Universitas Ahmad Dahlan
Bulletin of Applied Mathematics and Mathematics Education
Volume 5, Number 1, April 2025 pp. 81-96
https://dx.doi.org/10.12928/bamme.v5i1.12539
e-ISSN 2776-1029, p-ISSN 2776-1002

Pipeline on microarray data analysis: Pre-processing

Rohmatul Fajriyah[1]*, Noodchanath Kongchouy[2], Wanvisa Saisanan Na Ayudhaya[3], Rahmadi Yotenka[1], Ghiffari Ahnaf Danarwindu[1]

[1]Universitas Islam Indonesia, Jl. Kaliurang km 14.5 Sleman, Yogyakarta 55584, Indonesia
[2]Prince of Songkla University, Hat Yai Campus, Hat Yai, Songkhla 90112, Thailand
[3]Walailak University, 222 Thaiburi Sub-District, Thasala District, Nakhon Si Thammarat 80160, Thailand

*Corresponding e-mail: rfajriyah@uii.ac.id

ARTICLE INFO

Article History
Received 8 January 2025
Revised 2 July 2025
Accepted 14 July 2025

Keywords
Affymetrix
Bioinformatics
Microarray
Pre-processing

How to cite this article:
Fajriyah, R., Kongchouy, N., Ayudhaya, W. A. N., Yotenka, R., & Danarwindu, G. A. (2025) Pipeline on microarray data analysis: Pre-processing. *Bulletin of Applied Mathematics and Mathematics Education*, *5*(1), 81-96.

ABSTRACT

Bioinformatics is blooming and its data are store in some repository offline and or online. Yet some basic concepts are not fully disseminated. The paper intends to provide the reader with a review of one important concept in the pipeline bioinformatics data analysis of microarray, pre-processing. In pre-processing, there are four steps, background correction, normalization, probe correction and summarization. Each step consists of several methods. Differ from the previous works, this paper describes each method in each four steps of pre-processing. This is done to give a better understanding on how it works theoretically. We focused on microarray data from Affymetrix platform with single-color chip.

Introduction

Bioinformatics is an interdisciplinary field of biology, medicine, pharmacy, statistics, mathematics, computer or information engineering, chemistry, and physics. It has proliferated since 1990 through the microarray technology. There are many definitions of microarray technology. Microarray technology is a powerful tool seen as a general laboratory approach that involves binding thousands to millions of known nucleic acid fragments (probes) to an immobilized solid surface (chip).

The technology can measure thousands of gene expressions and detect single-nucleotide polymorphisms (SNPs) or specific DNA sequences in a single experiment. Microarray technology can be applied to medical diagnostics, drug discovery and development, and toxicogenomics. Microarrays has a different format, namely DNA, protein, cell, and tissue microarray

In Affymetrix design, a gene is represented on the array by a series of oligonucleotide probes. Probes are chosen based on specific nucleotide repositories. Each probe consists of a perfect match

(PM) oligonucleotide and a mismatch (MM) oligonucleotide (Miranda & Bringas, 2008).

Bioinformatics data is produced through microarray experiments, where the target is hybridized to the probe in the Affymetrix chips. It measures the presence of genes in the samples by their intensity values.

Grant et al. (2007) explained that microarray experiments require careful planning and choice of experimental design, data pre-processing, analysis tools, and data documentation to maximize the data generated and reproducibility. Once the experiment is finished, the raw data intensities will be available and must be handled appropriately.

For researchers who have already worked in bioinformatics, dealing with microarray data is accessible. There will be a hardship period for someone new to understanding how to deal with these data. Olson (2006) stated some reasons—the familiarity of analysis software and microarray data analysis have primarily been treated as separate steps.

Based on them, then first the basic knowledge of molecular biology, statistics, and programming and familiarity with the computer working environment are necessary. Second, understanding the object (data) in the research field, particularly for researchers from mathematics, statistics, and computer sciences, is a key to open more fruitful collaborative research with experts in this field. Third, understanding the pipeline in microarray data analysis is vital for reliable data analysis results.

Third, understanding the pipeline in microarray data analysis is vital for reliable data analysis results. The pipeline in microarray data analysis can be varied. However, one can refer to Figure 1 (Serin, 2011), Figure 2 (Microarray Galaxy User's Guide, 2023), and Figure 3 (Federico et al., 2022), namely quality control, pre-processing, filtering, data analysis and visualization, and functional analysis. An end-to-end workflow of Affymetrix microarray data analysis in R have been explained by Klaus and Reisenauer (2018).

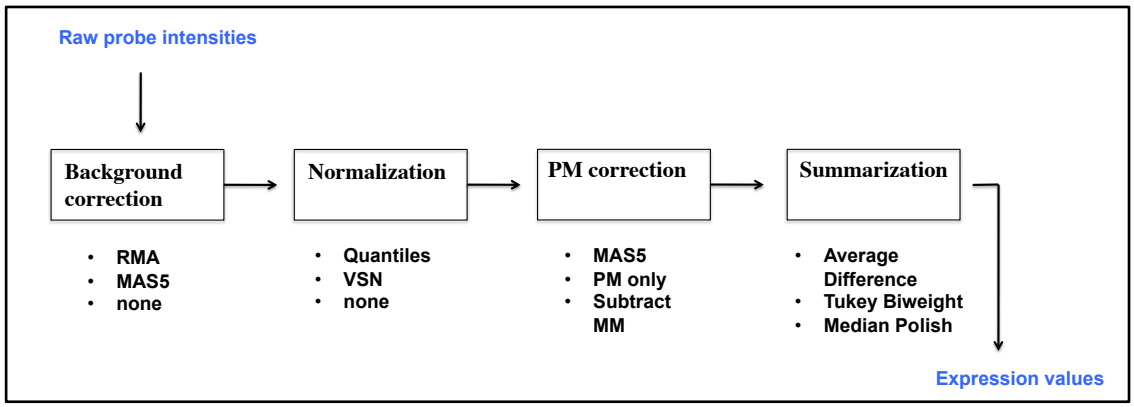

Figure 1. Pre-processing microarray data (Serin, 2011)

This paper is one of the publications about basic concept in bioinformatics series, where the initial paper was about an overview of microarray technologies (Fajriyah R. , 2021).

Pre-processing

Pre-processing is the first and most crucial step because it will remove unwanted variations (e.g., noise, artifacts, and systematics biases) in raw microarray data (TechMedBuddy, 2023).Pre-processing ensures that the data set is valid and reliable such that the conclusion from the analysis is trustworthy. Some steps are applied in the pre-processing of microarray data. They are background correction, normalization, probe correction, and summarization.

10.12928/bamme.v5i1.12539

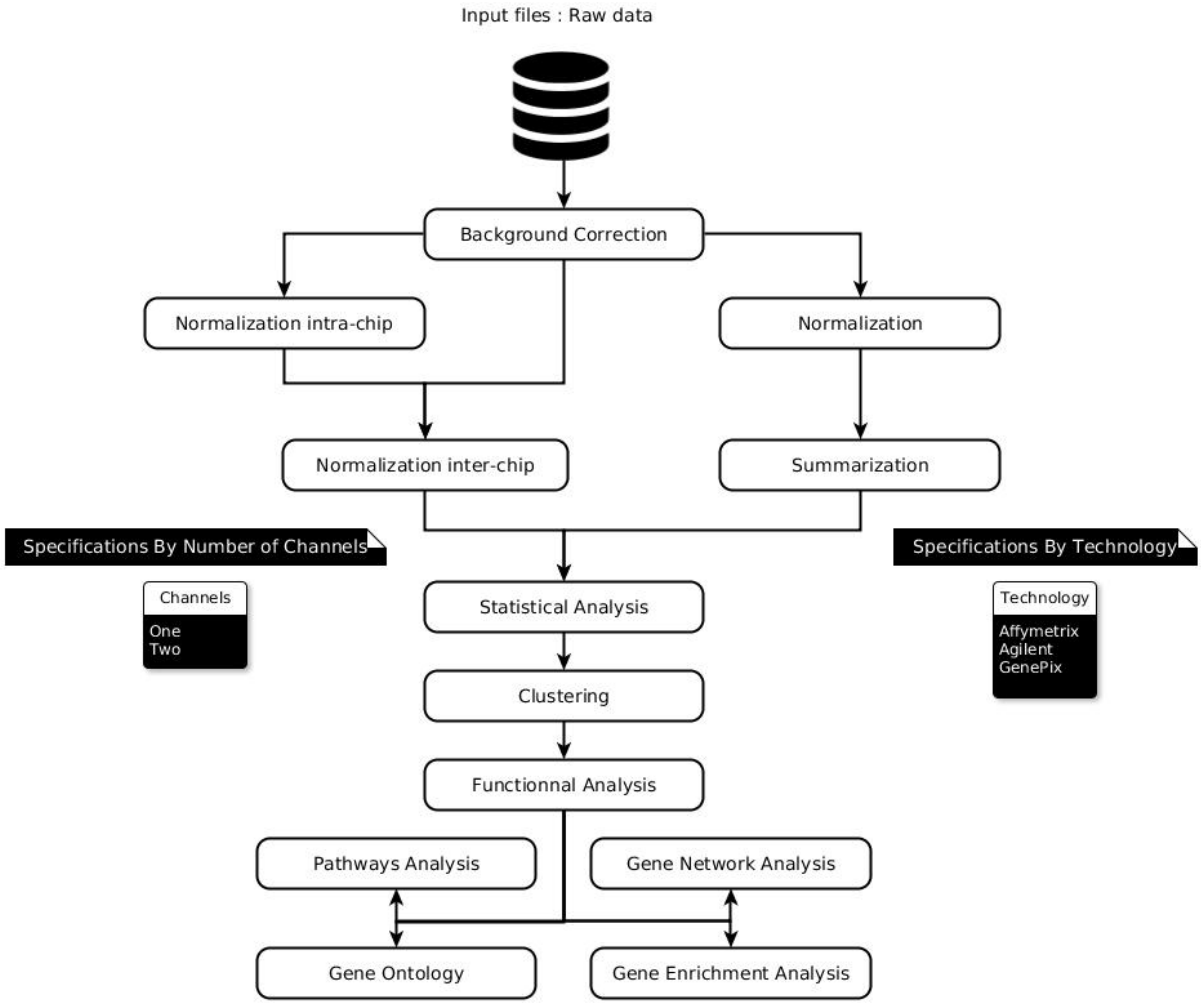

Figure 2. Microarray pipeline (Microarray Galaxy User's Guide, 2023)

Different methods have been developed for microarray Affymetrix data in terms of pre-processing. Some of them are mentioned in Kuyuk (2017), Munster et al. (2018), and Visentin et al. (2022) which are also described here. Concerning which pre-processing method is to be selected, one can follow the suggestion from other researchers, for example, by implementing the procedure from Dozmorov et al. (2010) or following Klaus and Reisenauer (2018).

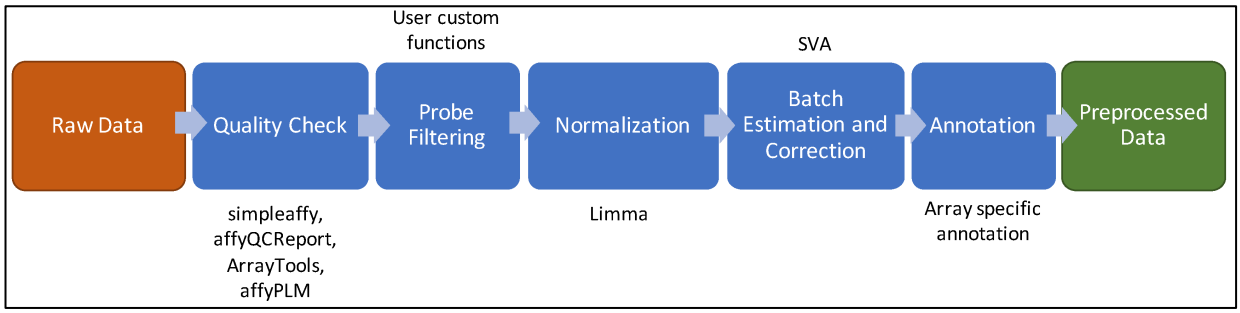

Figure 3. Microarray pipeline (Federico et al., 2022)

The challenge of choosing which pre-processing methods should be implemented for the data at hand, requires the understanding of the chip's platform. Because its platform has different

design. Moreover, missing values in bioinformatics data plays a key role as well, it will affect the data analysis results.

Background correction

Background correction is used to provide the actual intensity of the probe in the sample. In this case, it is modeled that the available intensity based on the measurement from the microarray experiment contains the non-specific noise (error). Statistically, it can be additive, multiplicative or both (hybrid) to model the error. Some background correction methods for microarray data and how they work are as follows.

MAS5.0

According to the Statistical Algorithm Description Document from (Affymetrix, 2002), the implemented background correction in MAS5.0 is as follows.

a. Chip is divided into k square areas ($k = 1,2, \dots ,16$).
   a.1. The smallest 2% is chosen as the background for area k, B_k.
   a.2. The standard deviation of that 2% is chosen as the noise in the area k, N_k.
b. The background adjustment in cell (x, y) is a weighted mean B_k where the weight depends on the distance between (x, y) and central point of the area b(x, y).
c. Compute the noise adjustment $n(x, y)$ as the background adjustment b(x, y).
d. The intensity after background adjusted is

$$A(x, y) = max(I(x, y) - b(x, y), 0.5 * n(x, y)) \quad (1)$$

RMA

The RMA method was first introduced (Irizarry et al., 2003) and it has been transformed into a Bioconductor affy package by (Gautier et al., 2004). The affy package is built based on (Bolstad, 2004) dissertation. RMA algorithms implement the simple heuristic estimator based on the histogram smoothing process from observed intensity values, and the distribution is divided by its mode.

The RMA model can be written as

$$S = X + Y \quad (2)$$

that S is the corrected intensity, X is the observed intensity based on the microarray experiment, and Y is the background measurement error. X is assumed to be exponentially distributed, and Y is truncated at 0 normally distributed. $E(X|S = s)$ gives the corrected background intensity.

GCRMA

The method is the extension of RMA, where the Guanine and Cytosine (GC) content is included in the model. In the GCRMA, the signal intensity is modeled by

$$PM = O_{PM} + N_{PM} + S \quad (3)$$

$$MM = O_{MM} + N_{MM} + \phi S \quad (4)$$

that O is optical noise, N is the background noise from non-specific binding, and S is a signal that degenerates the specific binding between the probe and its target. The parameter ϕ reflects that the MM signal contains some specific signal for several pair probes (Wu et al., 2004; Wu, 2009).

The background components $log(N_{PM})$ and $log(N_{MM})$ are assumed normally bivariate with means $\mu_{pm} = h(\alpha_{PM})$ and $\mu_{mm} = h(\alpha_{MM})$ where h is a smoothing function and α is a probe affinity defined (Naef & Magnasco, 2003) as

$$ln\left(\frac{B}{M}\right) = \sum_{k=1}^{25} \sum_{l \in (a,c,g,T)}^{\alpha} S_{lk} A_{lk} \quad (5)$$

10.12928/bamme.v5i1.12539

where B is a raw probe intensity value, M is a median of array intensities, l is a nucleotide index, k is the position of l in the probe, S is a boolean variable, and A is nucleotide affinity. Gharaibeh et al. (2008) have done the exact modeling with a dinucleotide. Their model is called GCRMA-NN, and its performance is better than GCRMA.

GCRMA gives two options in the background correction. The first is used the pre-computation α (it is called reference affinity) based on the non-specific binding (NSB) experiment of the GCRMA author, and the second one is where α is computed directly from the data (it is called local affinity).

Normexp

The Normexp background correction method is built based on the convolution model as follows

$$X = B + S \quad (6)$$

where X is the observed gene intensity value from the microarray experiment, B is the measurement error, and S is the real gene intensity value (Ritchie et al., 2007).

Table 1. Background correction methods for microarray data

Method	Key Features	Pros	Cons	References
MAS5.0 (Microarray Suite 5.0)	Affymetrix's default method; uses a statistical model to subtract background and scale data.	- Simple and widely used - Includes detection p-values - Good for single-array processing	- Less robust to noise - Can overcorrect low-intensity signals - Lower accuracy than RMA or GCRMA	Affymetrix (2002); Li & Wong (2001a, 2001b)
RMA (Robust Multi-array Average)	Background correction + quantile normalization + summarization via median polish.	- High reproducibility - Robust against outliers - No need for mismatch (MM) probes - Good for multi-array comparisons	- May underestimate low-intensity signals - Ignores MM probes, potentially losing specific information	Irizarry et al. (2003); Gautier et al. (2004); Bolstad (2004)
GCRMA (GC-content adjusted RMA)	Extension of RMA, adjusts for sequence-specific binding affinity using GC content.	- Improves background estimation using probe sequence - Better performance on low-intensity probes - Captures probe-specific biases	- Computationally more intensive - May overfit on small datasets - Requires probe sequence information	Wu et al. (2004); Wu (2009)
Modified GCRMA	Variant of GCRMA using modified parameter settings or models to improve accuracy.	- Better signal estimation in certain scenarios (e.g., low expression genes) - More flexible	- Not a standard method; requires careful tuning and validation	Gharaibeh et al. (2008); Wu (2009)
Normexp	Uses convolution model of normal + exponential distribution for background; often combined with offset.	- Smooth correction, especially for two-color arrays - Good theoretical foundation - Works well with Illumina arrays	- Assumes specific distribution shapes - Choice of offset affects results - Less used with Affymetrix data	Ritchie et al. (2007); Silver & Ritchie (2009)

In estimating S, the maximum likelihood (MLE) method has been used in (Silver JD, Ritchie ME, 2009). This method has been wrapped in the limma R-Bioconductor package. According to

(Silver JD, Ritchie ME, 2009) the normexp method is an adaptation method from (Irizarry et al., 2003). The summary of all background correction methods can be seen on Table 1.

Normalization

Normalization is a process to control technical variability between assays and preserve the biological variation to get the accurate analysis (Cheng et al., 2016). In other words, to eliminate the unwanted nonbiological variation that possibly exists in the microarray. The no-biological variations are, for instance, in Pelz et al. (2008), dyed and scanner setting for the microarray image.

We know that the chemical compound in dyes has different adhesive levels on the surface of objects, as in the microarray slide (J. Yang & Thorne, 2002). Therefore, gen expression measures via microarray technology will have the same situation. To normalize data microarray, keep in mind about array type, design of experiment, assumptions about data (e.g., 'genes are not expected to be differentially expressed in the test group relative to controls'), and packages that will be used to analyze the data.

There are two types of normalization, namely within array normalization: median, loess, print tip loess, composite, control, and robust spline; and between array normalization: scale, quantile, or cyclic loess.

Fujita et al. (2006) mentioned several normalization methods, such as Loess Regression, Splines Smoothing, Wavelets Smoothing, Kernel Regression, and Support Vector Regression. The VSN is mentioned in Barbacioru et al. (2006). New method, in which it was built and based on personalized-medicine workflow, can also be implemented to Affymetrix arrays (Piccolo et al., 2008).

Global/Affymetrix normalization

Affymetrix normalization method is often used in the beginning of the availability of microarray data based on microarray experiments. The average difference (AD) of the Affymetrix expression index is the foundation of this method, which is the average difference between perfect match (PM) and mismatch (MM) probes.

For each array, the trimmed mean from all AD probes is computed, and the AD factor normalization is determined by its ratio average or AD target average. The Affymetrix normalization assumes a linear relationship between arrays.

$$x_{ij} = \mathrm{AD}_{ij} * \frac{T}{M_i} \tag{7}$$

where T is the target intensity (500 or the median of all arrays) and M_i is summary statistics of AvgDiff in i-th array. It can be mean, median or (often) trimmed mean. Suppose a probe set has k probes then the AD is computed by

$$\mathrm{AD}_{ij} = \frac{1}{k}\sum_{l=1}^{k}\left(PM_{ijl} - MM_{ijl}\right) \tag{8}$$

Global Rank-Invariant

Global rank-invariant set normalization (GRSN) method is a generalized gene rank-invariant from Li, C. and Wong (2001a; 2001b) and Baans et al. (2019). In using this method, (Pelz et al., 2008) select a set of genes, a *globally rank-invariant set of endogenous genes*, which will be used to normalize all samples in the data set. The chosen gene is assumed to be expressed consistently in all available samples and has the same ranking in each sample. Although we will only describe the GRSN, it is worth mentioning that similar works has been proposed by (Wright Muelas et al., 2019), by using the Gini index.

10.12928/bamme.v5i1.12539

The GRSN procedures are as follows.

a. Do the pre-processing method in background correction with MAS5.0, RMA, or dChip.
b. Convert the results from the first step to the log2 scale.
c. Build a matrix where the row represents the transcript/probe and the column represents the sample.
d. Build a rank matrix with an equal dimension to the data matrix. All column is ranked based on their expression value.
e. Compute the variance from each row of the rank matrix.
f. The reference value for each transcript is the trimmed mean from all samples.
g. Other transcripts are the Global Rank-invariant Set (GRiS).
h. Repeat four times steps 4-7, where we exclude the 25% difference between transcript size and the chosen invariant transcript.
i. Remove the row where it has the highest variance of rank from the data matrix
j. For each sample, produce the calibration curve by comparing the reference value of GRiS to that of GRiS in that sample.
k. Use the lowess method to smooth the calibration curve for each sample.
l. Use the smoothed calibration curve to apply an intensity-dependent adjustment in all transcripts in the sample.
m. Repeat steps (j) – (l).

Contrast normalization

This method was introduced by (Astrand, 2003) and used to normalize microarray data by implementing a smoothing curve. The method is used to normalize PM and MM intensities, PM-MM, or others as long as they come from feature intensities.

Suppose x_{ij} is an intensity value of i-th probe in j-th array. Steps in contrast normalization are

a. Compute y_{ij} the log2 transformation of x_{ij}.
b. For each array I, compute its median and first quartile Q1.
c. Compute its local contrast, $\Delta_{ikj} = y_{ij} - y_{kj}$.
d. Choose one array as a reference array (the j-th array).
e. Compute the ratio of local and reference contrast array, $r_{ikj} = \frac{\Delta_{ikj}}{\Delta_{ikj^*}}$.
f. Do scaling contrast, s_j, where $s_j = \text{median}_{i,k}\left(\frac{\Delta_{ikj}}{\Delta_{ikj^*}}\right)$ and compute $\hat{y}_{ij} = \frac{y_{ij}}{s_j}$.
g. The normalization intensities are computed by inverse transformation formula of $\hat{x}_{ij} = 2^{\hat{y}_{ij}}$.

Quantile normalization

Bolstad et al. (2003), Bolstad et al. (2003), and Bolstad (2004) proposed the quantile normalization method for single-color array (Barbacioru et al., 2006) for two-color cDNA arrays. The method ensures that the intensities have equal empirical distribution across arrays.

The quantile normalization method ranks the intensity values per array/sample. Further, compute the reference value, for instance, its mean, per gene/probe. Replace the intensity value with its average. Order as the original value. These ordered values are the normalized intensities. Bolstad et al. (2003) explained in detail as follows.

Let $q_k = (q_{k1}, \dots, q_{kn})$ a kth quantile vector for all array n, $k = 1, \dots, p$ and $d = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$ is

a diagonal unit. The projection q into d is defined as

$$proj_d\, q_k = \left(\frac{1}{n}\sum_{j=1}^{n} q_{kj}\, , \ldots , \frac{1}{n}\sum_{j=1}^{n} q_{kj} \right) \tag{9}$$

This caused each array to have the same distribution by taking the average of its quantile and substituting the value into the original value. For many arrays, the algorithm is as follows.

a. Let us have n array, and each array has length p. Make a matrix X which has pxn dimension where each array is its column.
b. Sorted out each column X such that we got a new matrix X_{sort}.
c. Compute the mean for each row in X_{sort} matrix. This mean value then becomes the value in this row and computes X'_{sort}.
d. The X normalized is computed by re-arranged each column from X_{sort} where the ordered as the original data matrix X.

Scale normalization

The normalization method standard from Affymetrix is the scaling method. The method is implemented on expression measurement on a set of probes. Bolstad et al. (2003) proposed the scaling method on the probe level. First, choose the baseline array, which is the array that contains the median value from all median intensities. Then, all array is normalized with this 'baseline' array as follows.

Let x_{base} be the intensity of the baseline array, and x_i be the intensity value on other arrays. Next, compute

$$\beta_i = \frac{\tilde{x}_{base}}{\tilde{x}_i} \tag{10}$$

where $\tilde{x}_i$ is the trimmed (2%) mean intensity. The intensity value on the array after normalization is

$$x'_i = \beta_i x_i \tag{11}$$

Bolstad et al. (2003) explained that we could implement the scaling algorithm by using a probe from the probe subset, and it is chosen with specific stability criteria. These probe sets can be used for normalization.

Y. Yang et al. (2002) propose the normalization method in two color arrays, and further explanation can be found in (Smyth & Speed, 2003). The idea is to log-ratio-scaled $\left(M = log_2 \left(\frac{x_i}{x_j} \right) \right)$ from the existing arrays such that they all have the same median absolute deviation (Hartemink et al., 2001).

Cyclic loess

Cyclic loess is a normalization method between array by using the loess approach, a local regression. Originally it was for two-channels cDNA arrays but then can be used in single-channel Affymetrix arrays. This method is used for normalization of Affymetrix arrays after the expression is summarized into the gene expression, by other methods such as RMA, MBEI or FARMS.

Cyclic loess normalization works by plotting the value of M versus A, where M is the difference expression in log scale, and A is the average of expression values (Dudoit et al., 2002).

For any two arrays i, j with probe x_{ki} and x_{kj} where $k = 1,12, \ldots, p$ representing the probes, compute

10.12928/bamme.v5i1.12539

$$M_k = log_2\left(\frac{x_{ki}}{x_{kj}}\right) \quad (12)$$

$$A_k = \frac{1}{2} log_2(x_{ki}x_{kj}) \quad (13)$$

The normalization curve is fitted into $an\ M$ versus A plot with loess (local regression method) (Cleveland, 1979; Cleveland & Devlin, 1988). The value based on the normalization curve is referred to as $\widehat{M}_k$, and the value after normalization is $M'_k = M_k - \widehat{M}_k$. The probe intensity after normalization is

$$x'_{ki} = 2^{A_k + \frac{M'_k}{2}} \quad (14)$$

$$x'_{kj} = 2^{A_k - \frac{M'_k}{2}} \quad (15)$$

Variance Stabilization Normalization (VSN)

The VSN method overcomes the limitation of log transformation by accommodating the negative values and minimizing the variance whose values are increasing around low intensities. This calibrates the variation among features through shifting and scaling mechanisms where all data are adjusted.

Huber et al. (2002) and Durbin et al., (2002) are independently proposing the VSN approach, which is a variant of log-transform (glog2). (Serin, 2011) explained the VSN method as follows.

The variance of measured intensity X_i from the u gene depends on the mean of measurement intensities X_i. Hence, interpreting the fold change on raw data may lead to a different conclusion. The VSN method transforms the data so that the mean and variance are independent. The model is based on the standard error model.

$$Y = \alpha + \mu e^{\eta} + \epsilon \quad (16)$$

Y is the observed expression values, α is the offset, $and\ \mu$ the truth expression. The additive and multiplicative error terms, respectively, are ε and η. The expectation and variance of Y are

$$E(Y) = u = \alpha + m_\eta \mu \quad (17)$$

$$Var(Y) = v = s_\eta^2 \mu^2 + \sigma_\epsilon^2 \quad (18)$$

The mean and variance of eη respectively m_η and s_η^2, σ_ϵ^2 is a variance of ε. The estimator μ from equation (17) is $\left(u - \frac{\alpha}{m_\eta}\right)$. The reformulated variance from equation (18) in terms of $E(Y)$ is

$$v(u) = \frac{s_\eta^2}{m_\eta} u - \alpha^2 + \sigma_\epsilon^2 = (c_1 u + c_2)^2 + c_3 \quad (19)$$

The dependency between the variance v and the average u can be seen in equation (19). The delta method transforms Y into $h(Y)$ such that the mean is independent of its variance.

$$h(Y) = \int_0^y \frac{1}{\sqrt{v(u)}} du \quad (20)$$

Therefore, if the intensity variance v and mean u can be estimated for each probe, then we can compute the functions $v(u)$ and $h(y)$ to stabilize variance through equation (20).

The summary of all normalization methods can be seen on Table 2.

Table 2. Normalization methods for microarray data

Method	Key Features	Pros	Cons	References
Global/Affymetrix Normalization	Normalize expression by scaling arrays to a common target value	Easy to implement; Suitable for early Affymetrix arrays	Ignores intensity distribution; Overly simplistic	Affymetrix
Global Rank-Invariant Set Normalization	Uses a subset of genes (rank-invariant) to align distributions	Accounts for global and local variation; Better for non-random expression shifts	Selection of rank-invariant set can be subjective	Li & Wong (2001a,b); Pelz et al. (2008); Baans et al. (2019)
Contrast Normalization	Adjusts arrays by comparing intensities between channels (two-color arrays)	Suitable for dye-bias correction; Works well for paired comparisons	Limited to two-color platforms	Astrand (2003)
Quantile Normalization	Forces all arrays to have the same empirical distribution	Reduces technical variability; Simple and widely used	May distort biological variation if assumptions are violated	Bolstad et al. (2003); Baans et al. (2019)
Scale Normalization	Standardizes probe or array means/variances to a reference	Simple, fast; Preserves relative differences	May be insufficient for complex bias patterns	Bolstad et al. (2003)
Cyclic Loess	Pairwise loess regression between arrays in cycles	Preserves non-linear relationships; Effective on small sample sizes	Computationally intensive; Slower for large datasets	Dudoit et al. (2002)
Variance Stabilization Normalization (VSN)	Transforms data to stabilize variance across intensities	Handles heteroscedasticity; Improves downstream modeling	May distort data if over-applied	Huber et al. (2002); Durbin et al. (2002)

Probe correction

Perfect Match/Mismatch (PM/MM)

This method corrects for the differences in hybridization efficiency between perfect match (PM) and mismatch (MM) probes by subtracting the MM value from the PM value. The resulting PM-MM difference value represents the actual signal for the corresponding gene (Affymetrix, 2002).

Probe Logarithmic Intensity Error (PLIER)

PLIER is a PM correction method that uses a Bayesian framework to estimate the actual expression level of a gene based on the PM and MM probe intensities. PLIER effectively reduces the technical noise and improves the accuracy of gene expression measurements (Li, C. and Wong, 2001a, 2001b).

Model-Based Expression Index (MBEI)

MBEI is a PM correction method that uses a linear mixed-effects model to estimate the actual expression level of a gene based on the PM and MM probe intensities. MBEI effectively reduces the technical noise and improves the sensitivity of detecting differentially expressed genes (Wu et al., 2004).

10.12928/bamme.v5i1.12539

Linear Models for Microarray Data (LIMMA)

LIMMA is a PM correction method that uses a linear model to estimate the actual expression level of a gene based on the PM and MM probe intensities. LIMMA effectively reduces technical noise and improves gene expression measurements' accuracy (Smyth, 2006).

Summarization

Summarization is a process to produce gene expression values. The researcher can choose several methods of summarization. (Gondro, 2009) mentioned some of them, for instance, MAS 5.0, RMA, GCRMA, PLIER, VSN, and MBEI. Some summarization methods are explained in the following subsections.

Average difference

The average Difference or AvgDiff method is a method to measure the gene expression that Affymetrix has proposed in MAS4.0 software. In Affymetrix design, each gene is measured by a probe set, and each probe set contains many probes. Probes in Affymetrix design are of two types: Perfect Match (PM) and Miss Match (MM) probes (Affymetrix, 2002).

In AvgDiff for each probe set first, the difference of PM and MM intensities probes are computed. Later, the mean difference is computed and used as an estimator of gene expression. This method is rarely used because:

a. The gene expression can be negative.
b. For low gene expression, the error will be very high.
c. The affinity probes are not included.
d. The average value is not on a logarithmic scale.

Mean

The mean summarization method is implemented by taking the average of probe set expression as the gene expression estimator. The drawback of this method is that it is not robust because the average is affected by *outlier*s.

Tukey biweight, MAS5.0

Unlike mean methods, the Tukey Biweight is robust in computing the mean (Affymetrix, 2002). The steps in this method are

a. Compute the median.
b. Compute the distance from each data set to the median. This distance determines how much the contribution of the data toward the mean.
c. Compute the weight w_i for each data by using the equation (21).

$$w_i = \begin{cases} \left(1 - \left(\frac{u_i}{c}\right)^2\right)^2 & , \text{if } |u_i| < c \\ 0 & , \text{if } |u_i| \geq c \end{cases} \quad (21)$$

where $u_i = \frac{x_i - T}{MAD}$ and c is the cutoff parameter, usually $c = 4.685$.

d. Compute the biweight mean as following

$$\text{Biweight Mean} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad (22)$$

We can say that the Tukey Biweight method gives a smaller weight for the outlier data, which is the data that far from its mean.

Median Polish, RMA

For the k-th probe set, $i = 1,2,3, \dots, I$, array dan $j = 1,2,3, \dots, J$ probe, the model of median polish can be written as

$$log_2\left(PM_{ij}^{(k)}\right) = \alpha_j^{(k)} + \beta_i^{(k)} + \varepsilon_{ij}^{(k)} \quad (23)$$

where

$PM_{ij}^{(k)}$ represents the value after the background correction, normalization, and log2 transformed from the Perfect Match (PM) intensity;

$\beta_i^{(k)}$ represents i-th array expression value in log2 scale;

$\alpha_j^{(k)}$represents j-th probe log scale affinity effect; and

$\varepsilon_{ij}^{(k)}$ represents the random error.

(Giorgi, Federico & Bolger, Anthony & Lohse, Marc & Usadel, 2010) and (Irizarry et al., 2003)

FARMS

Factor Analysis for Robust Microarray Summarization (FARMS) method is built to summarize data on probe level for Affymetrix GeneChips. The factor analysis is the foundation of the method where a Bayesian maximum posterior method optimizes the model parameter, assuming that the noise is normally distributed. The RNA concentration is estimated based on the model. The method is introduced in (Hochreiter et al., 2006).

The FARMS model is

$$y_{ij} = w_i x_j + \varepsilon_{ij} \quad (24)$$

where

y_{ij} is the intensity (log PM_{ij} or the log difference between PM and MM, $PM_{ij} - MM_{ij}$) of i-th probe and j-th array;

w_i is the weight of i-th probe intensity;

x_j is the gene latent expression on the array j; and

ε_{ij} is a gaussian noise $\mathrm{N}(0, \sigma^2)$.

The model in equation (24) is used to estimate the x_j, w_i and σ^2 with the Gaussian prior of $w_i \sim \mathrm{N}(0, \tau^2)$ and $x_j \sim \mathrm{N}(0, \eta^2)$. The x_j expression is computed by

$$\hat{x}_j = \frac{\sum_{i=1}^{I} w_i y_{ij}}{\sum_{i=1}^{I} w_i^2} \quad (25)$$

Li-Wong, MBEI

Li-Wong MBEI model is based on the model

$$y_{ij} = \phi_i \theta_j + \varepsilon_{ij} \quad (26)$$

where

y_{ij} is PM_{ij} the intensity of i-th probe in j-th array

ϕ_i is the i-th probe affinity,

θ_j is the gene expression on the array j, and

ε_{ij} is a random error assumed to be normally distributed (Li, C. and Wong, 2001a, 2001b).

10.12928/bamme.v5i1.12539

The ϕ_i and θ_j is estimated by iterative optimization method, such as the least square, where the model is first represented in log-transform. The $\widehat{\theta_j}$ is the estimator of probe set expression value.

Table 3. Normalization methods for microarray data

Method	Key Features	Pros	Cons	References
Average Difference	Simple average of PM-MM for each probe pair (linear scale)	Easy to interpret; No transformation	Sensitive to outliers; Not robust to noise	Affymetrix (2002)
Mean	Arithmetic mean of probe intensities	Simple; Computationally efficient	Affected by extreme values; Less robust	Affymetrix (2002)
Tukey Biweight (MAS5.0)	Robust average using weighted contribution (less influence from outliers)	Reduces effect of outliers; Used in MAS5.0	Less effective if data is heavily skewed	Affymetrix (2002)
Median Polish (RMA)	Removes row/column effects iteratively; works on log2 scale	Robust to outliers; Captures probe effects well	More complex; Assumes additive model	Irizarry et al. (2003); Giorgi et al. (2010)
FARMS	Factor analysis-based; uses signal-to-noise and reliability weighting	Accurate summarization; Handles probe reliability	Requires good prior estimation; More complex	Hochreiter et al. (2006)
Li-Wong (MBEI)	Model-based estimation including probe affinity and expression level	Accurate; Adjusts for probe-specific effects	Assumes correct model; Computationally intensive	Li & Wong (2001a,b)
DFW	Distribution-free method using weighted contributions of probes	No assumption of data distribution; Good for irregular data	Still underused; Less documented	Chen et al. (2007)

Distribution Free Weighted, DFW

In the previous sections, we have talked about some summarization methods:

a. based on assumptions and require parameter estimation
b. only a few used the information about non-specific and cross-hybridization.

The DFW method is built based on no distributional assumptions and uses non-specific and cross-hybridization (Chen et al., 2007). This method calculates the summarized expression values (log base 2) of a probe set across arrays based on the weighted probe intensity values. It is explained as following:

a. For each probe set, calculate the weighted intensity value based on the log base 2 PM intensity and the weight from each probe within the probe set as $w_i = \left(1 - \left(\frac{x_i}{Max}\right)^2\right)^2$, where Max is the maximum absolute value of x_i.
b. The weighted intensity values are linearly transformed to be between 0 and 1 to give the transformed intensity values, $w_i = \frac{w(x_i)}{\Sigma_{j=1}^{J} w(x_j)}$, where j is the number of PMs in the probe set.
c. Calculate the probe's variability across arrays by the weighted range (WR), the range of the weighted intensity values for each probe.
d. Calculate the weighted standard deviation (WSD) as the weighted intensity values, where x_i the median-centered standard deviation across arrays replaces xi in step (a).
e. The expression value G_i for gene i across arrays is given by

$$ExpValue = \min + c(TIV)(WR)^m(WSD)^n \quad (27)$$

m and n are positive numbers (default values are $m = 3$ and $n = 1$), min is the minimum of the weighted intensity values before the linear transformation, and c is a scale parameter. The summary of all summarization methods can be seen on Table 3.

Remark

Bioinformatics data available once the microarray experiment has been done. This data is the raw data where the error measurement affects the real value. Before the data analysis is carried out, to remove the errors or unwanted variations then the pre-processing steps need to be implemented. We describe some methods in pre-processing on steps of background correction, normalization, probe correction and summarization. The methods have been wrapped in the Bioconductor-Biopython packages with R-Python programming languages. Some of the packages for microarray data analysis have been written in Matlab and Julia programming languages as well. There is another step needs to be taken before analyzing the data, the filtering. This will be addressed in the different paper.

References

Affymetrix, I. (2002). Statistical algorithms description document. *Technical paper*, *62*, 110.

Astrand, M. (2003). Contrast normalization of oligonucleotide arrays. *Journal of Computational Biology*, *10*(1), 95–102. https://doi.org/10.1089/106652703763255697

Baans, O. S., Jambek, A. B., & Said, K. A. M. (2019). Analysis of normalization method for DNA microarray data. *Asia-Pacific Journal of Molecular Biology and Biotechnology*, *27*(4), 30–37. https://doi.org/10.35118/apjmbb.2019.027.4.04

Barbacioru, C. C., Wang, Y., Canales, R. D., Sun, Y. A., Keys, D. N., Chan, F., Poulter, K. A., & Samaha, R. R. (2006). Effect of various normalization methods on Applied Biosystems expression array system data. *BMC Bioinformatics*, *7*, 1–14. https://doi.org/10.1186/1471-2105-7-533

Bolstad, B. M. (2004). *Bolstad_2004_Dissertation*. 156. papers2://publication/uuid/8B996D4A-CD91-4F11-9F50-7B5E60EFC00C

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). Gene Expression Omnibus A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics*, *19*(2), 185–193. http://www.ncbi.nlm.nih.gov/geo

Chen, Z., McGee, M., Liu, Q., & Scheuermann, R. H. (2007). A distribution free summarization method for Affymetrix GeneChip® arrays. *Bioinformatics*, *23*(3), 321–327. https://doi.org/10.1093/bioinformatics/btl609

Cheng, L., Lo, L. Y., Tang, N. L. S., Wang, D., & Leung, K. S. (2016). CrossNorm: A novel normalization strategy for microarray data in cancers. *Scientific Reports*, *6*, 1–2. https://doi.org/10.1038/srep18898

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, *74*(368), 829. https://doi.org/10.2307/2286407

Cleveland, W. S., & Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, *83*(403), 596. https://doi.org/10.2307/2289282

Dozmorov, M. G., Guthridge, J. M., Hurst, R. E., & Dozmorov, I. M. (2010). A comprehensive and universal method for assessing the performance of differential gene expression analyses. *PLoS ONE*, *5*(9), 1–11. https://doi.org/10.1371/journal.pone.0012657

Dudoit, S., Yang, Y. H., Speed, T. P., & Callow, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, *12*(1), 111–140. http://www1.cs.columbia.edu/~cleslie/cs4761/lectures/speed-statistical.pdf

Durbin, B. P., Hardin, J. S., Hawkins, D. M., & Rocke, D. M. (2002). A variance-stabilizing

 10.12928/bamme.v5i1.12539

transformation for gene-expression microarray data. *Bioinformatics (Oxford, England)*, *18 Suppl 1*, S105–S110. https://doi.org/10.1093/bioinformatics/18.suppl_1.s105.

Fajriyah R. (2021). Paper review: An overview on microarray technologies. Bulletin of Applied Mathematics and Mathematics Education, 1(1), 21-30.

Federico, A., Saarimäki, L. A., Serra, A., Giudice, G. Del, Kinaret, P. A. S., Scala, G., & Greco, D. (2022). *Microarray Data Preprocessing: From Experimental Design to Differential Analysis*. *24*(01), 79–100. https://doi.org/10.1007/978-1-0716-1839-4_7

Fujita, A., Sato, J. R., de Oliveira Rodrigues, L., Ferreira, C. E., & Sogayar, M. C. (2006). Evaluating different methods of microarray data normalization. *BMC Bioinformatics*, *7*, 1–11. https://doi.org/10.1186/1471-2105-7-469

Gautier, L., Bolstad, B. M., Cope, L., & Irizarry, R. A. (2004). Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405

Gharaibeh, R. Z., Fodor, A. A., & Gibas, C. J. (2008). Background correction using dinucleotide affinities improves the performance of GCRMA. *BMC Bioinformatics*, *9*, 1–12. https://doi.org/10.1186/1471-2105-9-452

Giorgi, F.M., Bolger, A.M., Lohse, M. (2010). Algorithm-driven Artifacts in median polish summarization of Microarray data. *BMC Bioinformatics, 11*, 553. https://doi.org/10.1186/1471-2105-11-553.

Gondro, C. (2009). Summarization methods and quality problems in Affymetrix microarrays. *Proc Assoc Advmt Anim Breed Genet*, *18*(February).

Grant, G. R., Manduchi, E., & Stoeckert, C. J. (2007). Analysis and management of microarray gene expression data. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *Chapter 19*, 1–30. https://doi.org/10.1002/0471142727.mb1906s77

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., & Young, R. A. (2001). Maximum-likelihood estimation of optimal scaling factors for expression array normalization. *Microarrays: Optical Technologies and Informatics*, *4266*(Ml), 132–140. https://doi.org/10.1117/12.427981

Hochreiter, S., Clevert, D. A., & Obermayer, K. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics*, *22*(8), 943–949. https://doi.org/10.1093/bioinformatics/btl033

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, *18*(SUPPL. 1). https://doi.org/10.1093/bioinformatics/18.suppl_1.S96

Irizarry, R. A., Bolstad, B., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, *31*(4), e15. https://doi.org/10.1093/nar/gng015

Klaus, B., & Reisenauer, S. (2018). An end to end workflow for differential gene expression using Affymetrix microarrays. *F1000Research*, *5*, 1–56. https://doi.org/10.12688/f1000research.8967.2

Kuyuk, S. A. (2017). Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics and Epidemiology International Journal*, *0*(0), 1–8. https://doi.org/10.30881/beij.00001

Li, C. and Wong, W. . (2001a). Model-based analysis of oligo- nucleotide arrays: expression index computation and outlier detection. *Computational Statistics & Data Analysis*, *a*(98), 31–36.

Li, C. and Wong, W. H. (2001b). *Model-based analysis of oligo- nucleotide arrays: model validation, design issues and standard error application*. *b*(2), 1-11.

Microarray Galaxy User's Guide. (2023). *Microarray Galaxy User's Guide*. http://www.ensat.ac.ma/mobihic/microarray-galaxy.html

Miranda, J., & Bringas, R. (2008). Analysis of DNA microarray data. Part I: Technological background and experimental design. *Biotecnologia Aplicada*, *25*(2).

Munster, S., VL, W., Hutchings, DC., B. D., & Nicholson, S. (2018). Comparison Study of Microarray and RNA-seq for Differential Expression. *Final Report*. https://doi.org/DOT/FAA/AM-

20/09

Wright Muelas, M., Mughal, F., O'Hagan, S. *et al.* The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. *Sci Rep* **9**, 17960 (2019). https://doi.org/10.1038/s41598-019-54288-7

Naef, F., & Magnasco, M. O. (2003). Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *68*(1), 4. https://doi.org/10.1103/PhysRevE.68.011906

Olson, N. E. (2006). The Microarray Data Analysis Process: From Raw Data to Biological Significance. *NeuroRx*, *3*(3), 373–383. https://doi.org/10.1016/j.nurx.2006.05.005

Pelz, C. R., Kulesz-Martin, M., Bagby, G., & Sears, R. C. (2008). Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. *BMC Bioinformatics*, *9*(January 2009). https://doi.org/10.1186/1471-2105-9-520

Piccolo, S. R., Ying Sun, Campbell, D, J., Lenburg, M. E., Bild, A. H., & W Evan Johnson. (2012). A single-sample microarray normalization method to facilitate personalized-medicine workflow. *Genomics*, *100*(6), 337–344. https://doi.org/10.1016/j.ygeno.2012.08.003

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., & Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, *23*(20), 2700–2707. https://doi.org/10.1093/bioinformatics/btm412

Serin, A. (2011). *Biclustering Analysis for Large Scale Data*. *September*. http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000035625?lang=en

Silver JD, Ritchie ME, & S. G. (2009). Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. *Biostatistics and Epidemiology International Journal*, *10*(2), 52–63. https://doi.org/10.1093/biostatistics/kxn042

Smyth, G. K. (2006). *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray*. *3*(1), 1–26.

Smyth, G. K., & Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, *31*(4), 265–273. https://doi.org/10.1016/S1046-2023(03)00155-5

TechMedBuddy,. (2023). Microarray Data Analysis in Bioinformatics: A Comprehensive Overview. https://www.linkedin.com/pulse/microarray-data-analysis-overview-techmedbuddy/

Visentin, L., Scarpellino, G., Chinigò, G., Munaron, L., & Ruffinatti, F. A. (2022). BioTEA: Containerized Methods of Analysis for Microarray-Based Transcriptomics Data. *Biology*, *11*(9), 1–14. https://doi.org/10.3390/biology11091346

Wright Muelas, M., Mughal, F., O'Hagan, S., Day, P. J., & Kell, D. B. (2019). The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. *Scientific Reports*, *9*(1), 1–21. https://doi.org/10.1038/s41598-019-54288-7

Wu, Z. (2009). A Review of Statistical Methods for Preprocessing. *Nih*, *71*(2), 233–236. https://doi.org/10.1177/0962280209351924.A

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., & Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, *99*(468), 909–917. https://doi.org/10.1198/016214504000000683

Yang, J., & Thorne, N. (2002). Normalization for Two-color cDNA Microarray Data. *Science and Statistics: A Festschrift for Terry Speed*, 403–418.

Yang, Y., S, D., P, L., DM, L., V, P., J, N., & TP, S. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, *30*(4).

 10.12928/bamme.v5i1.12539