Austrian Journal of Statistics
March 2016, Volume 45, 15–33.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i2.92

A Study of Convolution Models for Background Correction of BeadArrays

Rohmatul Fajriyah
Graz University of Technology and Universitas Islam Indonesia

Abstract

The robust multi-array average (RMA), since its introduction in Irizarry, Bolstad, Collin, Cope, Hobbs, and Speed (2003a); Irizarry, Hobbs, Collin, Beazer-Barclay, Antonellis, Scherf, and Speed (2003b); Irizarry, Wu, and Jaffee (2006), has gained popularity among bioinformaticians. It has evolved from the exponential-normal convolution to the gamma-normal convolution, from single to two channels and from the Affymetrix to the Illumina platform.

The Illumina design provides two probe types: the regular and the control probes. This design is very suitable for studying the probability distribution of both and one can apply a convolution model to compute the true intensity estimator.

In this paper, we study the existing convolution models for background correction of Illumina BeadArrays in the literature and give a new estimator for the true intensity, assuming that the intensity value is exponentially or gamma distributed and the noise has lognormal distribution.

Our study shows that one of our proposed models, the gamma-lognormal with the method of moments for parameters estimation, is the optimal one for the benchmarking data set with benchmarking criteria, while the gamma-normal model has the best performance for the benchmarking data set with simulation criteria.

For the publicly available data sets, the gamma-normal and the exponential-gamma models with maximum likelihood estimation method can not be used and our proposed models exponential-lognormal and gamma-lognormal have the best performance, showing a moderate error in background correction and in the parametrization.

Keywords: convolution, background correction, BeadArrays.

1. Introduction

There are various processes in producing data from microarray experiments and each process contributes noise to the data. The noise can be of two types, biological and non-biological. Non-biological noise should be avoided or at least minimized.

Sources for the non-biological noise are, for example, the chip itself, the scanner, or fluctuations in the electric network. Therefore, the data needs to be adjusted. The pre-processing will adjust the intensity value (Huber, Irizarry, and Gentleman 2005a; Huber, von Heydebreck, and Vingron 2005b) and provides an estimate of the true intensity.

To estimate the intensity value, researchers proposed additive and multiplicative models and also *additive-muliplicative error models*, see e.g. Huber, von Heydebreck, and Vingron (2004). In case of additive models, the underlying distribution is generally chosen as normal (log-normal), exponential, or a Gamma-t mixture in the parametric approach (Allen, Chen, and Xie (2009), Bolstad, Irizarry, Astrand, and Speed (2003), Chen, Xie, and Story (2011), Hochreiter, Djork-Arné, and Obermayer (2006), Irizarry *et al.* (2003a,b, 2006), and Plancade, Rozenholc, and Lund (2011, 2012)).

Irizarry *et al.* (2003a,b, 2006) and Bolstad *et al.* (2003), on the Affymetrix platform, have estimated the true intensity values based on a convolution model in the background correction step of their robust multi-array average (RMA) pre-processing method. They assumed that the true intensity is exponentially distributed and the background noise is normally distributed.

Plancade *et al.* (2011, 2012) showed that the RMA model (in Bolstad *et al.* (2003) and Irizarry *et al.* (2003a,b, 2006)) does not fit Illumina BeadArrays: using the exponential-normal convolution leads to a large distance between the observed and the modeled intensities. They proposed, instead, the implementation of a gamma distribution for the intensity value and normal distribution for the noise.

The simulation study of Plancade *et al.* (2011, 2012) showed that the gamma-normal model performs better than the existing exponential-normal convolution model, giving a more accurate and correct fit for the observed intensities in Illumina BeadArray.

Using a Gamma distribution for the intensity values in Illumina BeadArrays has been first suggested by Xie, Wang, and Story (2009).

The studies of Baek, Son, and MacLachlan (2007) (on the background correction of the image processing) and Chen *et al.* (2011) show that the noise distribution is usually skewed in different degrees. In their studies, based on simulated and real data sets, Baek *et al.* (2007) conclude that the gamma distribution is well suited for the noise. It accounts for the intensities with a positive lower bound and is very flexible in its shape, including asymmetric exponential type and symmetric normal type.

The proposed convolution of exponential-gamma distribution by Chen *et al.* (2011) improves the intensity estimation and the detection of differentially expressed genes in the case when the intensity to noise ratio is large and the noise has a skewed distribution.

In view of the remarks above, it is natural to model both the true intensity and the background noise in Illumina BeadArrays as gamma distributed. In an earlier version of this paper we have developed an estimator for the true intensity based on the gamma-gamma convolution model of RMA. However, this model does not fit very well the Illumina benchmarking data set. Independently, Triche, Weisenberger, Berg, Laird, and Siegmund (2013) proposed and applied the gamma-gamma model to pre-process Illumina methylation arrays.

In this paper we introduce a new model for background correction in Illumina BeadArrays where the true intensity value is exponentially or gamma distributed and the noise has a lognormal distribution. As we will see, this model avoids the difficulties with the gamma-gamma model and has an overall satisfactory performance.

We note that a new method reducing the bias of the maximum likelihood estimator of the shape parameter of the gamma distribution was proposed by Zhang (2013). But since our samples are very large, bias is not a problem in our studies.

We compare the performance of the models on the Illumina spike-in data set, based on various criteria: root and mean square error (RMSE), L_1 error, Kullback-Leibler (K-L) coefficient, and some adapted criteria from Affycomp (Cope, Irizarry, Jaffee, Wu, and Speed 2004). These criteria are measuring the reproducibility, accuracy, precision, specificity, and sensitivity of the expression measure of each model. We then provide a simulation study to measure the consistency of the error of background correction and the parametrization. The description and some details on these criteria and simulation can be found at `http://rfajriyah.staff.`

uii.ac.id/category/supplementary/

Our paper is organized as follows. In Section 2 we review previous work related to the background correction for Illumina BeadArrays. Our proposed models are described in Sections 3.1 and 3.2. Section 3.3 explains the benchmarking and the simulation studies on the benchmarking data set and Section 3.4 compares the performance of all models in the public data sets. Finally, Section 4 states the conclusions and indications of future work.

2. Previous work

Affymetrix is the pioneer and most widely used platform for microarray gene expression experiments. The tools and algorithms to handle the data are numerous, both free and commercial. Some methods for pre-processing are available. Examples for the background correction step are: MAS5.0 by Affymetrix, multiplicative model based expression index (MMBE) by Li and Wong (2001), RMA in Irizarry *et al.* (2003a,b, 2006) and Bolstad *et al.* (2003), GC-RMA by Wu, Irizarry, Gentleman, Martinez-Murillo, and Spencer (2004) and maximum likelihood estimation based on the normal-exponential convolution model by Silver, Ritchie, and Smyth (2009).

Illumina is one of the alternative platforms and is increasingly popular. A few statistical methods have been developed for BeadArray data and there is no consensus yet for the pre-processing steps (Shi, Oshlack, and Smyth 2010). Xie *et al.* (2009) mention that for the background correction step, Illumina bead studio gives two options (no background correction and background substraction) and the packages for BeadArrays in R provide three options (no background correction, background substraction and RMA background correction).

Ding, Xie, Park, Xiao, and Story (2008) extended the RMA model by proposing the model-based background correction method (MBCB) and showed that their model leads to a more precise determination of the gene expression and a better biological interpretation of Illumina BeadArray data.

The studies of Chen *et al.* (2011) and Plancade *et al.* (2011, 2012) show that their background correction models are made by adapting the RMA Affymetrix model. As Forcheh, Verbeke, Kasim, Lin, Shkedy, Talloen, Gohlmann, and Clement (2012), pointed out, most preprocessing methods for Illumina BeadArrays are taken from the Affymetrix microarray platform.

In general, the background correction is applied toward each array, where in each array there are probes, probesets and genes (terminology for the Affymetrix platform) or bead and bead-type level probes (terminology for the Illumina platform).

At the Illumina platform, each gene is only targeted by one bead-type, which has been represented by about 30 time replications. If we can have a raw benchmarking data set, then it is possible to have all bead-type level probes of the raw data intensities.

The current publicly available benchmarking data set for the Illumina platform is the raw data from the bead studio, which is the average of the bead-type level probes, not background corrected and of unnormalized intensity. Therefore, the background correction in this paper is applied to the gene intensity in each array.

Suppose we have J arrays and for each array there are I regular genes. Our convolution model is as follows:

$$P_{ij} = S_{ij} + B_{ij}\,, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J\,, \tag{1}$$

where P_{ij}, S_{ij}, and B_{ij} are the observed signal, true signal, and noise intensity, respectively. The S_{ij} are i.i.d. and so are the B_{ij}; the S_{ij} are independent of the B_{ij}. The P_{ij} are observable. Our task is to recover the unknown signals S_{ij} from the P_{ij}. To do this, for each array j we also have M observable noise intensities B_{0mj}, $m = 1, \ldots, M$; the B_{0mj} are i.i.d. with the same distribution as the B_{ij} and are independent of all of the S_{ij}, B_{ij}. The S_{ij} and B_{ij} have a known type of distribution (exponential, gamma, normal, lognormal) with unknown parameters.

2.1. Background correction by RMA

The RMA method was developed for the Affymetrix platform, where the design of arrays is different from the Illumina one. In this platform, one has perfect match and mismatch probe design. The RMA uses only the perfect match probe, which is the targeted probe of the intended gene. For those not familiar with the Affymetrix platform, refer to Bolstad (2004), Bolstad *et al.* (2003), and Irizarry *et al.* (2003a,b, 2006) for further information.

In modeling the intensity values, the RMA model (Bolstad *et al.* (2003) and Irizarry *et al.* (2003a,b, 2006)) assumes that the intensity values are affected by the noise of the chip. By referring to Equation (1), in the RMA model P_{ij} is the observed bead-type level probe intensity, S_{ij} is the true signal with $S_{ij} \sim f_1(s_{ij};\theta_j) = \mathrm{Exp}(\theta_j)$, $\theta_j > 0$, and B_{ij} is the background noise of the chip with $B_{ij} \sim f_2(b_{ij};\mu_j,\sigma_j^2) = \mathcal{N}(\mu_j,\sigma_j^2)$, $\mu_j \in \mathbb{R}$, $\sigma_j^2 > 0$. To avoid negative intensity values, we truncate B_{ij} at 0 from below, i.e. we replace B_{ij} by $\max\{B_{ij}, 0\}$; this will not change its density function $f_2(b_{ij};\mu_j,\sigma_j^2)$ for $b_{ij} > 0$.

Assuming independence, the joint density of the two-dimensional random variable (S_{ij}, B_{ij}) is

$$f_{S_{ij},B_{ij}}(s_{ij},b_{ij};\mu_j,\sigma_j^2,\theta_j) = \theta_j \exp\Big\{-\theta_j s_{ij}\Big\} f_2(b_{ij};\mu_j,\sigma_j^2), \qquad b_{ij}, s_{ij} > 0\,.$$

Furthermore, the transformation formula for two-dimensional densities gives that joint density of S and P is

$$f_{S_{ij},P_{ij}}(s_{ij},p_{ij};\mu_j,\sigma_j^2,\theta_j) = \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_{ij}-\mu_j)\right\} f_2(s_{ij};\mu_{S.P,j},\sigma_j^2), \qquad 0 < s_{ij} < p_{ij}, \quad (2)$$

where $\mu_{S.P,j} = p_{ij} - \mu_j - \theta_j\sigma_j^2$. From (2) we get the marginal density of P_{ij} and the conditional density of S_{ij} given P_{ij} as

$$f_{P_{ij}}(p_{ij};\theta_j,\mu_j,\sigma_j^2) = \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_{ij}-\mu_j)\right\}\left(\Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) - 1\right)$$

$$f_{S_{ij}|P_{ij}}(s_{ij}|p_{ij};\mu_{S.P,j},\sigma_j^2) = \frac{f_2(s_{ij};\mu_{S.P,j},\sigma_j^2)}{\Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) - 1}.$$

The background adjusted intensity is computed as the conditional expectation of the true signal given the observed intensity, i.e.

$$\mathsf{E}(S_{ij}|P_{ij}=p_{ij}) = \frac{1}{\Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) - 1}\int_0^{p_{ij}} s_{ij} f_2(s_{ij};\mu_{S.P,j},\sigma_j^2)ds_{ij}.$$

The substitution $s_{ij} = \mu_{S.P,j} + \sigma_j t$ yields

$$\int_0^{p_{ij}} s_{ij} f_2(s_{ij};\mu_{S.P,j},\sigma_j^2)ds_{ij} = \mu_{S.P,j}\left(\Phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right) + \Phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) - 1\right) + \sigma_j\left(\phi\left(\frac{\mu_{S.P,j}}{\sigma_j}\right) - \phi\left(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j}\right)\right)$$

and thus (see Bolstad *et al.* (2003))

$$\mathsf{E}(S_{ij}|P_{ij}=p_{ij}) = \mu_{S.P,j} + \sigma_j\frac{\phi(\frac{\mu_{S.P,j}}{\sigma_j}) - \phi(\frac{p_{ij}-\mu_{S.P,j}}{\sigma_j})}{\Phi(\frac{\mu_{S.P,j}}{\sigma_j}) + \Phi(\frac{p-\mu_{S.P,j}}{\sigma_j}) - 1}. \quad (3)$$

Note that modelling the noise as a truncated normal variable has the consequence that the noise equals 0 with a positive probability p_0, a rather unpleasant feature of the model. As

pointed out in Xie *et al.* (2009), however, in practical cases p_0 is rather small, so this problem can be disregarded. To avoid this difficulty, one can model the noise as the absolute value of a $\mathcal{N}(\mu, \sigma^2)$ variable, which changes the calculations above. However, since in this paper we will provide a background correction model fitting the reality considerably better, we do not give the details here.

2.2. Exponential-normal MBCB

Xie *et al.* (2009) use the same underlying distributions in (1) for background correction. The difference with the RMA (Bolstad *et al.* (2003) and Irizarry *et al.* (2003a,b, 2006)) are

1. Xie *et al.* (2009) use $+\infty$ as the upper bound of the integral to compute the marginal density function and the conditional expectation of the true intensity value. On the other hand, RMA uses p as the upper bound of the integration.

   The background corrected intensity value of Xie *et al.* (2009) is

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \mu_{S.P,j} + \sigma_j \frac{\phi(\frac{\mu_{S.P,j}}{\sigma_j})}{\Phi(\frac{\mu_{S.P,j}}{\sigma_j})}.$$

2. Under the convolution model (1), where the true intensity value is assumed exponentially distributed and the noise is normally distributed, we need to estimate the parameters θ_j, μ_j, and σ_j^2. Xie *et al.* (2009) offer three estimation methods: the method of moments, maximum likelihood estimation, and a Bayesian approach. On the other hand, RMA applies the *ad-hoc* method.

Ding *et al.* (2008) use the exponential-normal convolution model to correct the background of the Illumina platform by using Markov chain Monte Carlo simulation.

2.3. Gamma-normal convolution

Plancade *et al.* (2011, 2012) introduced gamma-normal convolution to model the background correction of Illumina BeadArray. The model is based on the RMA background correction of Affymetrix GeneChip. Plancade *et al.* (2011, 2012) assume that the intensity value is gamma distributed and the noise is normally distributed.

Under model (1), $f_{P_{ij}}$ is the convolution of $f_{S_{ij}}$ and $f_{B_{ij}}$. The background corrected intensity is computed as the conditional expectation of S_{ij} given $P_{ij} = p_{ij}$, i.e.

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{\int s_{ij} f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}}{\int f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}}, \tag{4}$$

where

$$f^{\text{gam}}_{\alpha_j,\beta_j}(s) = \frac{\beta_j^{\alpha_j} s^{\alpha_j - 1} \exp(-\beta_j s)}{\Gamma(\alpha_j)}, \qquad \alpha_j, \beta_j, s > 0$$

is the gamma density. When S_{ij} is gamma distributed and B_{ij} is normally distributed, then (4) has no analytic expression like (3). Plancade *et al.* (2011, 2012) implemented the Fast Fourier Transform to estimate the parameters and to correct the background. For the background correction with Fast Fourier Transform, the corrected intensity (4) is rewritten as

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{\alpha_j \beta_j \int f^{\text{gam}}_{\alpha_j+1,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}}{\int f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) f^{\text{norm}}_{\mu_j,\sigma_j}(p_{ij} - s_{ij}) ds_{ij}},$$

since $s_{ij} f^{\text{gam}}_{\alpha_j,\beta_j}(s_{ij}) = \alpha_j \beta_j f^{\text{gam}}_{\alpha_j+1,\beta_j}(s_{ij})$ is valid for every $s_{ij} > 0$.

2.4. Exponential-gamma convolution

Under model (1), Chen *et al.* (2011) proposed for the distribution of the true intensity and its noise the exponential and gamma distribution, respectively. Therefore, $S_{ij} \sim f_1(s_{ij};\theta_j) = \text{Exp}(\theta_j)$ and $B_{ij} \sim f_2(b_{ij};\alpha_j,\beta_j) = \text{Gamma}(\alpha_j,\beta_j)$, where $s_{ij}, b_{ij}, \theta_j, \alpha_j, \beta_j > 0$.

The corrected background intensity of Chen *et al.* (2011) is

$$\mathsf{E}(S_{ij}|P_{ij}=p_{ij}) = p_{ij} - \frac{\int_0^{p_{ij}} b_{ij}^{\alpha_j} \exp(-(\beta_j-\theta_j)b_{ij})db_{ij}}{\int_0^{p_{ij}} b_{ij}^{\alpha_j-1} \exp(-(\beta_j-\theta_j)b_{ij})db_{ij}}.$$

3. Results

Now we present the results of the two proposed convolution models: the exponential-lognormal convolution in Section 3.1 and the gamma-lognormal convolution in Section 3.2. In each section, the formula for the background corrected intensity value is derived and methods to estimate the parameters are explained. Section 3.3 present the benchmarking results, i.e. a performance comparison of all models at the Illumina spike-in data set. Section 3.4 present the performance comparison of all models for the public data sets. The simulation study results are presented in Sections 3.3 and 3.4.

For further details on the benchmarking criteria, supplemental plots and simulations see Supplementary_Materials at http://rfajriyah.staff.uii.ac.id/category/supplementary/.

3.1. Exponential-lognormal convolution

Background correction Consider model (1) when the true intensity S_{ij} is exponentially distributed, i.e.

$$S_{ij} \sim f_1(s_{ij};\theta_j) = \theta_j \exp(-\theta_j s_{ij}), \qquad \theta_j, s_{ij} > 0,$$

and the background noise B_{ij} is lognormally distributed, i.e.

$$B_{ij} \sim f_2(b_{ij};\mu_j,\sigma_j^2) = \frac{1}{b_{ij}\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log b_{ij}-\mu_j)^2}{2\sigma_j^2}\right), \qquad \mu_j \in \mathbb{R},\ \sigma_j^2, b_{ij} > 0.$$

Then the joint density function of S_{ij} and B_{ij} equals

$$f_{S_{ij},B_{ij}}(s_{ij},b_{ij}) = \frac{\theta_j \exp(-\theta_j s_{ij})}{b_{ij}\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log b_{ij}-\mu_j)^2}{2\sigma_j^2}\right),$$

and thus the joint density function of S_{ij} and P_{ij} is

$$f_{S_{ij},P_{ij}}(s_{ij},p_{ij}) = \frac{\theta_j \exp(-\theta_j s_{ij})}{(p_{ij}-s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij}-s_{ij})-\mu_j)^2}{2\sigma_j^2}\right), \qquad s_{ij} < p_{ij}.$$

Consequently, the marginal density function of P_{ij} equals

$$f_{P_{ij}}(p_{ij}) = \int_0^{p_{ij}} \frac{\theta_j \exp(-\theta_j s_{ij})}{(p_{ij}-s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij}-s_{ij})-\mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Using the substitution $\log(p_{ij} - s_{ij}) = z_{ij}$ we get

$$\begin{aligned}
f_{P_{ij}}(p_{ij}) &= \int_{-\infty}^{\log p_{ij}} \frac{\theta_j \exp(-\theta_j(p_{ij} - e^{z_{ij}}))}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\
&= \frac{\theta_j \exp(-\theta_j p_{ij})}{\sigma_j\sqrt{2\pi}} \int_{-\infty}^{\log p_{ij}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) \sum_{k=0}^{\infty} \frac{\theta_j^k e^{kz_{ij}}}{k!} dz_{ij} \\
&= \frac{\theta_j \exp(-\theta_j p_{ij})}{\sigma_j\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \int_{-\infty}^{\log p_{ij}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2} + kz_{ij}\right) dz_{ij} \\
&= \frac{\theta_j \exp(-\theta_j p_{ij})}{\sigma_j\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k}{2}\sigma_j^2\right)\right) \int_{-\infty}^{\log p_{ij}} \exp\left(-\frac{(z_{ij}-(\mu_j + k\sigma_j^2))^2}{2\sigma_j^2}\right) dz_{ij} \\
&= \theta_j \exp(-\theta_j p_{ij}) C_{a,j},
\end{aligned}$$

where

$$C_{a,j} = \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k}{2}\sigma_j^2\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + k\sigma_j^2)}{\sigma_j}\right).$$

The conditional density function of S_{ij} given $P_{ij} = p_{ij}$ is now obtained as

$$f_{S_{ij}|P_{ij}}(s_{ij}|p_{ij}) = \frac{\exp(\theta_j(p_{ij} - s_{ij}))}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}C_{a,j}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right)$$

with conditional mean

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{\exp(\theta_j p_{ij})}{C_{a,j}} \int_0^{p_{ij}} \frac{s_{ij}\exp(-\theta_j s_{ij})}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Using the substitution $\log(p_{ij} - s_{ij}) = z_{ij}$, this equals

$$\begin{aligned}
\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) &= \frac{p_{ij}}{C_{a,j}} \int_{-\infty}^{\log p_{ij}} \frac{\left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)\exp(\theta_j e^{z_{ij}})}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\
&= \frac{p_{ij}}{C_{a,j}} \left[C_{a,j} - \int_{-\infty}^{\log p_{ij}} \frac{\frac{e^{z_{ij}}}{p_{ij}}\exp(\theta_j e^{z_{ij}})}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-\mu_j)^2}{2\sigma_j^2}\right) dz_{ij}\right] \\
&= p_{ij} - \frac{\exp\left(\mu_j + \frac{\sigma_j^2}{2}\right)}{C_{a,j}} \int_{-\infty}^{\log p_{ij}} \frac{\exp(\theta_j e^{z_{ij}})}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-(\mu_j+\sigma_j^2))^2}{2\sigma_j^2}\right) dz_{ij} \\
&= p_{ij} - \frac{\exp\left(\mu_j + \frac{\sigma_j^2}{2}\right)}{C_{a,j}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k+2}{2}\sigma_j^2\right)\right) \\
&\qquad \times \int_{-\infty}^{\log p_{ij}} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij}-(\mu_j+(k+1)\sigma_j^2))^2}{2\sigma_j^2}\right) dz_{ij} \\
&= p_{ij} - \frac{C_{b,j}}{C_{a,j}} \exp\left(\mu_j + \frac{\sigma_j^2}{2}\right),
\end{aligned}$$

where

$$C_{b,j} = \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left(k\left(\mu_j + \frac{k+2}{2}\sigma_j^2\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + (k+1)\sigma_j^2)}{\sigma_j}\right).$$

Parameter estimation To estimate the parameters θ_j, μ_j, and σ_j^2, $j = 1, \ldots, J$ in the exponential-lognormal model we can use various methods.

1. Maximum likelihood estimation (MLE):
   This is implemented by applying the *optim* function in R to maximize the log-likelihood function of the jth array

$$\sum_{i=1}^{I}(\log\theta_j - \theta_j p_{ij} + \log C_{a,j;K}) + \sum_{m=1}^{M}\left(-\frac{(\log b_{0mj} - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2}\log(2\pi\sigma_j^2 b_{0mj}^2)\right),$$

   where p_{ij} and b_{0mj} are the observed values of P_{ij} and B_{0mj}. Note that $C_{a,j}$ in the log-likelihood function is defined by the infinite series at the end of the previous section. However, the terms of this infinite series decrease very rapidly and thus we can cut off the series at a proper index K giving $C_{a,j;K}$, making it suitable for its computation in R. The index K is chosen as the smallest integer for which $|(C_{a,j;K} - C_{a,j;K-1})/C_{a,j;K-1}| < 0.001$ holds.

2. Method of moments:
   Note that the method of moments estimator of the parameter θ in an exponential distribution is the reciprocal of the sample mean. Since the S_{ij} are not observable, but the B_{0mj} are, we consider Equation (1) and estimate $1/\theta_j$ by the difference $\text{mean}(p_{ij}) - \text{mean}(b_{0mj})$. Further, the parameters μ_j and σ_j^2 in the lognormal part are estimated by the sample mean and variance of the observed $\log b_{0mj}$ values.

3. Plug-in estimator:

   (a) We calculate the MLE of the parameter in the model of S_{ij} by utilizing the observed sample p_{ij} instead of s_{ij}. This is justified by the fact that, in some sense, the distribution of S_{ij} is similar to the distribution of P_{ij}.
   (b) We estimate μ_j and σ_j^2 through MLE based on B_{0mj} as described above.

3.2. Gamma-lognormal convolution

Background correction Consider now model (1) and assume that the true intensity S_{ij} is gamma distributed, i.e.

$$S_{ij} \sim f_1(s_{ij}; \alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1}\exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)}, \qquad \alpha_j, \beta_j, s_{ij} > 0,$$

and that the background noise B_{ij} is lognormally distributed, i.e.

$$B_{ij} \sim f_2(b_{ij}; \mu_j, \sigma_j^2) = \frac{1}{b_{ij}\sigma_j\sqrt{2\pi}}\exp\left(-\frac{(\log b_{ij} - \mu_j)^2}{2\sigma_j^2}\right), \qquad \mu_j \in \mathbb{R},\ \sigma_j^2 > 0.$$

The joint density function of S_{ij} and B_{ij} is

$$f_{S_{ij},B_{ij}}(s_{ij}, b_{ij}) = \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1}\exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)} \frac{1}{b_{ij}\sigma_j\sqrt{2\pi}}\exp\left(-\frac{(\log b_{ij} - \mu_j)^2}{2\sigma_j^2}\right)$$

and thus the joint density function of S_{ij} and P_{ij} is

$$f_{S_{ij},P_{ij}}(s_{ij}, p_{ij}) = \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1}\exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)} \frac{1}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}}\exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right).$$

Hence, the marginal density function of P_{ij} is obtained as

$$f_{P_{ij}}(p_{ij}) = \int_0^{p_{ij}} \frac{\beta_j^{\alpha_j} s_{ij}^{\alpha_j - 1} \exp(-s_{ij}\beta_j)}{\Gamma(\alpha_j)} \frac{1}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Using the substitution $\log(p_{ij} - s_{ij}) = z_{ij}$ we get

$$\begin{aligned} f_{P_{ij}}(p_{ij}) &= \int_{-\infty}^{\log p_{ij}} \frac{\beta_j^{\alpha_j} p_{ij}^{\alpha_j - 1} \left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)^{\alpha_j - 1} \exp(-p_{ij}\beta_j + e^{z_{ij}\beta_j})}{\Gamma(\alpha_j)\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij} - \mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\ &= \frac{\beta_j^{\alpha_j} p_{ij}^{\alpha_j - 1} \exp(-p_{ij}\beta_j)}{\Gamma(\alpha_j)\sigma_j\sqrt{2\pi}} \int_{-\infty}^{\log p_{ij}} \left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)^{\alpha_j - 1} \exp(e^{z_{ij}\beta_j}) \exp\left(-\frac{(z_{ij} - \mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\ &= \frac{\beta_j^{\alpha_j} p_{ij}^{\alpha_j - 1} \exp(-p_{ij}\beta_j)}{\Gamma(\alpha_j)} C_{c,j}, \end{aligned}$$

where

$$C_{c,j} = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j - 1}{k}}{p_{ij}^k n!} \beta_j^n \exp\left((k+n)\left(\mu_j + (k+n)\frac{\sigma_j^2}{2}\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\right).$$

The conditional density function is now obtained as

$$f_{S_{ij}|P_{ij}}(s_{ij}|p_{ij}) = \frac{\exp((p_{ij} - s_{ij})\beta_j) s_{ij}^{\alpha_j - 1}}{C_{c,j} p_{ij}^{\alpha_j - 1}(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right)$$

with respective conditional mean

$$\mathsf{E}(S_{ij}|P_{ij} = p_{ij}) = \frac{e^{p_{ij}\beta_j}}{C_{c,j} p_{ij}^{\alpha_j - 1}} \int_0^{p_{ij}} \frac{s_{ij}^{\alpha_j} e^{-s_{ij}\beta_j}}{(p_{ij} - s_{ij})\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(\log(p_{ij} - s_{ij}) - \mu_j)^2}{2\sigma_j^2}\right) ds_{ij}.$$

Substituting $\log(p_{ij} - s_{ij}) = z_{ij}$ this becomes

$$\begin{aligned} \mathsf{E}(S_{ij}|P_{ij} = p_{ij}) &= \frac{p_{ij}}{C_{c,j}} \int_{-\infty}^{\log p_{ij}} \frac{\left(1 - \frac{e^{z_{ij}}}{p_{ij}}\right)^{\alpha_j} \exp\left(e^{z_{ij}\beta_j}\right)}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(z_{ij} - \mu_j)^2}{2\sigma_j^2}\right) dz_{ij} \\ &= \frac{p_{ij} C_{d,j}}{C_{c,j}}, \end{aligned} \tag{5}$$

where

$$C_{d,j} = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j}{k}}{p_{ij}^k n!} \beta_j^n \exp\left((k+n)\left(\mu_j + (k+n)\frac{\sigma_j^2}{2}\right)\right) \Phi\left(\frac{\log p_{ij} - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\right).$$

Parameter estimation To estimate the parameters α_j, β_j, μ_j, and σ_j^2 in (5), we can use either of the following methods.

1. Maximum likelihood estimation:
   This is implemented by applying the *optim* function in R to maximize the log-likelihood function of the jth array

$$\begin{aligned} &\sum_{i=1}^{I} \left(\log(C_{c,j;K}) + (\alpha_j - 1)\log(p_{ij}) - p_{ij}\beta_j + \alpha_j \log(\beta_j) - \log(\Gamma(\alpha_j))\right) \\ &+ \sum_{m=1}^{M} \left(-\frac{(\log b_{0mj} - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2}\log(2\pi\sigma_j^2 b_{0mj}^2)\right). \end{aligned}$$

   Similarly to the exponential-lognormal model, in the computation of $C_{c,j}$ the cutoff index K is chosen according to the criteria $|(C_{c,j;K} - C_{c,j;K-1})/C_{c,j;K-1}| < 0.002$.

2. Method of moments:
The implementation of this method for the jth array is done by recalling that in case of a gamma distribution with parameters α_j and β_j, the method of moments estimator for α_j/β_j and α_j/β_j^2 is the sample mean and the sample variance, respectively. Thus, considering Equation (1) we get

$$\hat{\beta}_j = \frac{\text{mean}(s_{ij})}{\text{var}(s_{ij})} = \frac{\text{mean}(p_{ij}) - \text{mean}(b_{0mj})}{\text{var}(p_{ij}) - \text{var}(b_{0mj})},$$

as also

$$\hat{\alpha}_j = \hat{\beta}_j \text{ mean}(s_{ij}) = \frac{(\text{mean}(s_{ij}))^2}{\text{var}(s_{ij})} = \frac{(\text{mean}(p_{ij}) - \text{mean}(b_{0mj}))^2}{\text{var}(p_{ij}) - \text{var}(b_{0mj})}.$$

Furthermore, μ_j and σ_j^2 are estimated by the mean and variance of $\log b_{0mj}$.

3. Plug-in estimator:
In equation (1), P_{ij} and B_{0mj} are observable intensities. Therefore, the plug-in estimator is implemented by

   (a) estimating α_j and β_j through MLE based on the p_{ij} values, and
   (b) estimating μ_j and σ_j^2 through MLE based on the b_{0mj} values.

3.3. Benchmarking

Benchmarking data set Illumina platform has provided a benchmarking data set, the Illumina spike-in Dunning, Barbosa-Morais, Lynch, Tavaré, and Ritchie (2008). These spike-in probes are targeting bacterial and viral genes absent from the mouse genome. These were added at specific concentrations on each sample. Therefore the change in expression level of a particular spike between samples is known a priori. The expression levels of the non-spikes should not change between samples.

There are twelve different concentrations of spike: 1000 picomolar (pM), 300, 100, 30, 10, 3, 1, 0.3, 0.1, 0.03, 0.01, and 0.00 pM. It was replicated four times. Therefore, there are 48 samples and each sample has regular and control bead-type level probes.

There are approximately about 48000 bead-type level probes for each sample and in addition the 33 spike-in bead-type level probes are added into it. For the control, there are 1616 bead-type level probes. These control experiments are the benchmarking data sets of Illumina and are used to compare low-level analysis methods such as in Affymetrix platform.

Performance studies We compare all convolution models: Irizarry *et al.* (2003a,b, 2006) and Bolstad *et al.* (2003): RMA (Exponential-Normal), Plancade *et al.* (2011, 2012): Gamma-Normal, Chen *et al.* (2011): Exponential-Gamma, Xie *et al.* (2009): Exponential-Normal adjusted for Illumina BeadArrays with MLE for the parameters, Bayesian approach and the method of moments, and the proposed models: exponential-lognormal and gamma-lognormal.

We will call the methods above, respectively, as follows: ENr, GN, EG, ENm, ENm, ENn, ELNn, ELNm, ELNp, GLNn, GLNm, and GLNp. We use the MBCB package (Allen *et al.* (2009) and Xie *et al.* (2009)) to adjust the intensity values of these existing models ENr, ENm, ENmc and ENn. Except that, the GN uses the NormalGamma package (Plancade *et al.* (2011)).

Table 1 shows that the GLNn reproduces the Illumina concentration better than others. The ENr shows the closest performance toward the GLNn. Note that the computation of the Kullback-Leibler coefficient is implemented in each array j based on the nominal concentrations O in Table 1 and the observed intensities P in Table 2, and the value in each

Table 1: Reproducibility of each method toward the Illumina spike-in concentration

Model	RMSE	K-L
ENr	1.346	51310
ENn	1.407	41010
ENm	1.483	23170
ENmc	1.483	23170
EG	1.470	20660
GN	1.521	58480
ELNn	1.411	41200
ELNm	1.489	21280
ELNp	1.423	37800
GLNn	**1.323**	**4333**
GLNm	1.510	29630
GLNp	10.700	−115400

Table 2: Reproducibility of each method toward the Illumina spike-in based on the experiment data

Model	RMSE	K-L
ENr	7.251	1141000
ENn	7.127	1062000
ENm	6.927	926500
ENmc	6.927	926200
EG	6.919	907900
GN	7.100	1183000
ELNn	7.124	1062000
ELNm	6.904	911600
ELNp	7.092	1035000
GLNn	**6.825**	**793400**
GLNm	6.937	968400

table is the median Kullback-Leibler coefficient based on $J = 42$ arrays. The Kullback-Leibler coefficient for two positive sequences $(X_{1j}, \ldots, X_{Ij}), (S_{1j}, \ldots, S_{Ij})$ is computed as $\text{K-L}_j = \sum_{i=1}^{I} X_{ij} \log(X_{ij}/S_{ij})$, which can also be negative if $S_{ij} > X_{ij}$ for all or for most i. This is a sign that the S are overestimating X, where X could be O (Table 1) or P (Table 2).

Therefore we exclude the GLNp model from further comparisons. The behavior of GLNp which is different from other models, also shown at the supplemental plots.

Table 2 shows how each method reproduces the data from the experiment. We see that GLNn can be considered to reproduce it better than others, based on the RMSE, and the Kullback-Leibler coefficient. Tables 1 and 2 provide insight about how the performance comparison among the models would be conducted further.

In the first part, we compute the adopted Affycomp benchmarking criteria, based on the data after background correction and their log transformation. In the second part, in the simulation, the MSE_{bc} and the L_1 error will be computed based on the log transformation of the experiment and the nominal concentration data.

The log transformations that we use here, respectively, for the benchmarking and the FFPE data sets are as follows

$$y = \log_2\left(x + \sqrt{x^2+1}\right) \qquad \text{and} \qquad y = \log_2\left(x + 1 + \sqrt{x^2+1}\right),$$

Table 3: Median SD, IQR, and 99.9% percentiles of log fold change for non spike-in between replicates for each model.

Model	Median SD	IQR	99.9%
ENr	**0.027**	**0.062**	0.415
ENn	0.043	0.089	0.441
ENm	0.069	0.139	0.486
ENmc	0.069	0.139	0.486
EG	0.065	0.134	0.477
GN	0.051	0.098	0.520
ELNn	0.045	0.093	0.442
ELNm	0.071	0.145	0.489
ELNp	0.049	0.100	0.449
GLNn	0.038	0.075	**0.398**
GLNm	0.076	0.080	0.507

Table 4: The signal detect R^2 by regressing the nominal and observed value for each model for the Illumina spike-in.

Model	R^2	Low.R^2	Med.R^2	High.R^2
ENr	0.959	0.618	**0.698**	**0.559**
ENn	0.958	0.622	0.695	0.557
ENm	0.957	0.635	0.695	0.558
ENmc	0.957	0.635	0.695	0.558
EG	0.957	0.633	0.695	0.558
GN	0.956	**0.650**	0.697	0.555
ELNn	0.958	0.624	0.695	0.557
ELNm	0.957	0.636	0.694	0.558
ELNp	0.958	0.627	0.695	0.557
GLNn	**0.960**	0.609	0.696	0.558
GLNm	0.956	0.637	0.694	0.558

where x is the nominal concentration O or the observed intensity value P.

First part In Table 3 it is shown that the ENr method provides the smallest variation and IQR and the GLNn model provides the smallest 99.9% percentiles of log fold change for the non spike-in between replicates. The largest variation, IQR, and 99.9% percentiles, respectively, are observed for the GLNm, the ELNm, and the GN method.

In Table 4 it is shown that, in general, all methods perform similar to each other. The GLNn models have the highest signal detect R^2. The GN model has the highest R^2 at low concentration but has the lowest R^2 at high concentration. This means that the GN model works better at low concentration. On the other hand the ENr shows that it works better at medium and high concentrations, which is followed closely by GLNn model.

If we divide the concentrations into two categories, where high concentration means that the nominal concentration is at least 3 pM and low concentration means that the nominal concentration is at most 1 pM, the GLNn model has the highest R^2 (the data is not shown here). It means, in general and at high concentrations, the GLNn offers a better fit than other models. As in Table 4, Table 5 shows that all models have similar performance, although the GLNn model has the highest R^2 of nominal concentration against observed log-fold-change. Table 6 provides the results from the computation of the AUC value. The table shows that all models have a better accuracy at medium concentrations than at low and high concentrations. The ENr performs very poor at the low concentrations, but the GLNm performs best. At high

Table 5: The R^2 observed log-fold-change against nominal log-fold-changes for the spike-in genes.

Model	Obs-intended-fc.R^2	Obs-(low)-int-fc.R^2
ENr	0.976	0.989
ENn	0.974	0.990
ENm	0.972	0.985
ENmc	0.972	0.985
EG	0.972	0.986
GN	0.970	0.987
ELNn	0.974	0.990
ELNm	0.972	0.985
ELNp	0.973	0.990
GLNn	**0.978**	**0.991**
GLNm	0.971	0.984

Table 6: AUC value for each model.

Model	Low concentration AUC	Medium concentration AUC	High concentration AUC	Average AUC	All
ENr	0.450	0.987	**0.785**	0.585	0.886
ENn	0.518	0.987	0.764	0.631	0.899
ENm	0.573	0.987	0.741	0.667	0.911
ENmc	0.573	0.987	0.741	0.667	0.911
EG	0.567	0.987	0.746	0.664	0.910
GN	0.552	0.987	0.723	0.651	0.904
ELNn	0.524	0.987	0.763	0.635	0.900
ELNm	0.574	0.987	0.741	0.668	0.912
ELNp	0.534	0.987	0.761	0.642	0.902
GLNn	0.498	**0.987**	0.784	0.619	0.896
GLNm	**0.579**	0.987	0.730	**0.671**	**0.913**

concentrations, the ENr performs the best and it is followed by the GLNn. But in general, the highest AUC is achieved by all the model with the MLE parameter estimation methods: the GLNm, ELNm, and ENm.

The computation, which is based on all arrays, provides the results where all models have the AUC greater than 0.9. According to Zhu, Zeng, and Wang (2010), the AUC between 0.9 and 1.0 is classified as excellent in measuring the accuracy. Therefore, based on Table 6, we can identify that there are some models excellency accurate in predicting the gene expression.

Second part We do $N = 100$ simulations to assess the performance of each model. The bias of the background correction is assessed by the MSE_{bc}, and the bias of the parameterization is assessed by the L_1 error.

From the simulation results in Table 7 we can see that results for the EG model are not available, because the MBCB package did not work at the log transformation that we have chosen. The GN model performs best, by providing the smallest bias for the background correction and the parameters. A close performance is achieved by the ELN, particularly by the ELNn. The GLNn does not have an optimal performance on the MSE_{bc}, but we still can consider its performance good, concerning that the bias of the parameters are similar to other proposed models and GN.

One of the proposed models, GLNm has the highest bias on the MSE_{bc} and the parameter α. In our view this happens because we use an approximation in estimating the true intensity

Table 7: The simulation results on the spike-in data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	0.045		0.664	46.58	11.44
ENn	0.049		0.625	41.61	2.806
ENm	0.038		0.610	58.92	2.040
ENmc	0.036		0.610	62.77	2.039
GN	**0.030**	**0.0003**	**0.007**	0.013	**0.015**
ELNn	0.048		0.009	0.0004	0.018
ELNm	0.039		0.840	0.0004	0.018
ELNp	0.061		0.472	0.0004	0.018
GLNn	0.216	0.052	0.055	0.0004	0.018
GLNm	84.37	38.86	0.851	**0.0003**	0.017

value. The EN models (ENr, ENm, ENn and ENmc) have considerably better performance at MSE_{bc}, but are not good at the parametrization. The bias on the parametrization of the noise is higher than in other models.

3.4. The public data sets

Based on the results from Section 3.3, we compare the performance of these models on some public data sets. We would like to know how good these models are in real data samples. Here, we choose to use the formalin-fixed, paraffin-embedded (FFPE) data sets from Waldron, Ogino, Hoshida, Shima, Reed, Simpson, Baba, Nosho, Segata, Vargas, Cummings, Lakhani, Kirkner, Giovannucci, Quackenbush, Golub, Fuchs, Parmigiani, and Huttenhower (2012): the FFPE of tumors from colorectal cancer patients (GSE32651, 1003 samples), breast cancer metastases of the lymph node and autopsy tissues (GSE32490: GSE32489, 120 samples). Each sample has 24526 bead-type level probes.

The links for the data set are `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32651` and `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32490`.

Currently the FFPE archival samples are widely available in million and it is a great source of information in medical studies about some diseases, for example cancer. This data type is suffering from the RNA degradation, which leads to poor performance in array-based studies. However, the Illumina's DASL assays could provide high-quality data from this degraded RNA samples.

Comparing the performance of these background correction models certainly would help researchers to choose the appropriate background correction for their data, particularly if their data is the FFPE type.

The background correction for the FFPE data set is implemented in three steps:

step 1 Do the quality control (QC) to the raw FFPE data. In this paper, we used the *ffpe* package in R (Waldron (2013)).

step 2 Do the data transformation $\log_2(P_{ij} + 1 + \sqrt{P_{ij}^2 + 1}$ to the raw FFPE data after QC and estimate the background correction parameters based on it. The estimators of true intensity value and the background correction are based on the regular and negative control probe intensity data, respectively.

step 3 Compute the true intensity value (the adjusted intensity estimator) based on the BC parameters at step 2.

Table 8: Simulation results on the GSE32651 data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	0.058		0.657	297.67	22.61
ENn	0.281		0.572	1.984	2.627
ENm	0.086		0.591	28.05	1.715
ENmc	**0.036**		0.610	62.77	2.039
ELNn	0.275		**0.025**	0.001	0.018
ELNm	0.059		0.826	**0.0004**	0.018
ELNp	0.672		0.487	0.001	0.018
GLNn	0.838	**0.340**	0.527	0.001	0.018
GLNm	84.15	71.87	0.887	0.0005	**0.018**

Table 9: Simulation results on the GSE32489 data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	**0.093**		0.665	67.17	9.511
ENn	0.863		0.509	0.936	2.039
ENm	0.182		0.558	14.20	1.712
ENmc	0.184		0.556	14.18	1.049
ELNn	1.055		0.857	0.002	0.018
ELNm	0.116		0.781	0.001	0.018
ELNp	1.247		0.461	0.002	0.018
GLNn	1.348	**0.332**	**0.497**	0.002	**0.018**
GLNm	164.98	22.24	0.805	**0.0004**	**0.018**

The results of our computation are in Tables 8 and 9. From these tables we can see that there are no EG and GN models. Neither of these models can work on these data sets. For some samples in the data set, both models fail to compute the parameters which has the consequence that the true intensity value cannot be provided.

We decided to remove the EG and GN models from further comparisons in both FFPE data sets. Here we provide the results of the rest of the models only.

Tables 8 and 9 consistently show that the bias of the parameters of noise in the EN models are higher than the proposed models. For the parameter β, the ELNn has the smallest bias and it is followed by the ELNp and the GLNn. With regard to the bias of the background correction, the EN models show the smallest bias in both of FFPE data sets.

4. Conclusions and indication of future work

We have studied additive models of background correction for BeadArrays and proposed some new models where the true intensity is assumed to have exponential or gamma distribution and the noise is lognormally distributed. We have derived the estimator of the true intensity value of the proposed models.

Further, we compared the performance of all models, based on the benchmarking and public data sets. In the benchmarking data set we adopted the criteria from the Affycomp (Cope *et al.* 2004) and for the simulation study we used the criteria which have been used in Xie *et al.* (2009), Chen *et al.* (2011) and Plancade *et al.* (2011, 2012). For the public data sets, we only used the criteria for the simulation study.

We have seen in Sections 3.3 and 3.3 that EN, EG, GN and GLN perform rather similar. However, the GLNn model has provided the highest reproducibility in comparison to other

models. From the Affycomp criteria we can provide the following points:

1. The ENr and GLNn provide the lowest variation between replicates and all models using the MLE estimation method have a higher variation than others.

2. The GLNn model has the highest signal detect R^2 in general and in high concentration. This means the GLNn model is the best fitted for the gene expression.

3. The GLNn model, based on the MvA plot, produces the least number of genes which should not be expressed but are nevertheless expressed. On the other hand, the GN model provides the largest number of such genes.

4. All models with the MLE estimation method have a higher average AUC value, which means that they provide a better accuracy in predicting the gene expression.

5. The ENr and GLNn have the lowest IQR of log fold-change between replicates.

6. Points 1 and 2 show that the GLNn and ENr are more accurate and precise in modelling the gene expression and points 3 and 5 show that the specificity and sensitivity of the GLNn and ENr model are better than others.

In the simulation study, the best performance in estimating the signal by measuring its background correction and parametrization errors is achieved by the GN model. It is followed by our proposed ELN models. It has been shown that the GLNn does not perform optimally at the MSE_{bc} criterion, but for the parametrization this model still can be considered good.

In the FFPE public data set, the GN and EG models cannot be implemented. This is in strong contrast with the fact that in the simulation study of benchmarking data set, the GN model has the best performance.

The EN models show the highest bias in the parametrization in both public data sets and the lowest bias in the background correction. Our proposed models, except the GLNm, show the lowest bias in the parametrization in both data sets and a moderate bias in the background correction.

Based on the results from the benchmarking data and the public data sets, we would suggest researchers the following:

1. if the GN model works properly at the data set at hand (i.e. the estimated signals in all arrays can be computed by this model and the simulation criteria for this data at this model are low) then use the GN model to correct the background.

2. if the GN model fails, then use our proposed models, particularly the GLNn model. The reason for not choosing the ELN models is that the value of the parameter α from the benchmarking data set is less than 1, around 0.2. Therefore, the gamma model is more appropriate to model the true intensity distribution than the exponential one. We believe that the right approximative computation of the GLN models will lead to a better performance than the current approximation.

   The ELN models perform better than the original EN models, due to the fact that not only the regular probes, but also the control probes are skew-distributed (Chen *et al.* 2011). Therefore, these models could be the second choice after the GLN, when the GN model does not work.

3. With regard to the computation time, at the benchmarking data set the EN models are working faster than the others. They are followed by the ELNp, ELNn, and EG. The GLNn and the ENmc are the third fastest, then come the GN and the ELNm, which are followed by the GLNm as the slowest one.

One of the purposes of using microarray technology is finding the genes which are expressed differentially due to some disease or condition. Therefore, it is important to investigate the effect of bias of the background correction and the parametrization toward the differentially expressed genes. This will be our future work.

Acknowledgements

This paper is part of the author's Ph.D dissertation written under the direction of Prof. István Berkes. We would like to thank Herwig Friedl, Ernst Stadlober, Levi Waldron, and an anonymous referee for their comments leading to a substantial improvement of the paper. We thank Florian Endel for discussion and suggestion how to make the R programming more efficient. We thank Yevgeniy Grigoryev and Ken Laing for their generous permission to use their pictures in the Supplementary material at http://rfajriyah.staff.uii.ac.id/category/supplementary/. Financial support from the Austrian Science Fund (FWF), Project P24302-N18 is gratefully acknowledged.

References

Allen JD, Chen M, Xie Y (2009). "Model-Based Background Correction (MBCB): R Methods and GUI for Illumina Bead-array Data." *Journal of Cancer Science and Therapy*, **1**(1), 25–27.

Baek J, Son YS, MacLachlan GJ (2007). "Segmentation and Intensity Estimation of Microarray Images Using a Gamma-t Mixture Mode." *Bioinformatics*, **23**(4), 458–465.

Bolstad BM (2004). *Low Level Analysis of High-Density Oligonucleotide Array Data: Background, Normalization and Summarization*. Ph.D. thesis, University of California, California, Berkeley.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003). "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance." *Bioinformatics*, **19**(2), 185–193.

Chen M, Xie Y, Story MD (2011). "An Exponential-Gamma Convolution Model for Background Correction of Illumina BeadArray Data." *Communication in Statistics: theory and methods*, **40**(17), 3055–3069.

Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP (2004). "A Benchmark for Affymetrix GeneChip Expression Measures." *Bioinformatics*, **20**, 323–331.

Ding LH, Xie Y, Park S, Xiao G, Story MD (2008). "Enhanced Identification and Biological Validation of Differential Gene Expression via Illumina Whole-Genome Expression Arrays through the Use of the Model-Based Background Correction Methodology." *Nucleic Acids Research*, **36**(10: e58).

Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavaré S, Ritchie ME (2008). "Statistical Issues in the Analysis of Illumina Data." *BMC Bioinformatics*, **9**(85).

Forcheh AC, Verbeke G, Kasim A, Lin D, Shkedy Z, Talloen W, Gohlmann HW, Clement L (2012). "Gene Filtering in the Analysis of Illumina Microarrays Experiments." *Statistical Applications in Genetics and Molecular Biology*, **11**(2).

Hochreiter S, Djork-Arné, Obermayer K (2006). "A New Summarization Method for Affymetrix Probe Level Data." *Bioinformatics*, **22**(8), 943–949.

Huber W, Irizarry RA, Gentleman R (2005a). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter Preprocessing Overview. Springer.

Huber W, von Heydebreck A, Vingron M (2004). "Error Models for microarray Intensities." *Technical Report Paper 6*, Bioconductor Project Working Papers.

Huber W, von Heydebreck A, Vingron M (2005b). "An Introduction to Low-Level Analysis Methods of DNA Microarray Data." *Technical Report Paper 9*, Bioconductor Project Working Papers.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003a). "Summaries of Afymetrix GeneChip Probe Level Data." *Nucleic Acids Research*, **31**(4).

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003b). "Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics*, **4**(2), 249–264.

Irizarry RA, Wu Z, Jaffee HA (2006). "Comparison of Affymetrix geneChip Expression Measures." *Bioinformatics*, **22**(7), 789–794.

Li C, Wong WH (2001). "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection." *Proceeding national Academy of Sciences*, **98**(1), 31–36.

Plancade S, Rozenholc Y, Lund E (2011). "Improving Background Correction for Illumina BeadArrays: the Normal-Gamma Model."

Plancade S, Rozenholc Y, Lund E (2012). "Generalization of the Normal-Exponential Model: Exploration of a More Accurate Parameterisation for the Signal Distribution on Illumina BeadArrays." *BMC Bioinformatics*, **13**(329).

Shi W, Oshlack A, Smyth GK (2010). "Optimizing the Noise versus Bias Trade-off for Illumina Whole Enome Expression Beadchips." *Nucleic Acids Research*, **38**(22: e204).

Silver JD, Ritchie ME, Smyth GK (2009). "Microarray Background Correction: Maximum Likelihood Estimation for the Normal-Exponential Convolution Model." *Biostatistics*, **10**, 352–363.

Triche TJ, Weisenberger DJ, Berg DVD, Laird PW, Siegmund KD (2013). "Low-level Processing of Illumina Infinium DNA Methylation BeadArrays." *Nucleic Acids Research*, pp. 1–11.

Waldron L (2013). *ffpe: Quality Assessment and Control for FFPR Microarray Expression Data*, r package version 1.4.0 edition.

Waldron L, Ogino S, Hoshida Y, Shima K, Reed AEM, Simpson PT, Baba Y, Nosho K, Segata N, Vargas AC, Cummings M, Lakhani SR, Kirkner GJ, Giovannucci E, Quackenbush J, Golub TR, Fuchs CS, Parmigiani G, Huttenhower C (2012). "Expression Profiling of Archival Tumors for Long-term Health Studies." *Clinical Cancer Research*.

Wu Z, Irizarry RA, 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.

Xie Y, Wang X, Story MD (2009). "Statistical Methods of Background Correction for Illumina BeadArray Data." *Bioinformatics*, **25**(6), 751–757.

Zhang J (2013). "Reducing the Bias of the Maximum Likelihood Estimator of the Shape Parameter for the Gamma Distribution." *Computational Statistics*, **28**, 1715 – 1724.

Zhu W, Zeng N, Wang N (2010). "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations." Http://www.nesug.org/Proceedings/nesug10/hl/hl07.pdf.

Affiliation:

Rohmatul Fajriyah
Institute of Statistics
Graz University of Technology
8010 Graz, Austria
E-mail: fajriyah@student.tugraz.at

Department of Statistics
Universitas Islam Indonesia
Jl. Kaliurang Km. 14.4
55584 Jogjakarta, Indonesia
E-mail: rfajriyah@fmipa.uii.ac.id
URL: http://rfajriyah.staff.uii.ac.id/

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 45
March 2016

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-11-13
Accepted: 2014-11-13