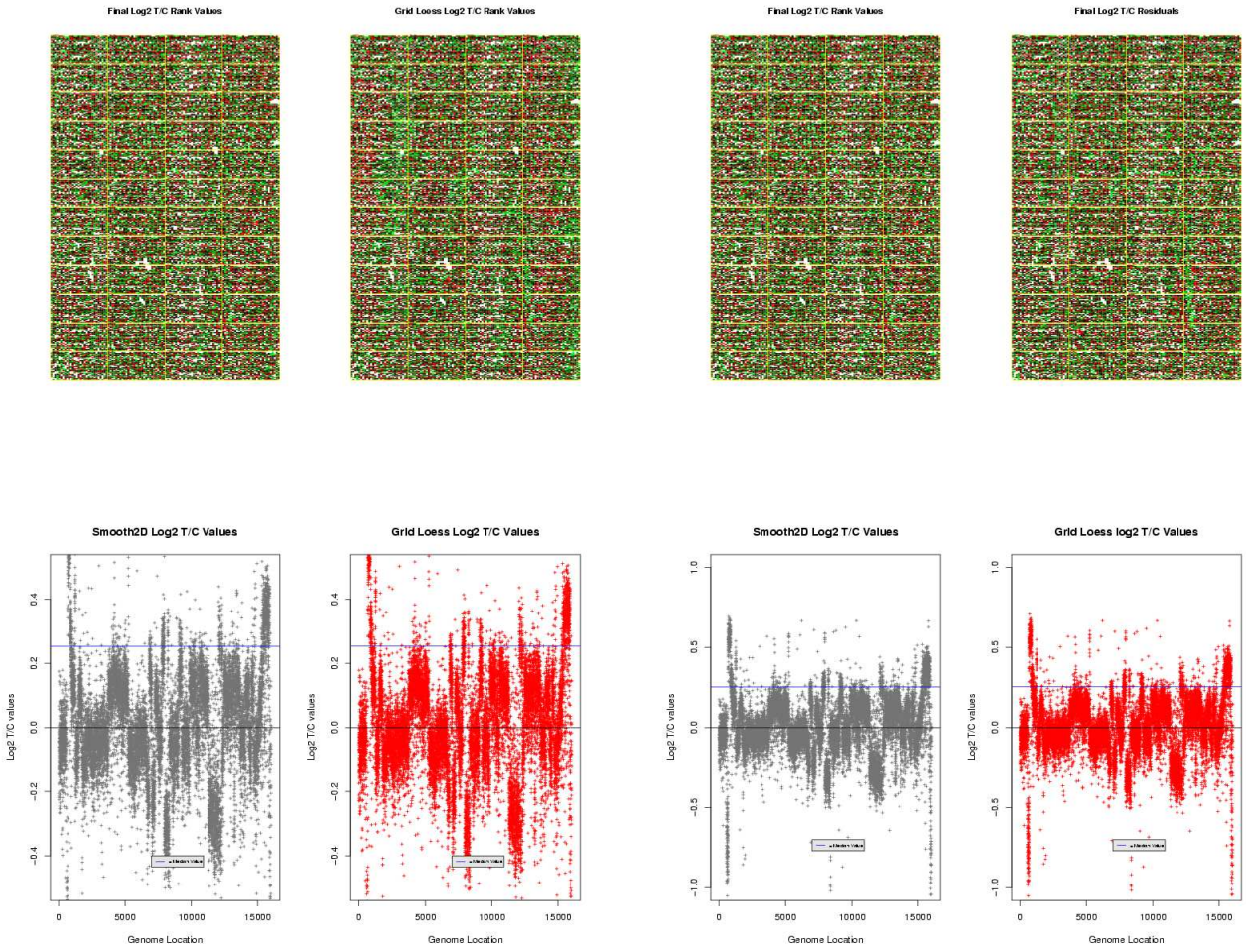
(a) (b)

Figure 9: **BF title**: Still not sure what to put here (a). Still not sure for (b)



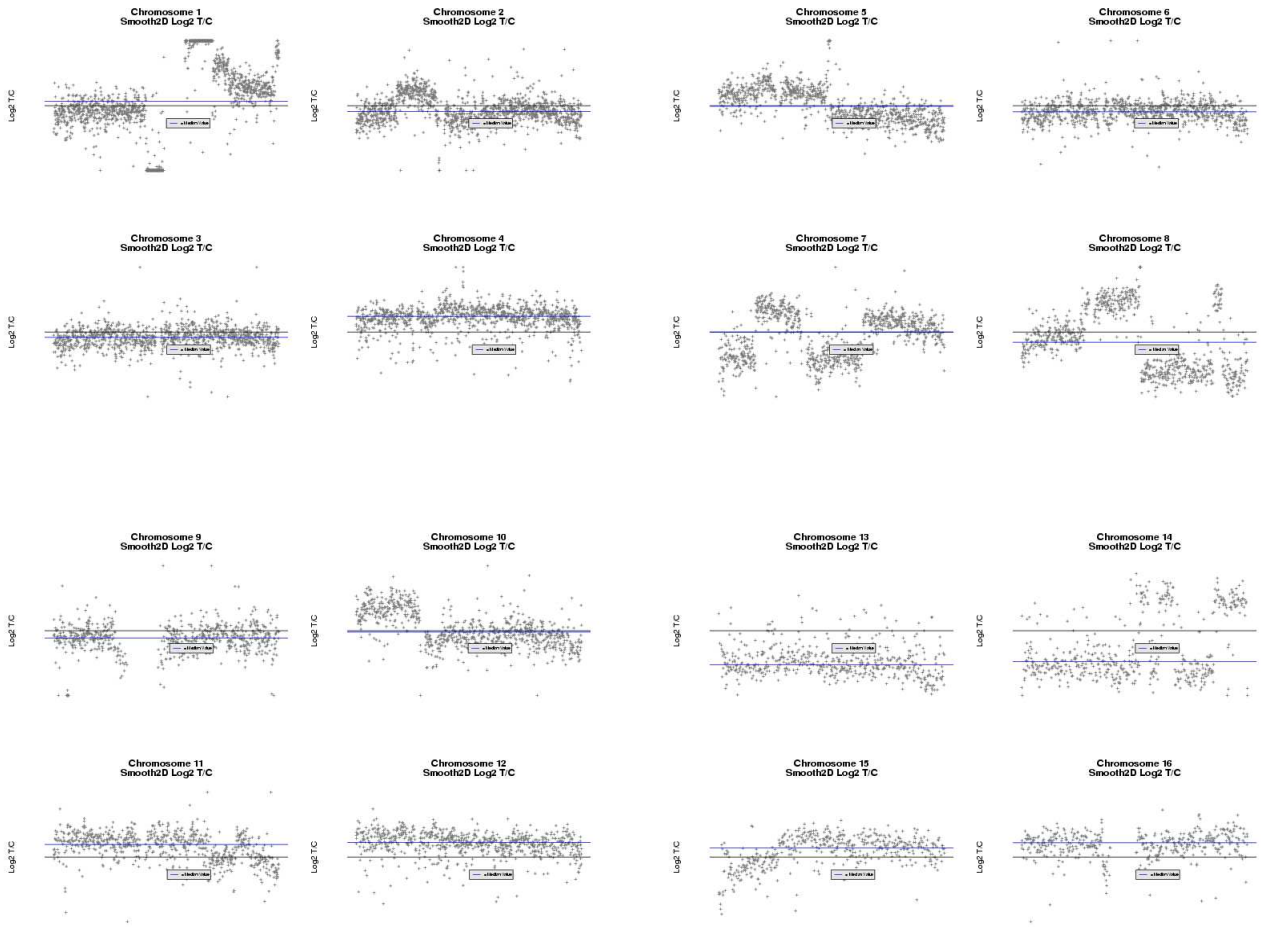
(a) (b)

Figure 10: **BF title**: Still not sure what to put here (a). Still not sure for (b)



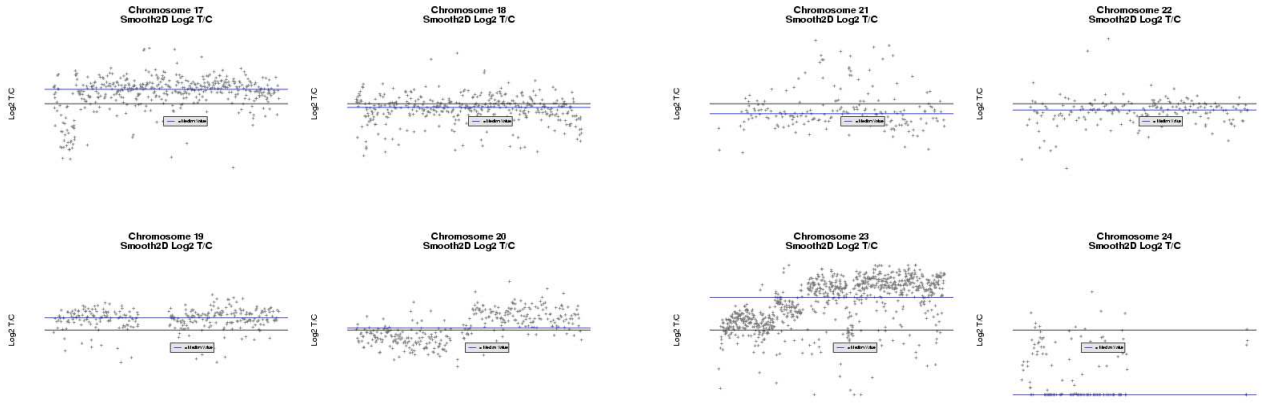
(a) (b)

Figure 11: **Final Results**: The final log2 T/C values from Smooth2d. Compare these results with the grid loess procedure.



(a) (b)

Figure 12: **Genome Plot 1**: Genome plots for the final log2 T/C values from Smooth2D)



(a) (b)

Figure 13: **Genome Plot 2** : Genome plots for the final log2 T/C values from Smooth2D

References