Statistical Applications in Genetics and Molecular Biology

Volume 9, *Issue* 1 2010 *Article* 4

Informative or Noninformative Calls for Gene Expression: A Latent Variable Approach

Adetayo Kasim, *Universiteit Hasselt & Katholieke Universiteit Leuven*
Dan Lin, *Universiteit Hasselt & Katholieke Universiteit Leuven*
Suzy Van Sanden, *Universiteit Hasselt & Katholieke Universiteit Leuven*
Djork-Arné Clevert, *Johannes Kepler University Linz & Charité - Universitätsmedizin Berlin*
Luc Bijnens, *Janssen Pharmaceutica N.V., Beerse*
Hinrich Göhlmann, *Janssen Pharmaceutica N.V., Beerse*
Dhammika Amaratunga, *Johnson & Johnson Pharmaceutical Research & Development, Raritan*
Sepp Hochreiter, *Johannes Kepler University Linz*
Ziv Shkedy, *Universiteit Hasselt & Katholieke Universiteit Leuven*
Willem Talloen, *Janssen Pharmaceutica N.V., Beerse*

Informative or Noninformative Calls for Gene Expression: A Latent Variable Approach*

Adetayo Kasim, Dan Lin, Suzy Van Sanden, Djork-Arné Clevert, Luc Bijnens, Hinrich Göhlmann, Dhammika Amaratunga, Sepp Hochreiter, Ziv Shkedy, and Willem Talloen

Abstract

The strength and weakness of microarray technology can be attributed to the enormous amount of information it is generating. To fully enhance the benefit of microarray technology for testing differentially expressed genes and classification, there is a need to minimize the amount of irrelevant genes present in microarray data. A major interest is to use probe-level data to call genes informative or noninformative based on the trade-off between the array-to-array variability and the measurement error. Existing works in this direction include filtering likely uninformative sets of hybridization (FLUSH; Calza et al., 2007) and I/NI calls for the exclusion of noninformative genes using FARMS (I/NI calls; Talloen et al., 2007; Hochreiter et al., 2006). In this paper, we propose a linear mixed model as a more flexible method that performs equally good as I/NI calls and outperforms FLUSH. We also introduce other criteria for gene filtering, such as, R2 and intra-cluster correlation. Additionally, we include some objective criteria based on likelihood ratio testing, the Akaike information criteria (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978).

Based on the HGU-133A Spiked-in data set, it is shown that the linear mixed model approach outperforms FLUSH, a method that filters genes based on a quantile regression. The linear model is equivalent to a factor analysis model when either the factor loadings are set to a constant with the variance of the latent factor equal to one, or if the factor loadings are set to one together with unconstrained variance of the latent factor. Filtering based on conditional variance calls a probe set informative when the intensity of one or more probes is consistent across the arrays, while filtering using R2 or intra-cluster correlation calls a probe set informative only when average intensity of a probe set is consistent across the arrays. Filtering based on likelihood ratio test AIC and BIC are less stringent compared to the other criteria.

KEYWORDS: gene filtering, factor analysis, linear mixed model

---

*Financial support from the IAP research network nr P6/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged. We are grateful to the reviewers for their insightful comments, which resulted in an improved manuscript.

1 Introduction

Microarray technology has been used extensively in biomedical research due to its ability to simultaneously measure the expression levels of thousands of genes in a biological sample. The strength and weakness of microarray technology can be attributed to the enormous amount of information which is generating. Not all genes are expected to be informative. First, many genes are not expressed at biologically meaningful or at detectable levels. Most tissues express only 30 - 40% of their genes (Su *et al.*, 2002). Second, even among the expressed genes, only a very small fraction is expected to be differentially expressed under different experimental conditions (Calza *et al.*, 2007). The noisy genes with irrelevant variation often lead to false positives in the identification of the differentially expressed genes (Dudoit *et al.*, 2002).

Given the importance of gene filtering, two methods have recently been proposed to assess the signal-to-noise ratio for every probe set. Talloen *et al.* (2007) proposed a filtering method based on the informative or non-informative calls (I/NI Calls) for probe sets. The I/NI Calls makes use of the concept of factor analysis, where probes in a probe set are assumed to measure the same latent variable (that is, the true expression level of a gene). Conditional on the observed data, the variance of the estimated latent variable should be less than 0.5 for the probe set to be called informative. The I/NI Calls is based on the factor analysis for robust microarray summarization (FARMS) method proposed by Hochreiter *et al.* (2006). Calza *et al.* (2007) proposed the Filtering Likely Uninformative Sets of Hybridizations (FLUSH) method. In contrast to the I/NI Calls, the FLUSH method is based on a probe set-specific linear model, where probes and arrays are treated as fixed effects. The FLUSH method captures the array-to-array variability with a Chi-squared statistic, expressed as a function of array-specific effects and its covariance matrix. The FLUSH eventually calls a gene informative based on a quantile regression comparing the array-to-array variability and the measurement error.

In this paper, we propose a gene filtering method based on a linear mixed model. We compare its performance to the I/NI Calls (Talloen *et al.*, 2007) and to the FLUSH (Calza *et al.*, 2007) and demonstrate how it is related to the factor analysis models. The probe set-specific linear mixed model treats the array-to-array variability as a random effect, but considers probe-specific effects to be fixed. The hierarchical formulation of the linear mixed model implies a marginal model, where total variability can be decomposed as the sum of the array-to-array variability and the measurement error. Hence, the array-to-array variability can be expressed as a proportion of the total variability for a probe set. This proportion is referred to as the intra-cluster correlation.

It is a measure of coherence between intensities measured by different probes in a probe set. In this paper, we show that the informative or non-informative calls based on the mixed model and the factor analysis models are equivalent.

This paper is organized as follows: a landmark dataset, the HG-U133A Spiked-in is presented in Section 2. In Section 3, we discuss the concept of latent variable models and present several criteria for filtering non-informative probe sets. The first one is similar to the conditional variance implemented for the I/NI Calls and uses the proportion of variability explained by the latent variables from the total variability as a filtering criterion. Other criteria, such as the AIC (Akaike, 1973), the BIC (Schwarz, 1978), and significance testing are discussed as well. In Section 4, we compare the performance of the different methods using the HG-U133A Spiked-in data. In Section 5, we discuss the settings and results from simulation studies carried out to assess the performance of the gene filtering methods. In Section 6, we end the paper with a discussion.

2 HGU-133A Spiked-in Dataset

The Affymetrix HGU-133A Spiked-in dataset is publicly available for the purpose of determining the sensitivity and specificity of various methods for the analysis of microarray data. The dataset have an advantage over real-life datasets because the true number of differentially expressed genes are known. It contains known genes that are spiked-in at 14 different concentrations ranging from 0pM to 512pM, arranged in a Latin squared design. There are 42 arrays and 42 spiked-in probe sets equally distributed over the 14 concentrations. In addition to the original spiked-in transcripts, McGee and Chen (2006) discovered 22 additional probe sets that have similar characteristics as the spiked-in probe sets. Thus, the HGU-133A spiked-in dataset contains 64 spiked-in probe sets out of the 22,300 probe sets. To clarify, we refer to probe sets other than the spiked-in as background mixtures. The distribution of the number of probes per probe sets is presented in Table 1. The majority of the probe sets have 11 probes. The spiked-in probe sets consist of 52 probe sets with 11 probes and 12 probe sets with 20 probes. For simplicity, we refer to probe sets as genes.

DOI: 10.2202/1544-6115.1460

Table 1: Number of probes per probe sets in the HGU-133A spiked-in dataset.

# probes	Background	Spiked-in	Total
8	1	0	1
10	1	0	1
11	21713	52	21765
13	4	0	4
14	4	0	4
15	2	0	2
16	482	0	482
20	28	12	40
69	1	0	1
Total	22236	64	22300

3 Methodology

In this section, we discuss several approaches for gene filtering. We briefly discuss the FLUSH and the I/NI Calls approaches in Sections 3.1 and 3.2, respectively. We discuss the mixed model and confirmatory factor analysis approaches for gene filtering in Sections 3.3 and 3.4.

3.1 Filtering Likely Uninformative Sets of Hybridization (FLUSH)

The FLUSH method, proposed by Calza *et al.* (2007), uses probe level data to filter genes based on the trade-off between array-to-array variability and variability due to the measurement error. It models the perfect match (PM) data (on the log2 scale) after background correction using the so called ideal mismatch (IMM) to ensure positive values for a specific gene (Calza *et al.*, 2007). Let PM_{ij} and IMM_{ij} $(i = 1, 2, ..., n;\ j = 1, 2, ..., K)$ be the perfect match and ideal mismatch for the j^{th} probe and the i^{th} array, respectively. Calza et al. (2007) proposed the following linear model:

$$log_2(PM_{ij} - IMM_{ij}) = \mu_j + \alpha_i + \varepsilon_{ij}, \tag{1}$$

where μ_j and α_i are probe- and array-specific effects. The non-informative genes are those with small array-to-array variability, which is captured by the χ^2 statistic defined as

$$\chi^2 = \hat{\boldsymbol{\alpha}}' \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{\alpha}}, \tag{2}$$

where $\hat{\boldsymbol{\alpha}}$ is the vector of estimated array-specific effects and $\hat{\boldsymbol{V}}$ is its estimated covariance matrix. A non-parametric quantile regression smoothing with a user-specified quantile is fitted on the χ^2 statistic (on the squared root scale) as a function of the logarithm of residual standard deviation. Likely non-informative probe sets are probe sets whose χ^2 statistics are below the fitted quantile regression line. It is worth noting that the linear model conceptually assumes that the probes in a probe set are independent, which is contrary to the domain knowledge of the Affymetrix platform.

3.2 The I/NI Calls for the Exclusion of Non-informative Genes Using the FARMS

The I/NI Calls depend on the domain knowledge that all the probes in a probe set are expected to measure the true expression level of the designated gene. Since the true expression level is not known, it can be assumed to be a common latent factor. The factor loadings are determined under the assumption that the latent factor is normally distributed with mean zero and variance one. The underlying factor analysis model is given by

$$log_2(\boldsymbol{PM}_j) = \mu_j + \lambda_j \boldsymbol{z} + \varepsilon_j. \tag{3}$$

Here, μ_j is the probe-specific effect, and λ_j is the factor loading on probe j (Hochreiter *et al.*, 2006). The common latent factor is denoted as $\boldsymbol{z} \sim N(0,1)$, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Psi})$ is the measurement error, and $\boldsymbol{\Psi}$ is a diagonal covariance matrix. $\boldsymbol{\varepsilon}$ and $\boldsymbol{z}$ are considered to be independent. Assuming that the probe set intensities are centered around zero, that is, $\boldsymbol{x}_j = log_2(\boldsymbol{PM}_j) - \mu_j$, the marginal distribution of $\boldsymbol{x}$ is given by:

$$\boldsymbol{x} \sim N(0, \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}), \tag{4}$$

where $\boldsymbol{x}$ is a matrix of probe level data (on log_2 scale) after correcting for the probe-specific effects. The term $\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}$ is the model based covariance matrix, measuring the total variability in the data. As shown in Talloen *et al.* (2007), the conditional variance of the latent factor given the data is defined as $v(\boldsymbol{z}|\boldsymbol{x}) = (1 + \boldsymbol{\lambda}' \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda})^{-1}$, which is bounded between 0 and 1.

DOI: 10.2202/1544-6115.1460

Based on a threshold, a gene is called informative if its conditional variance is less than the specified threshold. The model was implemented for the I/NI Calls using a Bayesian approach, where they specified $N(0, \sigma_\lambda^2)$ as a prior for λ. The consequence of this prior is that the conditional variance of the latent factor given the data shrinks towards zero for the informative probe sets. Consequently, a threshold of 0.5 was proposed to discriminate between informative and non-informative probe sets.

3.3 A Mixed Model Approach

Let PM_{ij} be the j^{th} probe intensity of the perfect match measured on array i in a given probe set. Similar to the approaches of Talloen *et al.* (2007), Hochreiter *et al.* (2006), and Calza *et al.* (2007), we assume that the $log_2(PM_{ij})$ consists of two sources of variability. The first is the variability due to measurement error and the second is an array-to-array variability. Therefore, the following linear mixed model (Verbeke and Molenberghs, 2000) is assumed:

$$\begin{aligned} log_2(PM_{ij}) &= \mu_j + b_i + \varepsilon_{ij}, \\ i &= 1, \cdots, n, \qquad j = 1, \cdots, k, \end{aligned} \tag{5}$$

where b_i is an array-specific effect, $b_i \sim \mathrm{N}(0, \sigma_b^2)$, μ_j is a probe-specific effect, and $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. The linear mixed model specified in (5) is a random intercept model, which can be re-written in matrix notation as

$$log_2(\boldsymbol{PM_i}) = \boldsymbol{X_i}\boldsymbol{\mu} + \boldsymbol{Z_i}b_i + \boldsymbol{\varepsilon}_i, \tag{6}$$

where $\boldsymbol{PM}_i$ is a vector of probe level data. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the design matrices for the fixed effects and the random effects with known covariates, respectively, $\boldsymbol{\mu}$ is a vector of fixed effects of the probes, $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$ is a vector of the array-specific effects, and $\boldsymbol{\varepsilon}$ is a vector of the measurement error. For our specific setting, the design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are given by

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{Z}_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}.$$

The marginal distribution of $\boldsymbol{PM}_i$, i.e., taking into account the two sources of variability, is a multivariate normal distribution with the covariance matrix given by

$\sigma_b^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\varepsilon^2 I$ (Verbeke and Molenberghs, 2000). For a probe set with k probes, it is a $k \times k$ matrix, for which the qp^{th} entry is given by

$$\left[\sigma_b^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\varepsilon^2 I\right]_{qp} = \begin{cases} \sigma_b^2 + \sigma_\varepsilon^2 & q = p, \\ \sigma_b^2 & q \neq p. \end{cases}$$

Within the mixed model framework, the probe intensities measured on the same array form a cluster, and it is expected that observations within a cluster are correlated if they all measure the same true expression levels of the probe set. The probe set-specific intra-cluster correlation (Verbeke and Molenberghs, 2000) is given by

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2}. \tag{7}$$

Note that for the case, in which σ_b^2 is relatively larger than σ_ε^2, i.e., the array-to-array variability is larger than the measurement error, ρ will be close to 1; while $\rho \to 0$ when $\sigma_b^2 << \sigma_\varepsilon^2$. For the latter, the probes intensities are independent and there is no coherence within the probe set.

A user-defined threshold is required to call a gene informative or non-informative using the intra-cluster correlation. Similar to the FLUSH and the I/NI Calls, such a threshold may not be readily available. An objective measure for calling a gene informative or non-informative can be based on a likelihood ratio test. Suppose there are two competing models.

$$\begin{aligned} M_0 &: log_2(PM_{ij}) = \mu_j + \varepsilon_{ij}, \\ M_1 &: log_2(PM_{ij}) = \mu_j + b_i + \varepsilon_{ij}. \end{aligned} \tag{8}$$

The first model (M_0) assumes that probes within in a probe set are independent; while the later (M_1) assumes that the probes within a probe set are correlated. Note that M_0 is nested within M_1, i.e., for the case that $\sigma_b^2 = 0$, M_1 reduces to M_0. Hence, one can use the likelihood ratio test to test the corresponding hypotheses

$$\begin{aligned} H_0 &: \sigma_b^2 = 0, \\ H_1 &: \sigma_b^2 > 0. \end{aligned} \tag{9}$$

Note that testing the hypotheses in (9) is equivalent to testing the null hypothesis $H_0 : \rho = 0$ versus the alterative $H_1 : \rho > 0$. As argued by Talloen *et al.* (2007) and Calza *et al.* (2007), probe sets with low array-to-array variability are not likely to carry an important biological signal and should be excluded from further analysis. Hence, the likelihood ratio test for the hypotheses formulated in (9) can be used to filter the probe sets. A probe

DOI: 10.2202/1544-6115.1460

set is declared as an informative probe set whenever the null hypothesis in (9) is rejected. Alternatively, we can use information criteria, such as the Akaike information criterion, (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978) to select one of the models (either M_0 or M_1) that has the best fit to the data. The likelihood ratio test, the AIC, and the BIC do not rely on a somewhat ad hoc selection of a threshold. To gain a better understanding of the informative/non-informative calls using the AIC and the BIC, we propose to calculate the posterior probability for a probe set to be called informative $P(M_1|\text{Data})$ using the information criteria. Following Burnham and Anderson (2002), the posterior probability based on the AIC for the models in (8) is defined as

$$P(M_1|\text{Data})_{AIC} = \frac{exp(-\frac{1}{2}\Delta AIC_{M_1})P(M_1)}{\sum_{r=0}^{1} exp(-\frac{1}{2}\Delta AIC_{M_r})P(M_r)}. \tag{10}$$

Let AIC_{M_1} and AIC_{M0} be the AIC values from models M_1 and M_0, respectively. Suppose we defined the minimum AIC from the two models as $AIC_{\text{min}} = \min(AIC_{M_1}, AIC_{M_0})$. The idea is to calculate the probability for a probe set to be called informative given the observed data. This probability depends on the relative loss of information when using model M_1 instead of the more plausible model. The lower the loss, the higher the probability. The relative loss of information for using model M_1 instead of the most plausible model out of models M_1 and M_0 is defined as $\Delta AIC_{M_1} = AIC_{M_1} - AIC_{\text{min}}$. Suppose the minimum of the AIC values from AIC_{M_1} and AIC_{M_0} is AIC_{M_1}, i.e, $AIC_{\text{min}} = AIC_{M_1}$. Then, $\Delta AIC_{M_1} = 0$ and there is no loss of information for using model M_1. Note that $P(M_1)$ is the prior probability for a probe set to be called informative. We assume a priori that a probe set is equally likely to be called informative or non-informative, which implies that $P(M_1) = P(M_0) = 0.5$. It is expected that an informative probe set will have a high posterior probability and otherwise for a non-informative probe set. The posterior probability of the models can similarly be obtained by using the BIC as well.

3.4 A Confirmatory Factor Analysis Approach

The I/NI Calls proposed by Talloen *et al.* (2007) is based on a factor analysis model using the Bayesian approach. We show that gene filtering can be done by using a confirmatory factor analysis model. Let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denote the covariance matrix from the confirmatory factor analysis model. Here, $\boldsymbol{\theta} = (\lambda_1, \cdots, \lambda_k, \sigma_z^2, \sigma_\varepsilon^2)$ is a vector of unknown variance-covariance compo-

nents. The first model considered is similar to the factor analysis model used in the I/NI Calls. We use the same notation as in Section 3.2 and assume that $\boldsymbol{z} \sim N(0, \sigma_z^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2)$. The implied covariance matrix for this model is a symmetric matrix given by

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{\theta}) &= [\sigma_z^2 \boldsymbol{\lambda}'\boldsymbol{\lambda} + \sigma_\varepsilon^2 I] \\ &= \left\{ \begin{array}{llll} \sigma_z^2\lambda_1\lambda_1 + \sigma_\varepsilon^2 & \sigma_z^2\lambda_1\lambda_2 & \cdots & \sigma_z^2\lambda_1\lambda_k \\ \sigma_z^2\lambda_2\lambda_1 & \sigma_z^2\lambda_2\lambda_2 + \sigma_\varepsilon^2 & \cdots & \sigma_z^2\lambda_2\lambda_k \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_z^2\lambda_k\lambda_1 & \sigma_z^2\lambda_k\lambda_2 & \cdots & \sigma_z^2\lambda_k\lambda_k + \sigma_\varepsilon^2 \end{array} \right\}. \end{aligned}$$

Next, we consider a factor analysis model for which $\sigma_z^2 = 1$. It is easy to see that under this assumption,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})_{qp} = \left[\boldsymbol{\lambda}'\boldsymbol{\lambda} + \sigma_\varepsilon^2 I\right]_{qp} = \left\{ \begin{array}{ll} \lambda_i^2 + \sigma_\varepsilon^2 & q = p, \\ \lambda_i\lambda_j & q \neq p. \end{array} \right.$$

We term this model "FA-Free" since $\lambda_q \neq \lambda_p$. Note that this factor analysis model is equivalent to the model used in the I/NI Calls, although, it is fitted in a frequentist way.

The third model we consider is termed "FA-Restricted" since the factor loading are constrained to be equal for all probes, that is, $\lambda_q = \lambda_p = \lambda$. Under this assumption, it follows that

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})_{qp} = \left[\lambda^2 J + \sigma_\varepsilon^2 I\right]_{qp} = \left\{ \begin{array}{ll} \lambda^2 + \sigma_\varepsilon^2 & q = p, \\ \lambda^2 & q \neq p. \end{array} \right.$$

Note that this model assumes a constant variance for all probes within a probe set, and the array-to-array variability is captured by the factor loadings, since $\sigma_z^2 = 1$. An alternative model is a model that relaxes the assumption that $\sigma_z^2 = 1$ and assumes $z \sim N(0, \sigma_z^2)$. This model treats the factor loadings as an offset variable, that is, $\lambda_q = \lambda_p = 1$. Under this assumption, the array-to-array variability is captured by the variance of the latent factor and the model based covariance matrix is given by

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})_{qp} = \left[\sigma_z^2 J + \sigma_\varepsilon^2 I\right]_{qp} = \left\{ \begin{array}{ll} \sigma_z^2 + \sigma_\varepsilon^2 & q = p, \\ \sigma_z^2 & q \neq p. \end{array} \right.$$

It is easy to see that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ from the "FA-Restricted" model is identical to the covariance matrix of the linear mixed model in (5) with $\sigma_b^2 = \lambda^2 = \sigma_z^2$. Hence, all models discussed in Sections 3.2, 3.3, and 3.4 belong to the same

DOI: 10.2202/1544-6115.1460

class of models, which use a latent variable (the factor or the random effect) to capture the array-to-array variability.

3.5 Latent Variable Models Versus Fixed Effects Models

Interestingly, as pointed out by Verbeke and Molenberghs (2000), the mixed model in (5) implies a marginal model for $\boldsymbol{PM_i}$, in which

$$log_2(\boldsymbol{PM}_i) \sim N(\boldsymbol{X}_i\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\Theta})) \quad \text{where} \quad \boldsymbol{\Sigma}(\boldsymbol{\theta})_{pq} = \begin{cases} \tau^2 + \sigma_\varepsilon^2 & q = p, \\ \tau^2 & q \neq p, \end{cases}$$

where τ^2 is the array-specific effect that captures the array-to-array variability and $\boldsymbol{X}_i\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)$ is a vector of probe-specific fixed effects. The decomposition of the total variability of the mixed effect model and the marginal model is identical (i.e, $\tau^2 = \sigma_b^2$). Note that the marginal model is different from the model of the FLUSH, which treated the arrays as fixed effects in the linear model. The fixed effects model in the FLUSH assumes that probes in a probe set are independent. This is implied by the structure of the model-based covariance matrix. The covariance matrix for the FLUSH is a diagonal matrix, while that of the marginal model has a compound symmetry structure with off-diagonal elements corresponding to the array-to-array variability. Consequently, the inference based on the marginal model for the array-to-array variability can be performed by comparing between the compound symmetry covariance matrix and a diagonal covariance matrix. It should be noted that for the FLUSH, the χ^2 statistic does not in any way account for the correlation between probes in a probe set. However, the formal inference for the FLUSH approach can be based on general linear hypothesis testing. The main goal is therefore to test the importance of array-to-array variability in the model. Suppose one starts with the full model as specified in (1). The idea is to investigate how much information will be lost by reducing the model to:

$$log_2(PM_{ij} - IMM_{ij}) = \mu_j + \varepsilon_{ij}. \tag{11}$$

The null and alternative hypotheses can be stated as:

$$H_0 : \boldsymbol{L\alpha} = \boldsymbol{0}, \qquad \text{versus} \qquad H_1 : \boldsymbol{L\alpha} \neq \boldsymbol{0}, \tag{12}$$

where $\boldsymbol{L}$ is a known matrix and $\boldsymbol{\alpha}$ is a vector of array-specific effects. The test statistic $\boldsymbol{\hat{\alpha}'L'V^{-1}L\hat{\alpha}}$ asymptotically follows a χ^2 distribution with rank($\boldsymbol{L}$) as the degrees of freedom. In order to account for extra variability due

to the estimation of the covariance parameters, the test is often approximated with an F-test with rank($\boldsymbol{L}$) as the numerator degrees of freedom and the denominator degrees of freedom is estimated from the data. The estimation can be based on Satterthwaite-type approximation (Verbeke and Molenberghs, 2000). Note that this test statistic is usually reported as a "Type III test" in the standard software for mixed models, such as SAS procedure mixed and lme() in R/S$^+$ package *nlme*.

3.6 Filtering Scores

The latent variable models decompose the total variability into the array-to array variability and measurement error. This decomposition allows us, similar to the I/NI Calls, to calculate a probe set-specific filtering score based on the ratio between the array-to-array variability and measurement error, denoted as R^2. In our setting, once the model-based covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is estimated, the filtering scores can be calculated using the parameter estimates from the covariance matrix. The filtering scores for the respective models are presented in Table 2. Note that unlike the conditional variance, the R^2 or intra-cluster correlation (ρ) does not depend on the probe set size (k).

Table 2: Criteria for gene filtering based on different models. The filtering scores are: conditional variance ($v(z|x)$), intra-cluster correlation (ρ), R^2, and χ^2 statistic ($\hat{\alpha}'\hat{\boldsymbol{V}}^{-1}\hat{\alpha}$).

Type	Model	Assumptions	Filtering Scores	
			$v(z\|x)$	R^2/ρ
latent	I/NI Calls/FA-Free	$\lambda_i \neq \lambda_j$ $z \sim N(0,1)$	$\sigma_\varepsilon^2 \Big/ \left(\sum_{j=1}^{k} \lambda_j^2 + \sigma_\varepsilon^2\right)$	$\sum_{j=1}^{k} \lambda_j^2 \Big/ \left(\sum_{j=1}^{k} \lambda_j^2 + k\sigma_\varepsilon^2\right)$
	FA-Restricted	$\lambda_i = \lambda_j = \lambda$ $z \sim N(0,1)$	$\sigma_\varepsilon^2 / \left(k\lambda^2 + \sigma_\varepsilon^2\right)$	$\lambda^2 / \left(\lambda^2 + \sigma_\varepsilon^2\right)$
		$\lambda_i = \lambda_j = 1$ $z \sim N(0, \sigma_z^2)$	$\sigma_\varepsilon^2 / \left(k\sigma_z^2 + \sigma_\varepsilon^2\right)$	$\sigma_z^2 / \left(\sigma_z^2 + \sigma_\varepsilon^2\right)$
	LMM	$b_i \sim N(0, \sigma_b^2)$	$\sigma_\varepsilon^2 / \left(k\sigma_b^2 + \sigma_\varepsilon^2\right)$	$\sigma_b^2 / \left(\sigma_b^2 + \sigma_\varepsilon^2\right)$
Type	Model	Assumptions	χ^2	ρ
fixed	Marginal Model	$\varepsilon \sim N(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$		$\sigma_b^2 / \left(\sigma_b^2 + \sigma_\varepsilon^2\right)$
	FLUSH	$\varepsilon \sim N(0, \sigma_\varepsilon^2)$	$\hat{\alpha}'\hat{\boldsymbol{V}}^{-1}\hat{\alpha}$	

DOI: 10.2202/1544-6115.1460

4 Application to HGU-133A

The results presented in this paper are based on the $log2$ transformation of the probes intensity values after quantile normalization. The analysis is carried out using only the perfect match (PM) with a Gaussian distribution. We refer to Hochreiter *et al.* (2006) for motivation for using only the PM and the choice of Gaussian distribution for the transformed intensity values. In Table 3, we present an overview of results for the methods discussed in previous sections. It can be observed from these results that the proposed linear mixed model outperforms the FLUSH and performs equally well as the I/NI Calls. Moreover, the importance of the Bayesian implementation of the factor analysis model in the I/NI Calls is evidenced when compared with the "FA-Free" model. The "FA-Free" model calls all probe sets informative, thereby resulting in 100% false positive rates; while the I/NI Calls results in 0.2% false positives.

Table 3: The summary of the performance of the different methods. The cut-off for $v(z|x)$ and ρ is 0.5

Criteria	Method	False Negatives	False Positives
quantile reg	FLUSH	0.3400	0.4000
	LMM	0.0800	0.4000
$v(z\|x)$	I/NI	0.0000	0.0020
	FA-Free	0.0000	1.0000
	FA-Restricted	0.0000	0.2400
ρ	FA-Free	0.0000	0.0600
	FA-Restricted	0.0000	0.0004
	LMM	0.0000	0.0004

4.1 The FLUSH and the Linear Mixed Model

The paradigm of an informative or non-informative call of a gene is hinged on the relationship between array-to-array variability and variability due to measurement error. The relationship between array-to-array variability and measurement error based on the FLUSH and the linear mixed model is pre-

sented in Figures 1a and 1b, respectively. Both methods show that array-to-array variability for the spiked-in probe sets is higher than that of the background mixtures. Also, a higher residual variance is observed for the spiked-in probe sets. This may be due to the variance-intensity relationship (Hochreiter *et al.*, 2006). The residual variance (Figure 1c) from the linear mixed model is higher than that of the FLUSH. This implies that FLUSH underestimates the variance of the measurement error for both the spiked-in probe sets and the background mixtures. The informative probe sets are expected to have higher array-to-array variability than the predicted values from the 60% quantile regression. The plot of the proportion of false negatives (Figure 1d) shows that both methods result in false negatives. However, the linear mixed model leads to less false negatives as compared to the FLUSH.

4.2 Confirmatory Factor Analysis Models, the I/NI Calls, and Linear Mixed Model

We compare results based on the latent variable models. Note that the "FA-Free" and "FA-Restricted" models are factor analysis models using the frequentist approach.

4.2.1 Conditional Variance

The I/NI Calls defines a probe set as informative based on the proportion of the total variability associated with the measurement error. This proportion is referred to as the conditional variance of the latent factor given the observed data. It is bounded between 0 and 1. The informative genes are those with small values of the conditional variance and large value for the non-informative genes. In Figure 2, we present the conditional variance for factor analysis models and show their relationship with intra-cluster correlation from the linear mixed model. The histogram of the conditional variance from the I/NI Calls (Figure 2b) is bimodal. For the conditional variances based on the "FA-Free" (Figure 2a) and "FA-Restricted" (Figure 2c), the choice of threshold value may be data dependent. The bimodal distribution for the conditional variance from the I/NI Calls is due to the shrinkage of the factor loadings towards zero for the non-informative probe sets.

Since the intra-cluster correlation obtained from the linear mixed model is based on the relationship between the array-to-array variability and the variability due to the measurement error, we expect it to be associated with the conditional variance from the factor analysis models. These relationships are

DOI: 10.2202/1544-6115.1460

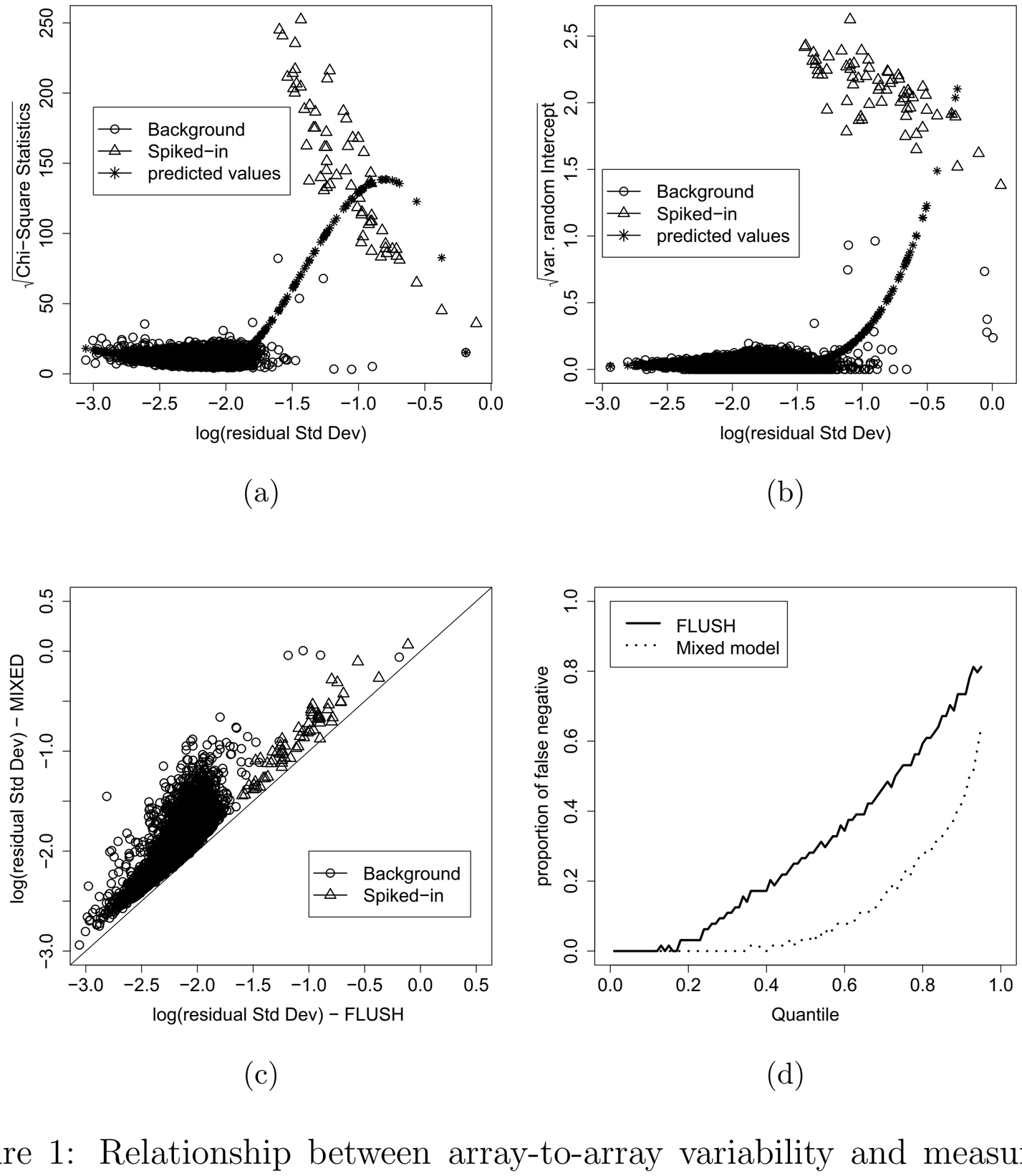


Figure 1: Relationship between array-to-array variability and measurement errors based on the FLUSH and the linear mixed model: (a) array-to-array variability versus measurement errors obtained from the FLUSH, (b) array-to-array variability versus measurement errors obtained from the linear mixed model , (c) estimated measurement errors from the FLUSH versus that of the linear mixed model, and (d) proportion of false negatives for varying values of quantile regression

shown in Figures 2d-2f. There is no obvious relationship between the conditional variance from the I/NI Calls and the intra-cluster correlation due to the zero informative prior placed on the factor loadings using Bayesian approach.

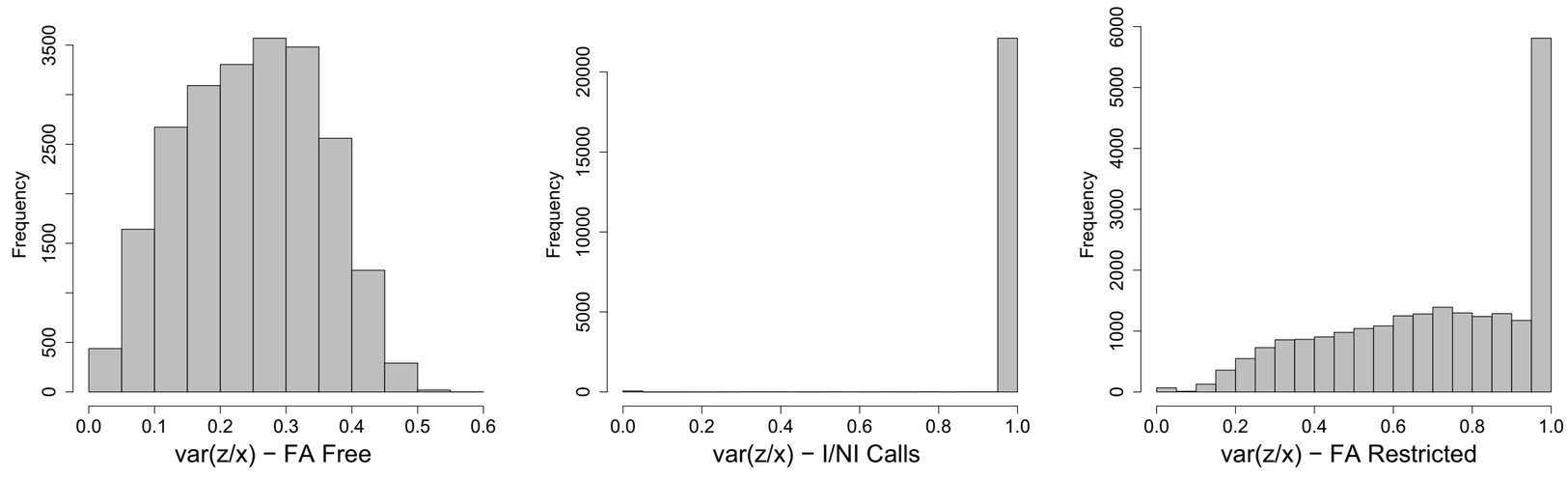


Figure 2: Distributions of the conditional variance based on the factor analysis models: histogram plot of conditional variance based on (a) the "FA-Free" model, (b) the I/NI Calls, (c) the "FA-Restricted model"; Relationship between (d) conditional variance from "FA-Free" model and intra-cluster correlation from the mixed model, (e) conditional variance from the I/NI Calls and intra-cluster correlation from the mixed model, and (f) conditional variance from the "FA-Restricted" model and intra-cluster correlation from the mixed model.

The effect of such a prior is that the conditional variance for informative and non-informative probe sets shrinks toward zero and one, respectively. A reciprocal relationship is observed between the conditional variance from the "FA-Free" model and the intra-cluster correlation. A one-to-one relationship can be observed between them. However, there are situations, where a value of intra-cluster correlation corresponds to multiple values of conditional variance. This is due to the dependence of conditional variance on the number of probes in a probe set. From all the models, it can be observed that the spiked-in probe sets have higher intra-cluster correlations and smaller conditional variances.

DOI: 10.2202/1544-6115.1460

4.2.2 Proportion of Variability Explained by the Latent Factor

As an alternative criterion to the conditional variance for the factor analysis models, we propose to use the proportion of variability in the observed data explained by the latent factor (denoted by R^2). This is equivalent to the intra-cluster correlation from the linear mixed model. Similar to the conditional variance, the distributions of R^2 for both the "FA-Free" and "FA-Restricted" models are not the same (3a and 3b). The relationship between R^2 and the intra-cluster correlation are presented in Figures 3c and 3d. The scatter plots of R^2 from the "FA-Restricted" model versus the intra-cluster correlation show, as expected, a perfect linear relationship between them. As we mentioned in Section 3.4, the two models are equivalent. This is not the case for the plot of R^2 from the "FA-Free" model and the intra-cluster correlation (Figure 3d). However, the informative probe sets are those with higher R^2 and intra-cluster correlations.

4.2.3 Proportion of False Negatives and False Positives Based on Conditional Variance and Intra-cluster Correlation

The plots of the false positive and false negative rates for the latent factor models using conditional variance are presented in Figures 4a and 4b. The percentage of false positive (4b) is more pronounced for the "FA-Free" model and less pronounced for the I/NI Calls (4a). All the models have a similar false positive rate when the threshold is around 0.05. Especially for the I/NI Calls, it appears that it does not matter which threshold is used. The proportion of false negative is zero for a threshold as small as 0.1. The percentages of false positives and false negatives based on R^2 or intra-cluster correlation are presented in Figures 4c and 4d. The percentages of false positive (4d) are the same for the "FA-Restricted" and the linear mixed model. Figure 4c shows zero false negatives for all the methods until a threshold around 0.6.

4.2.4 Informative or Non-informative Probe Sets Based on the I/NI Calls and Linear Mixed Model

In the previous sections, we establish the relationships between the factor analysis models and the linear mixed model. In this section, we zoom in on genes called informative by either the I/NI Calls or linear mixed model using the conditional variance or intra-cluster correlation with a threshold of 0.5. The probe level data of an informative probe set and non-informative probe set identified by both the I/NI Calls and the linear mixed model are presented in Figures 5a and 5b, respectively. There is a strong coherence between all

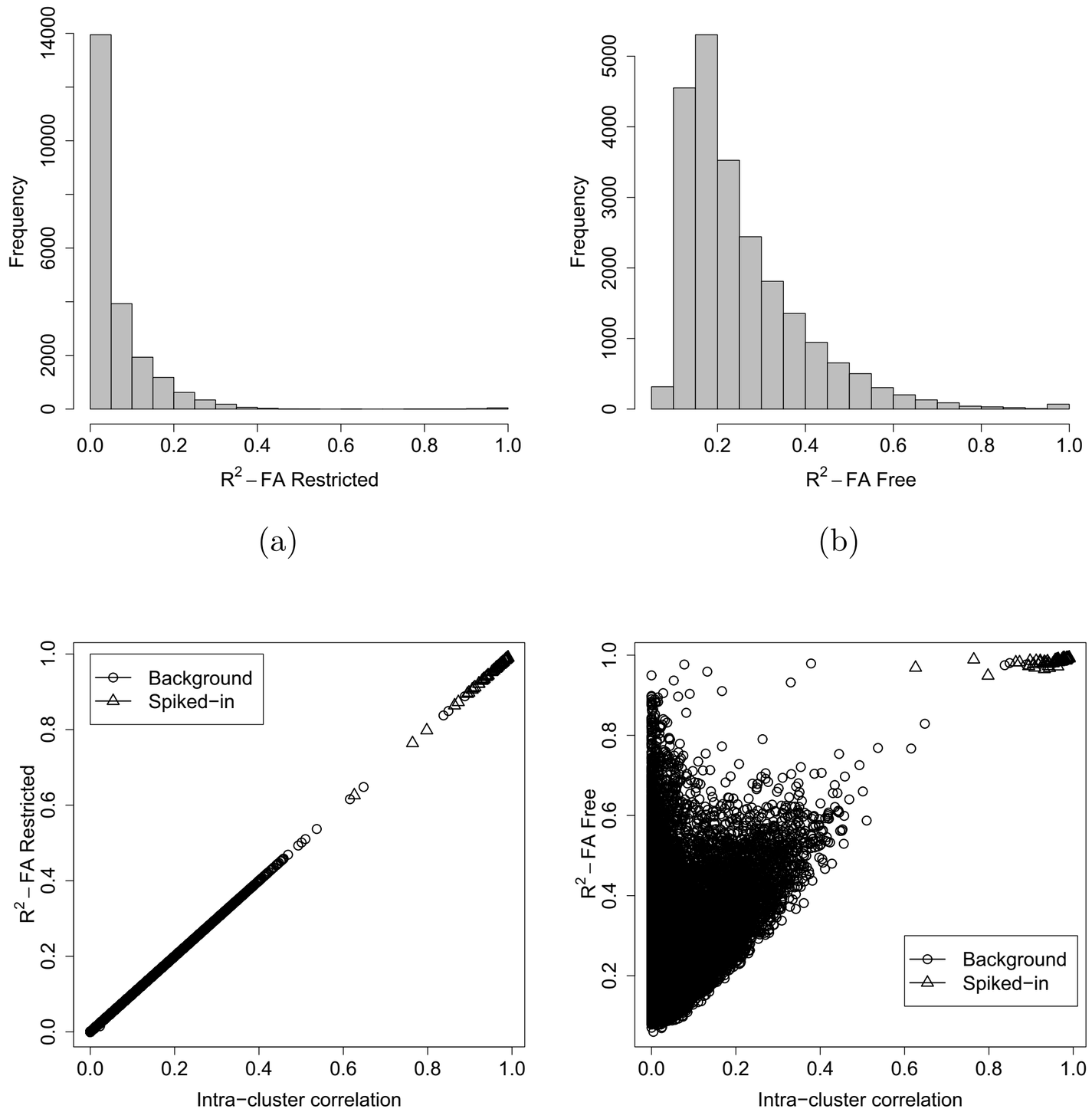


Figure 3: Relationship between R^2 from the factor analysis models and the intra-cluster correlation from the linear mixed model: (a) distribution of R^2 from the "FA-Free" model, (b) distribution of R^2 from the "FA-Restricted" model, (c) intra-cluster correlation versus R^2 from the "FA-Free" model, (d) intra-cluster correlation versus R^2 from the "FA-Restricted" model

the probes of the informative probe set. For the non-informative probe set, there is little or no coherence between the intensity measures of the probes. To illustrate this further, we present in Table 4 the factor loadings and R^2 from the "FA-Free" and "FA-Restricted" models. For the informative probe set, the latent factor explains about 99% of the variability of all the probes; while for the non-informative probe set, the proportion of variability in any

DOI: 10.2202/1544-6115.1460

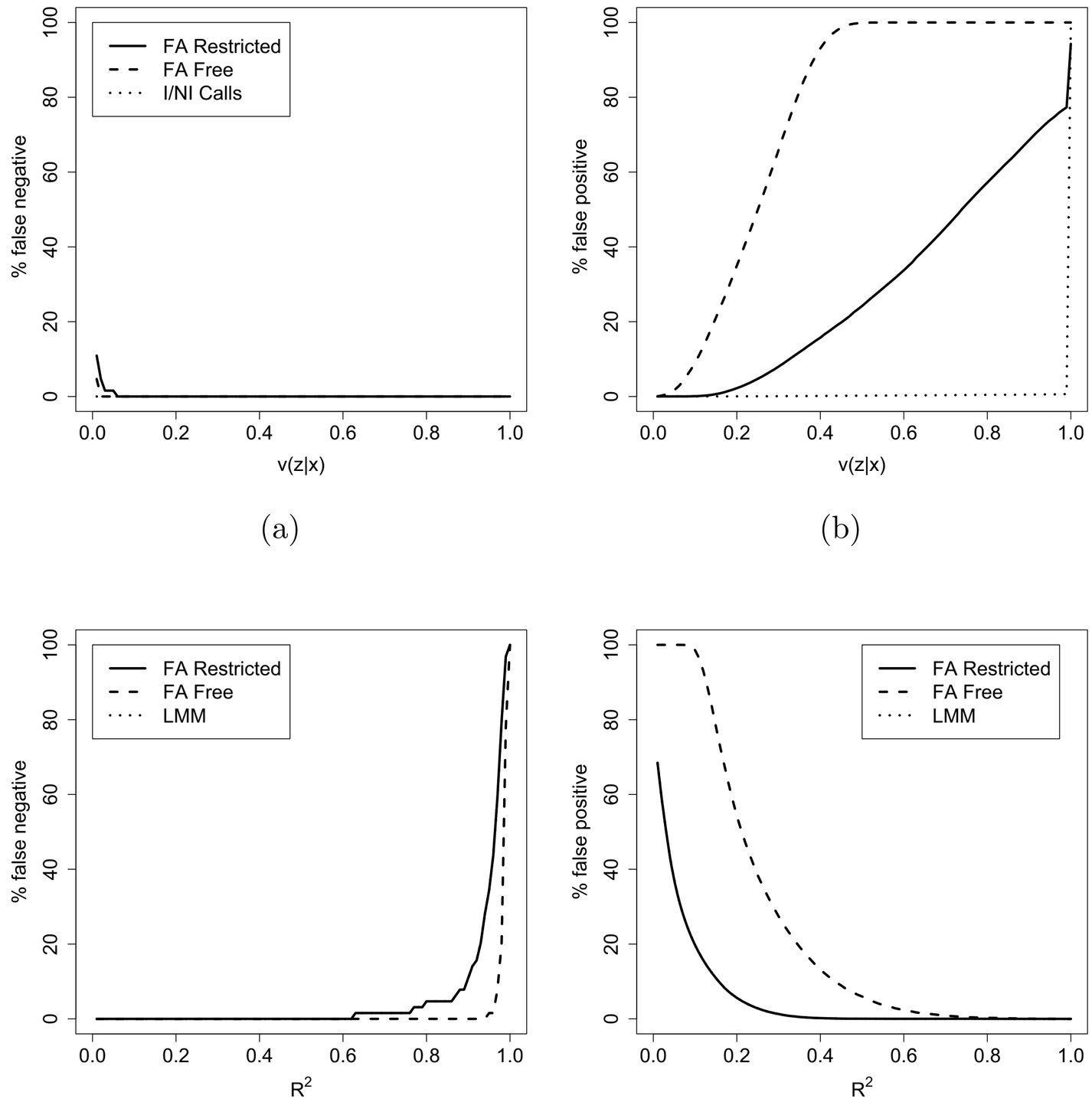


Figure 4: Proportion of false positives and false negatives based on the confirmatory factor analysis models and the linear mixed model: (a) false negatives using conditional variance , (b) false positives using conditional variance, (c) false negatives using R^2 or intra-cluster correlation, (d) false positives using R^2 or intra-cluster correlation

of the probe explained by the latent factor is less than 50%. Based on the "FA-Restricted" model, the proportion of variability explained by the latent factor for the informative probe set is 99%, and 0% for the non-informative probe.

For the threshold of 0.5, there are 112 probe sets called informative by the I/NI Calls and 78 probe sets called informative by the linear mixed model.

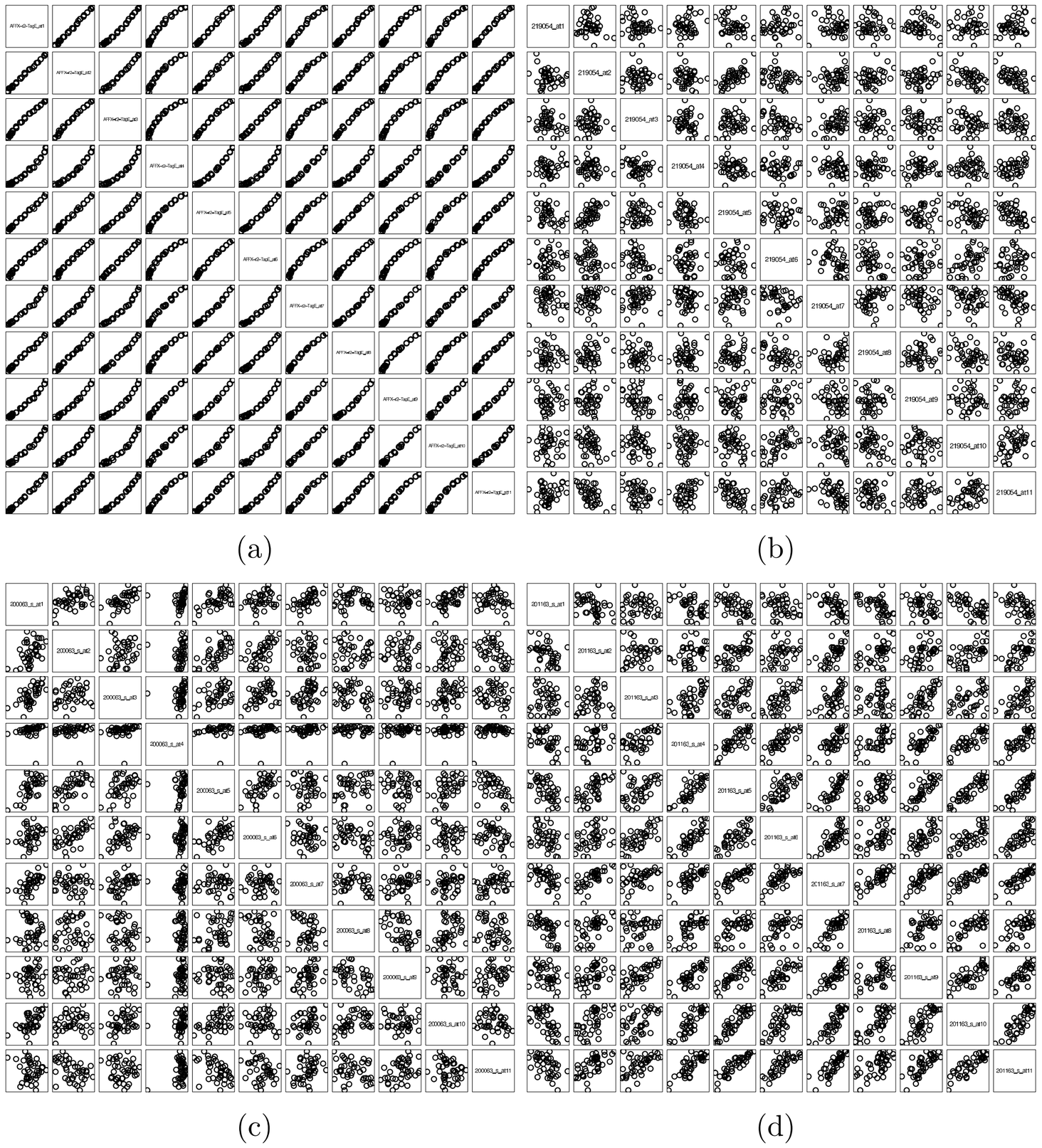


Figure 5: Probe level intensities for probe sets called informative or non-informative by both the I/NI Calls and the linear mixed model: (a) called informative by both the I/NI Calls and the linear mixed model, (b) called non-informative by both the I/NI Calls and the linear mixed model, (c) called informative only by the I/NI Calls, and (d) called informative only by the mixed model.

In both cases, the informative probe sets include all the spiked-in probe sets. Examples of probe level data for an informative probe set by either the I/NI Calls or the linear mixed model are presented in Figures 5c and 5d, respectively. For the majority of the probes in a probe set called informative by the I/NI Calls, there is little or no coherence between the arrays; while those called

DOI: 10.2202/1544-6115.1460

Table 4: Factor loadings (λ) and R^2 from "FA - Free" and "FA-Restricted" models for probe sets called informative (I) or non-informative (NI) by both the I/NI Calls and the linear mixed model.

		λ		R^2	
Model	Probes	I	NI	I	NI
"FA-Free"	1	2.458	0.022	0.992	0.063
	2	2.419	-0. 074	0.992	0.436
	3	2.536	-0.064	0.993	0.367
	4	2.333	0.025	0.992	0.084
	5	2.492	-0.061	0.992	0.343
	6	2.504	0.031	0.992	0.117
	7	2.455	-0.068	0.992	0.394
	8	2.417	-0.028	0.991	0.097
	9	2.407	0.010	0.991	0.013
	10	2.381	0.054	0.992	0.294
	11	2.347	0.073	0.991	0.429
"FA-Restricted"		2.416	0	0.991	0

informative by the mixed model appear to depend on the average coherence within a probe set. To illustrate this further, we present in Table 5 the factor loadings (λ) and R^2 for these probe sets, based on the "FA-Free" and "FA-Restricted" model. For the probe sets called informative by the I/NI Calls, the proportion of the variability explained by the latent factor is dominated by one of the probes within the probe set. For example, the latent factor explains about 96% of the variability of probe 4 , but less than 2% of the variability from other probes. For the probe set called informative only by the linear mixed model, the latent factors explain more than 50% of the variabilities in the majority of the probes. By looking at the "FA-Restricted" model, it can be observed that the proportion explained by the latent factor for the probe set called informative by only the I/NI Calls is less than 1% and more than 50% for the probe sets called informative only by the linear mixed model.

4.3 Likelihood Ratio Test and Information Criteria

The plots of the proportion of false positives and false negatives in the previous sections indicate that the choice of threshold for the quantile regression, conditional variance, R^2, and intra-cluster correlation may depend on the data at hand, and are therefore highly subjective. In this section, we consider other

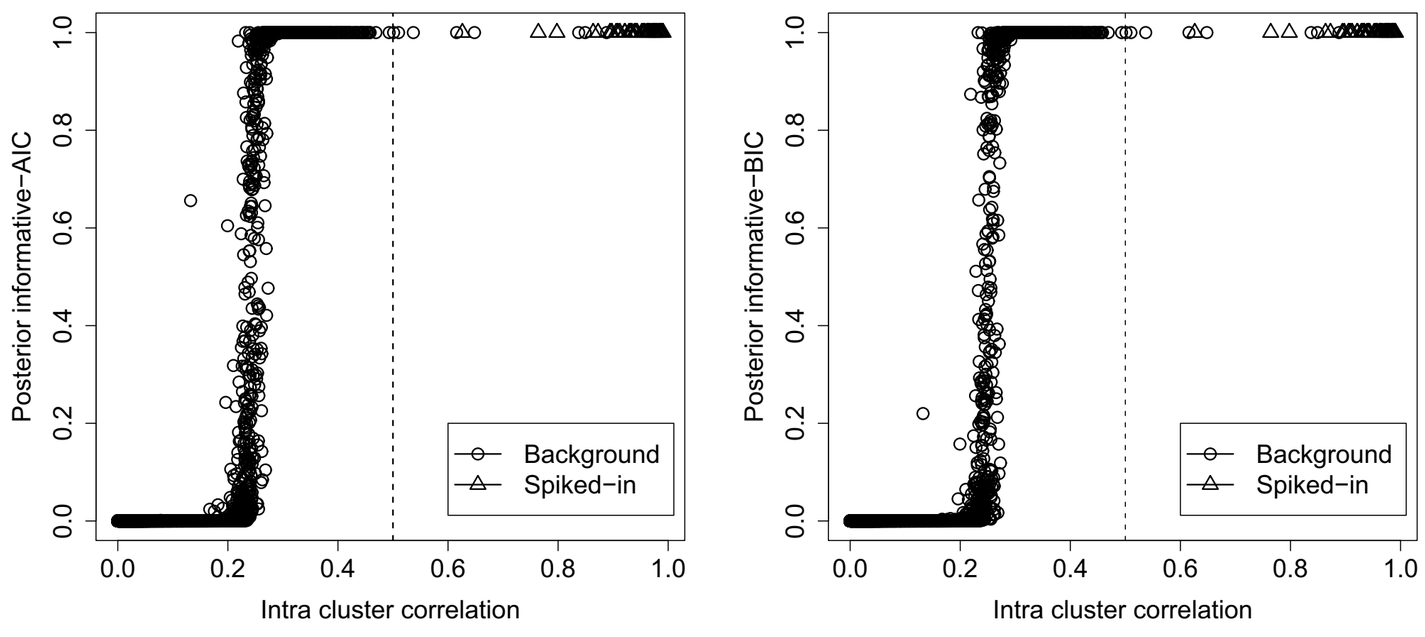
Table 5: Factor loadings(λ) and R^2 from the "FA - Free" and "FA-Restricted" models for probe sets called informative by either I/NI Calls or the Linear mixed model.

Model	Probes	λ		R^2	
		I/NI Calls	Mixed Model	I/NI Calls	Mixed Model
"FA-Free"	1	0.047	0.033	0.108	0.079
	2	0.015	-0.048	0.012	0.148
	3	0.022	-0.084	0.019	0.349
	4	0.068	-0.180	0.963	0.713
	5	0.040	-0.284	0.081	0.860
	6	0.015	-0.268	0.013	0.845
	7	0.007	-0.251	0.003	0.827
	8	0.011	-0.115	0.006	0.502
	9	0.017	-0.279	0.016	0.856
	10	0.025	-0.317	0.033	0.884
	11	0.022	-0.169	0.027	0.686
"FA-Restricted"		0.072	-0.174	0.084	0.537

criteria, such as using a likelihood ratio test or the AIC, and BIC to call genes informative or non-informative. To use any of these criteria, two models (M_0 and M_1 in (8)) are required. The likelihood ratio test is the difference between the -2loglikelihood obtained from the models M_0 and M_1. The likelihood ratio test for testing whether the variance of the random intercept is zero requires the correction for a boundary problem (Verbeke and Molenberghs, 2000), since the null hypothesis is tested on the boundary of the parameter space. As a result, a mixture of $\chi^2_{0,1}$ is used to obtain the p-value. It is observed that the majority of the probe sets have p-values close to 1. At the 5% level of significance, the number of genes called informative is 549, with the Benjamini and Horchberg procedure (Benjamin and Hochberg, 1995) for multiple testing adjustment. A gene may also be called informative if the model with the random array effect has the minimum AIC or BIC among the models in (8). Both criteria account for the goodness of fit and the complexity of each model. Out of the 22,300 probe sets, 742 and 663 are called informative by using the AIC and BIC, respectively. Probe sets called informative by either the likelihood ratio test or the AIC or the BIC include all the spiked-in probe sets, which means zero false negative. It is observed that spiked-in probe sets have small p-values, small conditional variances and high intra-cluster correlation. Additionally, Figure 6 shows the posterior probability for a probe set to be

DOI: 10.2202/1544-6115.1460

called informative, $P(M_1|\text{Data})$, by using the AIC and BIC. The spiked-in probe sets have posterior probability of 1. When the posterior probability is compared with the intra-cluster correlation, it is noted that using intra-cluster correlation of 0.5 as a threshold may be too stringent in real life data where the intra-cluster correlation of truly informative genes are unknown. There are genes with intra-cluster correlation between 0.2 and 0.4, which have a posterior probability of 1. This means for these probe sets, there is strong a evidence from the data to call them informative. Note that in the case of the HGU-133A dataset, considering these genes as informative will increase the number of false positive.



(a) (b)

Figure 6: Plot of the posterior probability of a probe set called informative by using the AIC and BIC versus the intra-cluster correlation; (a) $P(M_1|\text{Data})_{AIC}$ and (b) $P(M_1|\text{Data})_{BIC}$

5 Simulation Study

Two simulation studies are conducted in order to investigate the performance of the gene filtering methods, namely, the FLUSH, the I/NI Calls, and the gene filtering based on the linear mixed model (LMM). The simulation studies investigates: (1) the performance of the gene filtering methods, (2) the assumption that all probes in a probe set quantify the expression levels of the same target.

5.1 Investigation of Performance of Gene Filtering Methods

This simulation study asseses the performance of the FLUSH, the I/NI Calls, and the gene filtering based on the linear mixed model under the assumption that all probes in a probe set quantify the expression levels of the same target. 100 datasets were generated. Each dataset contains 100 informative and 1900 non-informative probe sets. The probe level data are generated from a multivariate normal distribution i.e., $\boldsymbol{Y}_l \sim N(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}(\Theta_l)$, where $\boldsymbol{Y}_l$, $\boldsymbol{\mu}_l$, and $\boldsymbol{\Sigma}(\Theta_l)$ are the probe level data, the probe-specific effects, and the covariance matrix for probe set l, respectively. Note that the covariance matrix depends on the parameter vector $\boldsymbol{\Theta}_l = \{\sigma^2_{b_l}, \sigma^2_{\varepsilon_l}\}$. To obtain values for $\boldsymbol{\mu}_l$ and $\sigma^2_{\varepsilon_l}$, 100 informative probe sets and 1900 non-informative probe sets are randomly sampled from the spiked-in dataset. The variance component associated with array-to-array variability is calculated by

$$\sigma^2_{b_l} = \begin{cases} \frac{\rho_I \sigma^2_{\varepsilon_l}}{1-\rho_I}, & \text{an informative probe set,} \\ \\ \frac{\rho_{NI} \sigma^2_{\varepsilon_l}}{1-\rho_{NI}}, & \text{a non-informative probe set.} \end{cases}$$

Four different types of datasets were generated; (1) a noisy dataset with $\rho_I = 0.3$ and $\rho_{NI} = 0.1$, (2) a weak signal dataset with $\rho_I = 0.55$ and $\rho_{NI} = 0.1$, (3) a strong signal dataset with $\rho_I = 0.9$ and $\rho_{NI} = 0.1$ and (4) a dataset for which $\rho_I = U(0.5, 1)$ and $\rho_{NI} = U(0, 0.5)$. Note that the last type represent a more realistic scenario, in which the correlation among the probes is not constant for all informative and non-informative probe set.

5.1.1 Effects of Sample Size n and Probe Set Size k

For each simulation setting, the performance of the gene filtering methods under increasing sample size/the number of arrays n = (3, 6, 9, 18, 21, 27, 30, 42, 50, 100) or the probe set size k= (8, 10, 12, 14, 16, 20, 30, 40, 50, 69) are investigated. In order to investigate the effect of sample size, 100 datasets are generated for each combination of the different types of dataset and the sample size; the probe set size is fixed at 20. Similarly, to investigate the effect of probe set size, 100 datasets are generated for each combination of different types of dataset and probe set size; while the sample size is fixed at 42 (the number of samples in the spiked-in dataset). We refer to Section 2 of the supplementary document for an elaborate discussion of the simulation setting and results. The simulation results for the setting $\rho_I = U(0.5, 1)$ and $\rho_{NI} = (0, 0.5)$ are

DOI: 10.2202/1544-6115.1460

presented in Table 6. For $n = 3$, the FLUSH method results in 47% and 39% false negatives and false positives, respectively. The I/NI Calls results in 7% false negatives and 60% false positives. The linear mixed model results in 33%, 25% and 19% false negatives and 12%, 2% and 2% false positives based on ρ, AIC and BIC respectively. For $n = 100$, the FLUSH method results in 46% and 39% false negatives and false positives, respectively. The I/NI Calls results in 10% false negatives and 1% false positives. The linear mixed model results in 4%, 0% and 0% false negatives and 3%, 75% and 74% false positives based on ρ, AIC and BIC, respectively. For the effect of probe set size, when $k = 8$, the FLUSH method results in 47% and 39% false negatives and false positives, respectively. The I/NI Calls results in 20% false negatives and 1% false positives. The linear mixed model results in 8%, 0% 0 19% false negatives and 5%, 46% and 44% false positives based on ρ, AIC and BIC respectively. When $k = 100$, the FLUSH method results in 48% and 39% false negatives and false positives, respectively. The I/NI Calls results in 4% false negatives and 4% false positives. The linear mixed model results in 4%, 0% and 0% false negatives and 4%, 67% and 66% false positives based on ρ, AIC and BIC, respectively.

Figures 7a and 7b show the effects of sample size on the proportion of false negatives and false positives. The I/NI Calls and the linear mixed model yield fewer false negatives and false positives than the FLUSH method. However, the linear mixed model using the AIC or the BIC as a filtering score gives a higher proportion of false positives than the FLUSH method. Note that the proportion of false negatives decreases with an increase in sample size. The results of false negatives can be explained by the overlap between the informative and non-informative probe sets when the probe level data are generated with intra-cluster correlation in the neighborhood of 0.5.

Figures 7c and 7d show the effects of the probe set size on the proportion of false negatives and false positives. The I/NI Calls and the linear mixed model give better results than the FLUSH method. The proportion of false negatives from all the methods decreases with an increase in probe set size. Note that, though the proportion of false positives from the I/NI Calls are consistently lower than that of the linear mixed model with intra-cluster correlation as the filtering score, they increase with an increase in the probe set size. This suggests that conditional variance used as the filtering score for the I/NI Calls favours a larger probe set size.

Table 6: Effect of sample and probe set size on the performance of the gene filtering methods. $\rho_I = U(0.5, 1)$ and $\rho_{NI} = (0, 0.5)$ for informative and non-informative probe sets, respectively.

		False Negatives					False Positives				
			I/NI	LMM				I/NI	LMM		
		FLUSH	Calls	ρ	AIC	BIC	FLUSH	Calls	ρ	AIC	BIC
Sample size	3	0.477	0.073	0.329	0.245	0.188	0.393	0.595	0.119	0.015	0.023
	6	0.475	0.107	0.212	0.081	0.077	0.393	0.214	0.093	0.050	0.052
	9	0.471	0.117	0.168	0.027	0.028	0.392	0.129	0.079	0.109	0.107
	18	0.468	0.112	0.111	0.001	0.001	0.393	0.059	0.059	0.309	0.298
	21	0.468	0.114	0.104	0.000	0.000	0.392	0.051	0.057	0.362	0.351
	27	0.467	0.103	0.083	0.000	0.000	0.392	0.038	0.051	0.448	0.437
	30	0.467	0.109	0.084	0.000	0.000	0.392	0.033	0.047	0.481	0.469
	42	0.465	0.109	0.069	0.000	0.000	0.392	0.022	0.041	0.574	0.564
	50	0.464	0.109	0.064	0.000	0.000	0.392	0.019	0.039	0.616	0.607
	100	0.463	0.101	0.042	0.000	0.000	0.392	0.006	0.028	0.749	0.742
Probe set size	8	0.468	0.197	0.082	0.000	0.000	0.392	0.010	0.049	0.459	0.436
	11	0.466	0.160	0.078	0.000	0.000	0.392	0.015	0.047	0.512	0.495
	13	0.468	0.141	0.076	0.000	0.000	0.392	0.017	0.046	0.535	0.520
	15	0.467	0.133	0.076	0.000	0.000	0.392	0.019	0.044	0.554	0.541
	20	0.467	0.112	0.072	0.000	0.000	0.392	0.023	0.043	0.585	0.576
	30	0.468	0.096	0.073	0.000	0.000	0.392	0.030	0.043	0.621	0.615
	40	0.470	0.083	0.070	0.000	0.000	0.392	0.033	0.041	0.640	0.635
	50	0.464	0.053	0.046	0.000	0.000	0.392	0.037	0.042	0.634	0.631
	69	0.477	0.041	0.039	0.000	0.000	0.393	0.038	0.040	0.666	0.663

5.2 Investigation of the Assumption that All Probes in a Probe Set Quantify the Expression Levels of the Same Target.

In Section 5.1, we assume that probe sets can either be informative or non-informative. In this section, we investigate the effect of cross-hybridization. We assume that for an informative probe set, a certain proportion of the probes $(1 - p)$ are non-informative; while p of the probes are informative. Hence, the proportion of probes that quantify expression level of the same target is p. The values $p = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$ are investigated. For the informative probe sets, p of the probes in a probe set are generated with ρ_I and the remaining $1 - p$ of the probes in the same probe set were generated with ρ_{NI}. Note that in this simulation setting the sample size and probe set size are fixed at 42 and 20.

DOI: 10.2202/1544-6115.1460

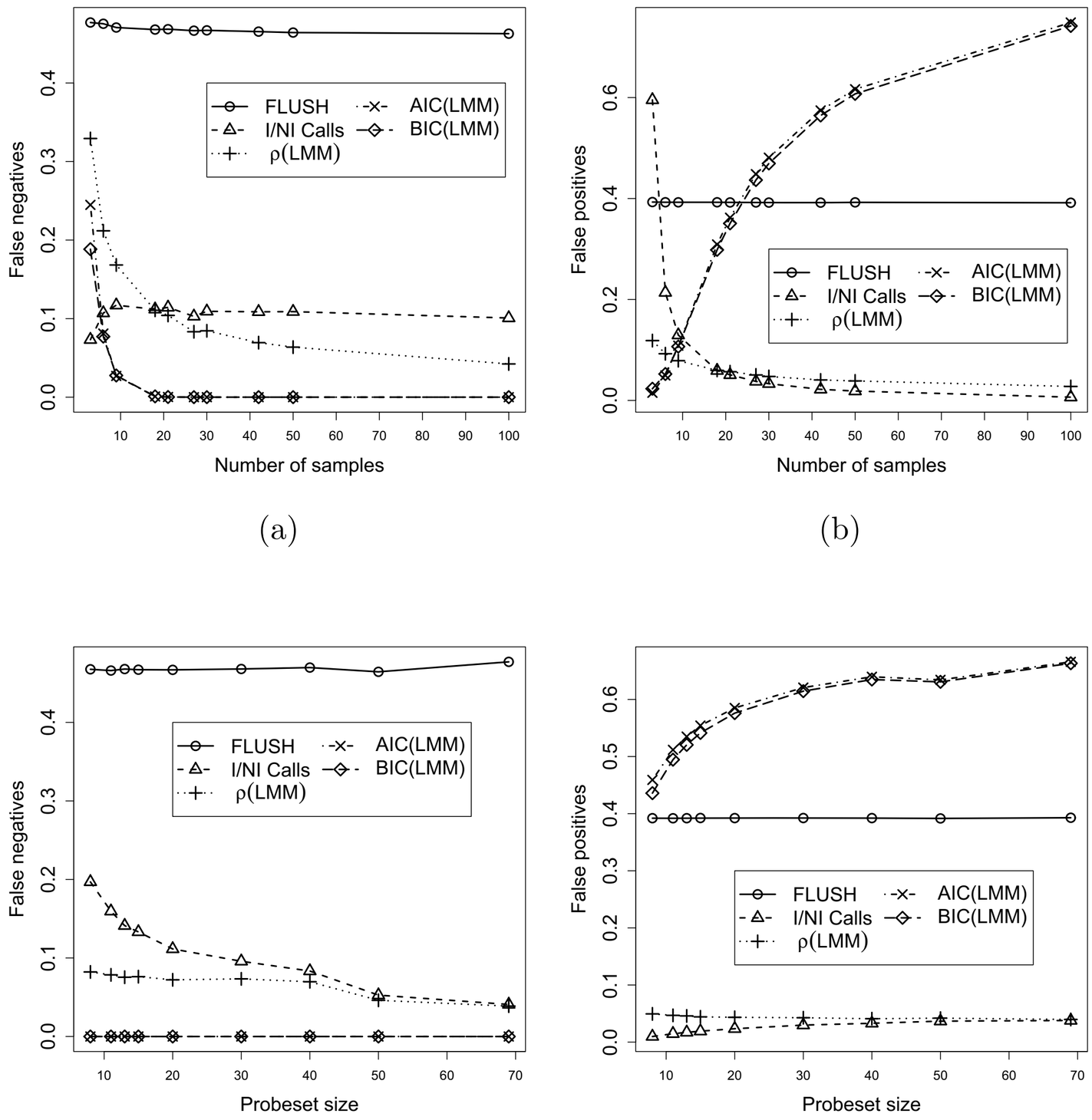


Figure 7: Investigation of the performance of gene filtering methods; (a) Effect of sample size on false negatives, (b) effect of sample size on false positives, (c) effect of probe set size on false negatives, and (d) effect of probe set size on false positives.

The results for the settings $\rho_I = 0.9$ and $\rho_{NI} = 0.1$, and $\rho_I = U(0.5, 1)$ and $\rho_{NI} = (0, 0.5)$ are presented in Table 7. The results based on $\rho_I = 0.9$ and $\rho_{NI} = 0.1$ indicate that all the methods correctly identify the informative probe sets when $p \geq 0.8$. However, for the setting $\rho_I = U(0.5, 1)$ and $\rho_{NI} = (0, 0.5)$, the FLUSH method results in a lower proportion of false negatives than the I/NI Calls and the linear mixed model for $p < 0.5$. The I/NI Calls and linear

Table 7: Effect of cross-hybridization on the performance of the gene filtering methods.

		False Negatives					False Positives				
			I/NI	LMM				I/NI	LMM		
	p	FLUSH	Calls	ρ	AIC	BIC	FLUSH	Calls	ρ	AIC	BIC
$\rho = (0.10, 0.90)$	0.1	0.902	1.000	1.000	1.000	1.000	0.415	0.000	0.000	0.001	0.000
	0.2	0.814	1.000	1.000	0.990	0.994	0.410	0.000	0.000	0.001	0.001
	0.3	0.407	1.000	1.000	0.236	0.295	0.389	0.000	0.000	0.001	0.001
	0.4	0.236	1.000	1.000	0.000	0.001	0.380	0.000	0.000	0.001	0.000
	0.5	0.252	0.035	1.000	0.000	0.000	0.381	0.000	0.000	0.001	0.001
	0.6	0.202	0.000	0.715	0.000	0.000	0.378	0.000	0.000	0.001	0.000
	0.7	0.077	0.000	0.002	0.000	0.000	0.372	0.000	0.000	0.001	0.001
	0.8	0.000	0.000	0.000	0.000	0.000	0.368	0.000	0.000	0.001	0.001
	0.9	0.000	0.000	0.000	0.000	0.000	0.367	0.000	0.000	0.001	0.000
$\rho \sim$ (U(0,0.5), U(0.5,1.0))	0.1	0.623	0.992	0.999	0.416	0.434	0.400	0.023	0.042	0.570	0.559
	0.2	0.634	0.998	1.000	0.287	0.317	0.401	0.023	0.041	0.570	0.559
	0.3	0.642	1.000	1.000	0.142	0.162	0.401	0.022	0.041	0.569	0.559
	0.4	0.643	0.994	1.000	0.062	0.073	0.402	0.022	0.041	0.570	0.560
	0.5	0.634	0.646	1.000	0.016	0.019	0.401	0.022	0.042	0.570	0.559
	0.6	0.618	0.426	0.769	0.003	0.003	0.400	0.023	0.041	0.570	0.559
	0.7	0.587	0.277	0.545	0.000	0.000	0.399	0.022	0.041	0.569	0.559
	0.8	0.549	0.173	0.330	0.000	0.000	0.396	0.023	0.041	0.569	0.559
	0.9	0.509	0.103	0.148	0.000	0.000	0.394	0.022	0.042	0.570	0.559

mixed model treat a probe set with less than half of its probes quantifying the expression level of its designated gene as non-informative. We refer to Section 3 of the supplementary document for an elaborate discussion of the simulation setting and results.

6 Discussion

The strength and weakness of microarray technology can be attributed to the enormous amount of information generated by this technology. To fully enhance the benefit of microarray technology for testing differentially expressed genes, there is a need to minimize the amount of irrelevant genes present in a microarray dataset prior to testing. In this paper, our major interest is to use the linear mixed model for informative or non-informative calls for gene expression data.

We propose alternative methods for informative or non-informative calls based on a linear mixed model with a random intercept and confirmatory

DOI: 10.2202/1544-6115.1460

factor analysis models. The linear mixed model appears to out-performs the FLUSH method. However, similar to the FLUSH, the linear mixed model results in both false positive and false negatives when quantile regression is used. The random intercept linear mixed model performs as well as the I/NI Calls. Also, the confirmatory factor analysis model is equivalent to a random intercept linear mixed model, when either the factor loadings are set to a constant with the variance of the latent factor equal to one, or the factor loadings are set to one with the variance of the latent factor left unconstrained, but non-negative. In addition, we have shown that informative or non-informative calls can be made based on formal hypothesis testing, that are more objective compared to an arbitrary cut-off, such as 0.5. Note that the likelihood ratio test requires corrections for boundary problems and multiplicity. For the spiked-in data, it is noted that conditional variance and intra-cluster correlation result in a smaller number of informative genes as compared to the likelihood ratio test. This is expected since the likelihood ratio test is used to test the null hypothesis that $\rho = 0$, while the filtering score uses a cut-off point of $\rho = 0.5$.

Two simulation studies have been carried out to compare the performance of the methods. The first simulation study compares the three methods under increasing sample size and probe set size. The results indicates that the I/NI Calls and the linear mixed performed better than the FLUSH method in terms of both the numbers of false negatives and false positives. This is especially the case when there is a large array-to-array variability in the dataset. The second simulation study compares the performance of the methods under the violation of the assumption that all probes in a probe set quantify the expression level of the same target. All the three methods result in a high proportion of false negatives, especially when the proportion of probes that quantify the expression level of the same target is less than half of the probe set size.

We have shown that the I/NI Calls, the confirmatory factor analysis models, and the linear mixed model capture the array-to-array variability using a latent factor. The difference between these models is that the I/NI Calls and the "FA-Free" model use the factor loadings to capture array-to-array variability; while the "FA-Restricted" model and the linear mixed model use the variance of the latent factor. The I/NI Calls is based on the conditional variance. We have shown that filtering scores based on the proportion of the variability explained by the latent factor (R^2 and ρ) can be used as well. We recommend to use the latent variable models, i.e., the I/NI Calls or the linear mixed model for gene filtering for the Affymetrix microarray platform. Our reservation for the FLUSH method is based on its violation of the domain

knowledge of the Affymetrix platform, as well as its empirical results. We also recommend the use of R^2 over the conditional variance as a filtering score for the I/NI Calls since the conditional variance has the tendency to favor informative calls for a larger probe set size.

References

Affymetrix (2002) Algorithms Description Document. *Affymetrix Santa Clara*, **CA**

Akaike, H. (1973) Information thoery and an extension of the maximum likelihood principle, B. Petrov and B. Csaki (eds), *Second International Symposium on Information Theory*, Academiai Kiado, Budapest, 267-281.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. B Met.*, **57**, 289-300.

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, New-York, Springer.

Calza, S., Raffelsberger, W., Ploner, A., Sahel, J., Leveillard, T., Pawitan, Y.(2007) Filtering genes to improve sensitivity in oligonucleotide microarray data analysis *Nucleic Acids Research*, **35(16)**, **e102**

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data.*Journal of the American Statistical Association*, **98**, 77-87.

Hochreiter, S., Clevert, D., Obermayer, K.(2006) A new summarization method for Affymetrix probe level data *Bioinformatics*, **22(8)**, 943-949.

McGee, M., Chen, Z. (2006) New Spiked-In Probe Sets for the Affymetrix HGU-133A Latin Square Experiment. [http://biostats.bepress.com/cobra/ps/art5] *COBRA Preprint Series* , Article **5**.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M (2002) Large-scale analysis of the human and ouse transcriptomes. *PNAS*, **99**, 4465-4470.

DOI: 10.2202/1544-6115.1460

Talloen, W., Clevert, D., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S., G*ö*hlmann, W.H.H (2007) I/NI Calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data *Bioinformatics*, **23**, 2897-2902.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data* , Springer, New York.