Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays

Leo Lahti, *Member, IEEE,* Laura L. Elo, Tero Aittokallio, and Samuel Kaski *Senior Member, IEEE*

***Abstract*— Probe defects are a major source of noise in gene expression studies. While existing approaches detect noisy probes based on external information such as genomic alignments, we introduce and validate a targeted probabilistic method for analyzing probe reliability directly from expression data and independently of the noise source. This provides insights into the various sources of probe-level noise and gives tools to guide probe design.**

***Index Terms*— Applications, Biology and genetics, Parameter learning, Probabilistic algorithms**

I. Introduction

GENE expression profiling is widely used to explore gene function in various biological conditions, and vast collections of microarray data are available in public repositories. These large-scale data sets contain valuable information of both biological and technical aspects of gene expression studies [1], [2], [3], [4]. However, gene expression data is notoriously noisy. A better understanding of the technical aspects of the measurement process could ultimately lead to enhanced measurement techniques and improved analytical procedures, providing more accurate biological results in future studies.

Short oligonucleotide arrays of Affymetrix [5] are one of the most widely used gene expression profiling platforms. These arrays utilize multiple (typically 10-20) 25-mer probes, the so-called probe set, to measure the expression level of each transcript target. The probes within an individual probe set are designed to target the same gene, and ideally they should detect the same gene expression signal. Use of several probes for each target leads to more robust estimates of transcript activity, but the reliability of individual probes is known to vary and may significantly affect the results of a microarray study [6]. For example, it has been noticed that a considerable number of probes on short oligonucleotide arrays do not uniquely match their intended targets [7], [8], [9]. Single-nucleotide polymorphisms, alternative splicing and non-specific hybridization add biological variation in the data [10], [11]. Other factors in the measurement process that cause probe-specific effects include RNA extraction and amplification, binding affinities, and experiment-specific variation [12], [13].

Many preprocessing algorithms utilize probe-specific parameters to obtain probeset-level summaries of gene expression. These include MBEI/dChip [14], RMA [15], gcRMA [16], FARMS [17], gMOS [18], and BGX [19]. Despite the importance of probe-specific effects in gene expression analysis and probe design [6], [20], the various sources of probe-level noise are still poorly understood. Only a few studies have systematically analyzed the factors affecting probe reliability. The existing approaches typically rely on external information such as genomic sequence data [8], [9], [11] or physical models [21], [22], [23], and cannot reveal probes that are less reliable due to so far unknown reasons.

We introduce and validate a targeted computational tool for probe reliability analysis. In contrast to previous probe quality studies, the proposed model is independent of external information or physical models. This can advance the understanding of the various factors that affect probe reliability. Our approach is closely related to preprocessing methods that utilize probe-specific parameters to obtain probeset-level summaries of gene expression. A key difference in our work is that we assign an explicit probabilistic measure of reliability to each probe and demonstrate how this information can be used to assess probe performance. Explicit estimates and analysis of probe-specific noise have been missing in preprocessing studies. The method is applied to gene expression data sets from two human genome arrays, HG-U95A/Av2 and HG-U133A, and the results are validated by comparisons to known probe-level error sources: errors in probe-genome alignment, interrogation position of a probe on the target sequence, GC-content, and the presence of SNPs in the target sequences of the probes. Implementation of the method is available in R[1] at http://www.cis.hut.fi/projects/mi/software/RPA.

II. Modeling of probe reliability

The reliability of a probe is ultimately determined by its ability to measure the expression level of the target transcript. As the true expression level is unknown in most practical situations, the collection of probes measuring the same transcript can provide the ground truth for assessing probe performance (Supplementary Fig. 1). Our model captures the most coherent signal of the probe set, and the reliability of individual probes is estimated with respect to this signal across a large number of arrays. We provide an explicit probabilistic model for probe-level observations, and derive the posterior distribution for the model parameters describing probe reliability and differential gene expression. While probe-level preprocessing algorithms aim at summarizing probe-level measurements [14], [15], [17], [18], [19], we have specifically targeted a more detailed analysis of

- L. Lahti and S. Kaski are with Helsinki Institute for Information Technology and Adaptive Informatics Research Centre, Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland. E-mails: leo.lahti@tkk.fi, samuel.kaski@tkk.fi.
- L.L. Elo and T. Aittokallio are with Department of Mathematics, University of Turku, FI-20014 Turku, Finland and Turku Centre for Biotechnology, P.O. Box 123, FI-20521 Turku, Finland. E-mails: laliel@utu.fi, teanai@utu.fi.

[1] http://www.r-project.org/

Digital Object Indentifier 10.1109/TCBB.2009.38

1545-5963/$25.00 © 2009 IEEE

probe reliability. This avoids certain problems encountered in the preprocessing context as discussed in the next section.

A. Model assumptions

Our approach is based on a Gaussian model for probe effects. This is a reasonable starting point for modeling heterogeneous and partially unknown sources of probe-level noise. The feasibility of related models has already been demonstrated in the preprocessing context [15], [17]. In a nutshell, we assume normally distributed probe effects, and identify probe reliability with its variance over a large number of arrays. In contrast to many probe-level preprocessing methods, where the mean is the important quantity, we use probe-level observations of differential expression. Then the mean cancels out, and the model can focus on estimating the variances (see Methods for details).

Variance reflects the noise level of the probe and is the main focus in our analysis. This is different from probe-level preprocessing methods that focus on estimating probe affinities, corresponding to the mean parameter of the Gaussian noise model. For example, the probe-specific parameters in MBEI [14] and RMA [15] preprocessing models describe probe affinities. These are constant shifting factors and as such not informative of probe reliability. Moreover, unidentifiability of probe affinities is a known problem in preprocessing studies [15], [24]. The recently suggested FARMS preprocessing algorithm [17] has a more complex model for probe effects than RMA and contains implicitly a similar probe-specific variance parameter as our model. However, FARMS does not provide explicit estimates of the probe-related parameters and is therefore not applicable to probe reliability analysis.

We avoid the modeling of unidentifiable probe affinities by using probe-level observations of differential gene expression. Probe effects are captured in a single probe-specific variance parameter in the resulting model. The number of probe-related parameters in the model is halved, and faster and more robust inferences concerning the parameters of interest can be obtained. Use of a single parameter for probe effects also leads to more straightforward interpretations of probe reliability. Cancellation of the probe affinity parameters in our analysis can partly explain the previous observations that calculating differential expression at probe-level improves the analysis of differential gene expression [25], [26]. However, these methods differ from our approach in that they are non-probabilistic preprocessing methods that do not aim at quantifying the uncertainty in the probes.

B. Comparison to known error sources

The model is applied to six publicly available gene expression data sets, including four large-scale studies on human samples [27], [28], [29], [30], referred to as ALL and GEA data sets, and two spike-in data sets from Affymetrix (www.affymetrix.com), referred to as SPIKE data sets (Table 1). The data sets have been measured using two popular human genome arrays, HG-U95A/Av2 and HG-U133A. To validate our model and to analyze probe reliability on these arrays, we test the overrepresentation of the following probe-level error sources among the least reliable probes predicted by our model.

TABLE I
GENE EXPRESSION DATA SETS IN THIS STUDY

Name	Platform	Arrays	Author
ALL-95Av2	HG-U95Av2	37	Yeoh et al. (2002)
GEA-95A	HG-U95A	85	Su et al. (2002)
SPIKE-95Av2	HG-U95Av2	59	Affymetrix
ALL-133A	HG-U133A	37	Ross et al. (2003)
GEA-133A	HG-U133A	158	Su et al. (2004)
SPIKE-133A	HG-U133A	42	Affymetrix

1) Probe-genome alignments: Ideally, each probe has a unique sequence match to its target gene. In practice, a number of probes do not uniquely match their intended mRNA target. Filtering of probes with erroneous genome alignments has previously been shown to improve the accuracy and comparability of microarray results [8], [9], [11], [26], [31]. A good model for estimating probe reliability should detect such erroneous probes.

2) Interrogation position on the target sequence: RNA degradation, typically starting from the 5' end of the transcript, has been reported to affect the results in microarray studies [32], [33]. Hence, the binding location of the probe on the target sequence, i.e., its *interrogation position*, is likely to affect probe reliability.

3) GC-content: Various hybridization effects that are based on the nucleotide content of the probes have been reported [21], [22], [23], [34]. For example, the G/C nucleotides have a higher binding affinity since G-C pairs form three hydrogen bonds whereas the A-T pairs form two. Therefore the GC-content of a probe is expected to affect its reliability.

4) SNP associations: Probes that target sequences with common single-nucleotide polymorphims (SNPs) can produce misleading results in microarray studies [10], [35], [36]. Each probe can measure accurately at most one of the polymorphic target sequences and therefore gene expression differences between two individuals can be observed in some probes due to sequence polymorphism rather than real expression changes. This would add noise to microarray data. It is expected that SNPs located in the central region of the target sequence will have a greater influence on probe reliability than other SNPs due to a larger impact on probe affinity [21], [37].

C. Connection to preprocessing

The reliability of a probe is ultimately measured by its ability to capture the real underlying gene expression signal. This is unknown in most practical situations, however, and needs to be estimated from the probe-level observations. Probe reliability estimates are sensible only if the true signal is estimated accurately in our model. To guarantee this, the performance of the proposed model in estimating relative gene expression changes was compared to four alternative approaches: MAS5.0 (www.affymetrix.com) and RMA [15] are among the most widely applied methods for assessing probe set-level signals (which are then used to calculate the expression changes); FARMS [17] represents the previously introduced probe-level models; and PECA [38] shares the idea of directly utilizing probe-level expression changes. Note that the other methods do not provide explicit estimates of probe reliability, while our method provides only estimates of relative gene expression changes. A general difference between preprocessing algorithms and our method is that preprocessing methods have been designed to summarize probe-level information, whereas our model is specifically targeted at estimating certain probe-specific effects that are then used to analyze probe reliability.

III. Methods

A. *Probabilistic model*

In the following, we describe a probabilistic model for probe reliability and differential gene expression. In the calculations, we use the logarithmized perfect match (PM) intensities of the Affymetrix arrays, and investigate each probe set separately. Affymetrix arrays also contain so-called mismatch (MM) probes that have an altered nucleotide in the middle (13th) position of the probe. These were originally designed to measure cross-hybridization from unrelated sequences. Some widely used preprocessing algorithms, such as RMA, ignore the MM probes due to the lack of efficient models for utilizing this information [15].

1) Conditional likelihood for the observations: Let us consider a probe set targeted at measuring the expression level of target transcript g. We model probe-level observations as a sum of the true expression signal that is common for all probes, and probe-specific Gaussian noise. A probe-level observation for probe j on array i can then be written as $s_{ij} = g_i + \mu_j + \varepsilon_{ij}$. The mean parameter μ_j describes the systematic probe affinity effect, and the stochastic noise component is distributed as $\varepsilon_{ij} \sim N(0, \tau_j^2)$.

The variance parameters $\{\tau_j^2\}$ are of interest in probe reliability analysis. To focus on these parameters we take advantage of the fact that the unidentifiable probe affinity parameters $\{\mu_j\}$ cancel out when the signal log-ratio between a randomly selected 'control' array and the remaining arrays is computed for each probe. The differential expression signal between arrays $t = \{1, \ldots, T\}$ and the control array c for probe j is then $m_{tj} = s_{tj} - s_{cj} = g_t - g_c + \varepsilon_{tj} - \varepsilon_{cj} = d_t + \varepsilon_{tj} - \varepsilon_{cj}$. Using vector notation, the differential gene expression profile of probe j across the arrays $\{t\}$ is now $\mathbf{m}_j = \mathbf{d} + \boldsymbol{\varepsilon}_j$, where the two noise terms have been combined into a single variable $\boldsymbol{\varepsilon}_j$. Note that the control-related noise ε_{cj} is constant across the comparisons whereas the second noise component ε_{tj} depends on the array t.

To identify the probe-specific variance parameter, shared by the two noise components in $\boldsymbol{\varepsilon}_j$ for each probe j, we consider the control-related noise ε_{cj} a hidden variable in our model. This can be marginalized out by assuming that the probe-level observations $\mathbf{m}_j$ of the true underlying signal $\mathbf{d}$ are independent given the model parameters. Let us also denote the collection of probe-level signals of a probe set by $\mathbf{m} = \{\mathbf{m}_j\}$. The likelihood for the observations is then

$$\begin{aligned} P(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2) &= \prod_{tj} \int N(m_{tj}|d_t - \varepsilon_{cj}, \tau_j^2) N(\varepsilon_{cj}|0, \tau_j^2) d\varepsilon_{cj} \\ &\sim \prod_j (2\pi\tau_j^2)^{-\frac{T}{2}} exp(-\frac{\sum_t (m_{tj} - d_t)^2 - \frac{[\sum_t (m_{tj} - d_t)]^2}{T+1}}{2\tau_j^2}). \end{aligned} \tag{1}$$

2) Posterior distribution of the model parameters: The posterior density for the model parameters is computed from the conditional likelihood of the data (Eq. (1)) and the prior according to Bayes rule:

$$P(\mathbf{d}, \boldsymbol{\tau}^2|\mathbf{m}) \sim P(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2) P(\mathbf{d}, \boldsymbol{\tau}^2). \tag{2}$$

We use a non-informative prior for $\mathbf{d}$, and conjugate priors for the variance parameters in $\boldsymbol{\tau}^2$ (inverse Gamma distribution, see [39]). Using a standard assumption that $\mathbf{d}$ and $\boldsymbol{\tau}^2$ are independent with $P(\mathbf{d}|\boldsymbol{\tau}^2) \sim 1$, the prior takes the form $P(\mathbf{d}, \boldsymbol{\tau}^2) \sim \prod_j invgam(\tau_j^2; \alpha_j, \beta_j)$, where α_j and β_j are the parameters of the inverse Gamma distribution. These parameters are probe-specific and allow incorporation of prior information about probe reliability into the analysis.

The final model for probe intensities is hence described by two sets of parameters; the vector of underlying differential gene expression signals $\mathbf{d} = [d_1 \ldots d_T]$, and the probe-specific variance parameters $\boldsymbol{\tau}^2 = [\tau_1^2 \ldots \tau_J^2]$. High variance τ_j^2 would indicate that the probe-level observation $\mathbf{m}_j$ is strongly deviated from the estimated true signal $\mathbf{d}$. The Bayesian formulation quantifies the uncertainty in the model parameters, and allows incorporation of prior information about probe reliability into the analysis. We refer to this procedure as *Robust Probabilistic Averaging (RPA)*.

3) Implementation: In this paper, we use the posterior mode as a point estimate for the model parameters. This is searched for by iteratively optimizing $\mathbf{d}$ and $\boldsymbol{\tau}^2$ in Eq. 2. The model is initialized to give equal prior weight for each probe by setting $\tau_j^2 = 1$ for each probe j. A mode for $\mathbf{d}$, given $\boldsymbol{\tau}^2$, is searched for by a standard quasi-Newton optimization method [40]. The variance parameters τ_j^2 follow an inverse Gamma distribution with parameters $\hat{\alpha}_j = \alpha_j + \frac{T}{2}$ and $\hat{\beta}_j = \beta_j + \frac{1}{2}(\sum_t (m_{tj} - d_t)^2 - \frac{(\sum_t (m_{tj} - d_t))^2}{T+1}))$ given $\mathbf{d}$. The mode is then given by $\tau_{j,new}^2 := \hat{\beta}_j / (\hat{\alpha}_j + 1)$. We use non-informative priors with $\alpha_j = \beta_j = 10^{-5}$.

B. *Data*

Only the common probe sets of the HG-U95A and HG-U95Av2 platforms were used, referred to as HG-U95A/Av2. Probe intensities were quantile-normalized, and the AFFX control sets excluded before the analysis.

1) Leukemia data (ALL): The public ALL data sets from the microarray studies of Ross *et al.* [27] and Yeoh *et al.* [30] contain expression data from patients with various leukemia subtypes. A total of 360 patient samples have been hybridized to HG-U95Av2 arrays and 132 of the same samples are additionally hybridized to HG-U133A arrays. For our analyses we selected 37 samples that were hybridized to both array types and represent homogeneous patient groups with five distinct leukemia subtypes and control patients (Table 1). We refer to these two data sets as ALL-95Av2 and ALL-133A, respectively.

2) Gene expression atlas (GEA): The gene expression atlases of Su *et al.* [28], [29] cover a diverse set of biological conditions measured on the human array platforms HG-U95A and HG-U133A (Table 1). We refer to these two data sets as GEA-95A and GEA-133A, respectively. Some samples in the HG-U95A data were ignored because no biological replicates were available.

3) Affymetrix spike-in data (SPIKE): The Affymetrix HG-U95Av2 and HG-U133A spike-in data sets were downloaded from the Affymetrix web pages (www.affymetrix.com). We refer to these data sets as SPIKE-95Av2 (59 hybridizations) and SPIKE-133A (42 hybridizations). A total of 14 and 42 genes have been spiked-in at known concentrations on the HG-U95Av2 and HG-U133A arrays, respectively, and arrayed in a Latin Square format. Recently, it has been demonstrated that 22 additional probe sets in the SPIKE-133A data set should also be considered as spiked [41]. Accordingly, we utilized the extended set of 64 spiked probe sets when evaluating the performance of the different analysis approaches in the SPIKE-133A data.

4) Probe sequence data: Probe sequences and their best-match tables were downloaded from the Affymetrix web pages (www.affymetrix.com). Other array-wise information on probes and probe sets, including probe locations on the array, were acquired from the annotation data packages of the Bioconductor project [42]. Human genomic mRNA sequences were downloaded from Entrez Nucleotide [43] on August 16, 2006, excluding EST, STS, GSS, working draft and patents sequences, and sequences with a 'XM_*' tag, as in [8], [26].

5) Probe-genome alignment: To identify probes having errors in the genomic alignment, all probes on the HG-U95A and HG-U133A arrays were aligned to the nucleotide sequences from Entrez Nucleotide, and assigned GeneIDs according to their matched sequence. Perfect matches of the probes to mRNA sequences were sought with BLAT v. 26 [44], following the same procedure as in [8], [26], but using updated genomic sequence data. The Entrez mRNA sequences were assigned to GeneID identifiers by using the 'gene2accession' conversion file obtained from NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA, August 10, 2006). The percentage of probes with no GeneID match was 9.4% and 10.1% for the HG-U95A and HG-U133A arrays, respectively. Multiple GeneID matches were detected for 4.6% (HG-U95A) and 4.8% (HG-U133A) of the probes.

6) Single-nucleotide polymorphisms: Information about the probe-SNP associations was provided by the CustomCDF BioConductor package [10] that contains SNP mapping for the probes based on data from the dbSNP database [43]. The mappings have been used to investigate SNP effects in microarray data in recent studies [36], [45]. To focus on common SNPs, we considered only SNPs with a minimum population frequency of 5%.

IV. Results

The RPA algorithm was applied on gene expression data sets from two commonly used microarray platforms to validate the model and to assess the differences between known probe-level noise sources. First, we compared probe reliability estimates to known probe-level error sources. Second, preprocessing comparisons were used to test the preprocessing performance of RPA and, importantly, to guarantee the validity of the probe reliability measures that depend on accurate estimation of the differential gene expression signal.

A. Comparison to known error sources

1) Probe-genome alignment: Mistargeted probes that did not uniquely match the GeneID target of the probe set were significantly enriched ($p < 0.05$; hypergeometric test) among the least reliable 1% of the probes detected by our model (Fig. 1; Tab. S1). The mistargeted probes were 1.1-1.7 times more common in the HG-U95A/Av2 data sets than expected, and 2.2-3.1 times more common in HG-U133A. The enrichment of mistargeted probes was the highest for the probes that were consistently unreliable in the independent GEA and ALL data sets. On the HG-U133A array, mistargeted probes could explain 20.4% of the least reliable probes while the expected proportion was 6.7%. Consistently unreliable probes were detected by using the average rank of the probes obtained in the two experiments. Detection of probes having errors in their genomic alignment was expected because such probes do not necessarily have any correlation with the probe set-level signal. This supports the validity of our model.

2) Interrogation position: The interrogation position of a probe on the target sequence was significantly associated with probe reliability ($p < 0.05$; χ^2-test). Probes closest to either end of the target sequence were enriched among the least reliable probes; the observed counts deviated 73 - 138% from the expectation, depending on the interrogation position (Fig. 2(a); Tab. S1). Enrichment of 5'-binding probes was expected due to RNA degradation starting from this end of the transcript. Enrichment of 3' probes is supported by previous findings of Dai et al., who noticed that 3'-focused probe sets have often increased noise levels [10]. Probes closer to the 3' end detect, on average, a higher absolute signal. A higher signal is often associated with higher noise in microarray studies [46], which could explain our observation. Alternative transcription may also cause differences between 3' probes and the other probes [47], [48].

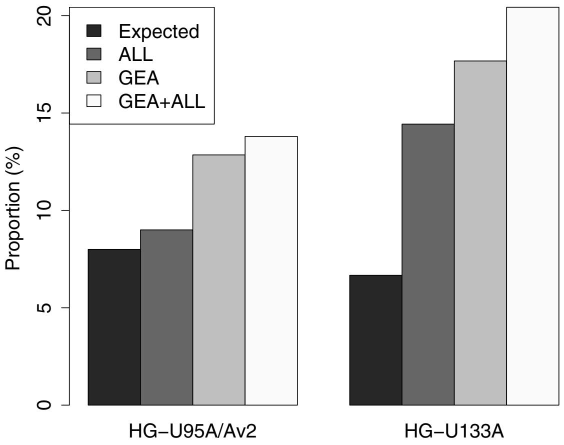
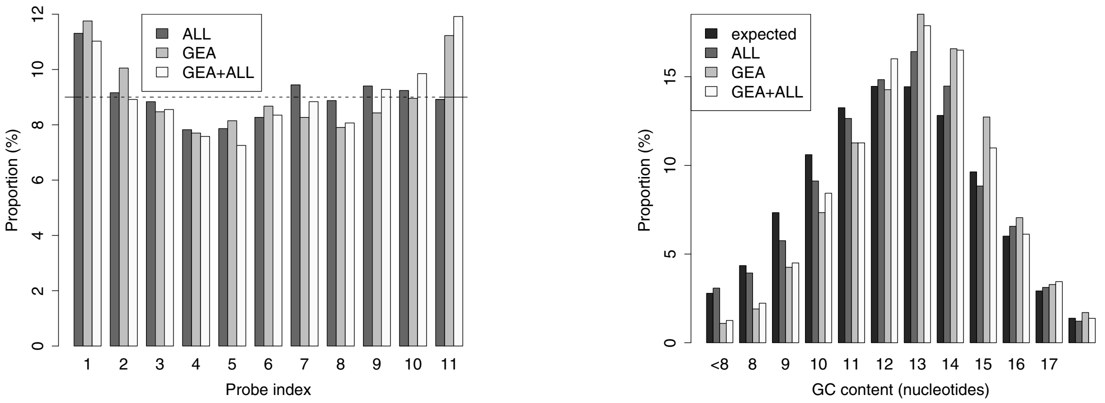


Fig. 1. Genomic alignment and probe reliability. Mistargeted probes that do not uniquely match the GeneID target of the probe set were enriched among the least reliable probes ($p < 0.05$; hypergeometric test). Black bars show the expected proportion of mistargeted probes i.e. their proportion on the whole array. Grey bars show the proportion of mistargeted probes among the least reliable 1% of the probes detected by our model (dark: ALL; light: GEA; white: combined results).

3) GC-content and probe reliability: GC-rich probe sequences were enriched among the least reliable probes of our model in all data sets except ALL-95Av2 and GEA-95A (Fig. 2(b); Tab. S1). The observed counts for the different GC contents deviated 39-132% from the expectation in the investigated data sets ($p < 0.05$; χ^2-test). To guarantee the assumptions of the χ^2-test, probes with most extreme G/C or A/T contents were combined in the test. One explanation for our observation is that high-affinity probes may have higher likelihood of cross-hybridization to nonspecific targets [21]. This would add noise to the probe-level signal.

4) Single-nucleotide polymorphisms: Probes whose target sequences have common SNPs were enriched among the least reliable probes on the HG-U133A platform and in the combined results from HG-U95A/Av2 platform (Supplementary Fig. 3; Tab. S1). In these data sets, the SNP-associated probes were 1.7 - 1.9 times more common among the least reliable probes than expected ($p < 0.05$; hypergeometric test). It is interesting to notice that the association between probe reliability and SNPs is observed only when information from the ALL-95Av2 and GEA-95A is combined; a similar observation was made with the GC-rich probes. A likely explanation is that the systematic effects from the SNP-associated, or GC-rich probes are more



(a) Interrogation position

(b) GC-content

Fig. 2. Probe reliability vs. interrogation position and GC-content on the HG-U133A platform. **(a)** Probes that bind to either the 5' or the 3' end of the target transcript were enriched among the least reliable (1%) probes ($p < 0.05$; χ^2-test). Probe index indicates the relative interrogation position of the probe on the target sequence, starting from the 5' end of the transcript. The grey bars show the proportion for each interrogation position among the least reliable probes in the inspected data sets (dark: ALL; light: GEA; white: combined results). The expectation is illustrated by the dashed line. There are 11 probes per probe set on the HG-U133A arrays. **(b)** GC-rich probes were enriched among the least reliable (1%) probes ($p < 0.001$; χ^2-test). The GC-content of a probe is indicated by the number of G/C nucleotides on the 25-mer probes. Grey bars show the proportion of each GC-content among the least reliable probes (dark: ALL; light: GEA; white: combined results). Consistently less reliable probes (GEA+ALL) had the highest deviation from the expectation (black bars). To guarantee the assumptions of the χ^2-test, we combined probes with most extreme G/C or A/T contents for testing. Results for the HG-U95A/Av2 data sets are shown in Supplementary Fig. 2.

effectively observed when the data sets are combined and the data set specific noise cancels out. In general, the SNP-associated probes were less reliable than the other probes in all investigated data sets ($p < 0.05$; Wilcoxon test). As expected, probes having a single SNP in the central 13bp region of the 25-mer probe were less reliable than probes with a single SNP in either end of the target sequence on HG-U133A ($p < 0.05$; Wilcoxon test) but, interestingly, not on the HG-U95A/Av2 platform.

5) Relative contribution of the known error sources: Probes that are associated with the investigated noise sources had 7-39% increase in average variance, detected by RPA, in the studied data sets except ALL95-Av2 (Fig. 3). Mistargeted probes had the highest variances on HG-U133A, whereas probes with the most 5'/3' interrogation positions had the highest variances on HG-U95A/Av2. High GC-content led to a more moderate increase in probe-specific variance than the other investigated sources. However, GC-rich probes are more common (28-33% of the probes) than mistargeted probes (6-8%), probes with common SNPs (3-3.4%), or probes in the most 5'/3' positions of the target sequence (10-18%) and have therefore a remarkable contribution to the overall probe-level noise. Interestingly, many (35-60%) of the least reliable probes detected by RPA were not associated with the investigated sources, including many probes that have systematically low reliability in independent data sets.

B. General observations of probe reliability

Examples of the least reliable probes in the GEA-95A data set are shown in Supplementary Fig. 4. Comparison of the results from independent ALL and GEA data sets revealed many probes with consistently poor reliability, although the comparability of the results was affected by data set-specific effects: Spearman correlations of the probe-specific variances $\{\tau_j^2\}$ between the ALL and GEA data sets were 0.28 (HG-U95A/Av2) and 0.52 (HG-U133A). Surprisingly, the least reliable probes in the ALL data sets showed almost identical expression profiles (Supplementary Fig. 5), although they are located in independent probe sets and expected to capture uncorrelated signals. The noise probably originates in the biological samples that have been hybridized on both array types in the ALL-95Av2 and ALL-133A data sets. The specific source of this contamination remains unclear.

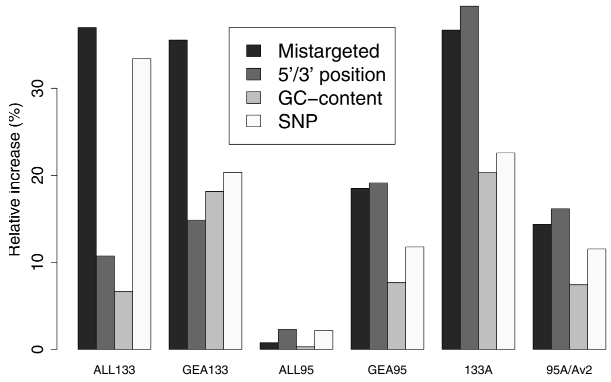


Fig. 3. Increase in the average variance of the probes associated with the investigated noise sources: mistargeted probes having errors in the genomic alignment, most 5'/3' probes of each probe set, GC-rich, and SNP-associated probes. The variances were estimated by RPA and describe the noise level of the probes. The results are shown for the individual ALL and GEA data sets, and for their combined results on both platforms (133A and 95A/Av2).

C. Preprocessing comparisons

The validity of probe reliability estimates depends on accurate estimation of the probe set-level signal. We compared RPA to other preprocessing methods to test its preprocessing performance and to guarantee the validity of probe reliability estimates.

1) Spike-in data: In spike-in data sets, the true expression changes are known and, hence, the different preprocessing approaches can be compared in terms of their receiver operating characteristics (ROC). RPA and PECA were more successful in detecting the spiked genes than MAS5.0 or RMA (Fig. 4). FARMS was found to outperform the other methods when a large number of genes are inspected. The good performance of FARMS in the spike-in data may, however, be favoured by the particular design of the spike-in experiments, in which the expression changes always occur in the same genes. This was supported by the observation that, unlike the other methods, FARMS produced nearly perfect ROC-curves even when replicated samples were compared with each other, although in these comparisons no changes should be detected and the gene rankings should be random (Supplementary Fig. 6).

2) Technical replicates: We also assessed the performance of the different preprocessing methods in real research settings using the ALL and GEA data sets. Since in these data sets the true expression changes were not known, the performance of the different methods was evaluated in terms of their consistency across replicated measurements for both genes and biological samples. Following the approach of Reverter *et al.* (2005) [49], we first measured the consistency of the expression changes within each data set (Supplementary Fig. 7). Specifically, for each GeneID represented by at least two probe sets on an array, the average Pearson correlation of the expression profiles between all the matching probe sets was calculated. Based on our probe-genome alignments, there were 1470 and 3774 such GeneIDs on the HG-U95A/HG-U95Av2 and HG-U133A arrays, respectively. In each data set, RPA produced the highest correlations ($p < 0.05$; paired Wilcoxon test), and PECA and RMA also clearly outperformed not only MAS5.0 but, notably, also FARMS.

To further investigate the performance of the methods, we evaluated the consistency of the expression changes across the two separate data sets, ALL-95Av2 and ALL-U133A, in which the same biological samples have been hybridized (Fig. 5). The consistency was measured by the Pearson correlation between the pairs of arrays, to which the same sample was hybridized. This indicates the performance of the methods, as the technical replicates are assumed to produce effectively the same results on both array versions. The so-called 'bestmatch' tables, provided by the array manufacturer (www.affymetrix.com), were utilized to combine the data across the arrays. The results from this analysis supported the earlier findings. In particular, RPA and PECA outperformed the other approaches; RMA performed better than MAS5.0 and FARMS; and MAS5.0 showed the poorest performance ($p < 0.05$; paired Wilcoxon test). Interestingly, the simple PECA yielded better consistency between the data sets than RPA ($p < 0.05$). While the main focus of this paper is in probe reliability analysis, the preprocessing comparisons confirmed that RPA compares favourably with the other methods in estimating differential gene expression. This guarantees the validity of probe reliability estimates in our model.

V. Discussion

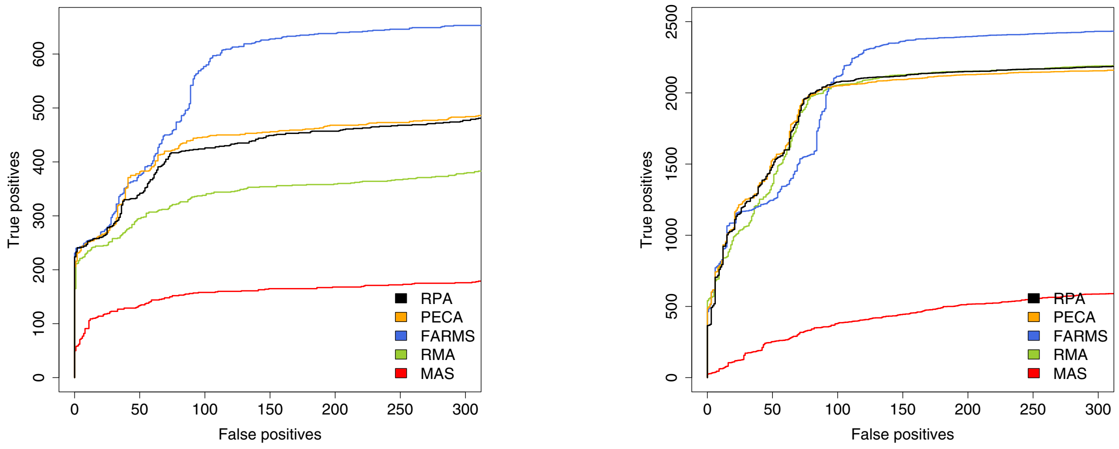
Previous probe-level models have focused on preprocessing of gene expression data, whereas we have specifically targeted a more detailed analysis of probe reliability. Enrichment of known probe-level error sources among the less reliable observed probes validates our model; many of the findings were explained by errors in genomic alignment, probe interrogation position, GC-content, or common SNPs. However, any single source of error seems to explain only a fraction of the probes that have consistently poor reliability in independent data sets. Therefore methods that remove probe-level noise based on external information such as genomic alignments are likely to ignore a large number of the least reliable probes. For example, a probe set designed to measure a certain transcript may additionally detect unknown alternatively spliced transcripts which may have different expression patterns [12], or cross-hybridize with mRNAs having closely similar ($> 18/25$ bp) but not perfectly matching sequences [11]. Various laboratory- and experiment-specific effects are also known to add experimental noise in microarray studies [12], [13]. The proposed model can detect poorly performing probes that are susceptible to noise from such sources.

A Gaussian model for probe effects is a reasonable starting point for modeling heterogeneous and partially unknown sources of probe-level noise. The feasibility of similar models has already been demonstrated in the preprocessing context. For example, the RMA preprocessing algorithm [15] has a Gaussian model for probe effects with probe-specific mean (affinity) parameters and a shared variance parameter for the probes. We avoid the estimation of probe affinities and instead focus on estimating probe-specific variances. The recently suggested FARMS preprocessing algorithm [17] is closely related to our approach but has a more complex model for probe effects. The model can be written as $s_{ij} = z_i\lambda_j + \mu_j + \varepsilon_{ij}$. Here z_i captures the underlying gene expression, and the model has three parameters $\{\lambda_j, \mu_j, \varepsilon_{ij}\}$ for each of the 10-20 probes in a probe set. In contrast, our model has a single variance parameter for each probe. The use of a more complex model in FARMS is justified as it aims at summarizing the absolute values of logarithmized PM intensities. This is a hard task since large systematic differences are known to exist between probes [14], [46]. We have shown that by computing differential gene expression at probe-level avoids the need to estimate unidentifiable probe affinity parameters. Use of a single parameter for probe effects leads to more straightforward interpretations about probe reliability and makes the model potentially less prone to overfitting. This is supported by the observation that RPA and PECA compared favourably with other preprocessing methods in the analysis of differential gene expression. The distinguishing feature of the two methods is that they compute differential gene expression at the probe-level. However, only the probabilistic RPA estimates probe reliability.

While for most probe sets, different preprocessing methods give largely consistent results, their differences can be especially large for probe sets containing several inconsistent probe-level signals. The main contribution of the current study is to introduce and apply a probabilistic model with explicit modeling assumptions to analyze probe reliability on short oligonucleotide arrays. At the same time the model provides a principled framework for incorporating prior information of the probes in differential gene expression analysis. This is a potential topic for future studies.

VI. Conclusion

We have introduced a probabilistic framework for analyzing the reliability of individual probes directly from gene expression data, and validated the model using gene expression data sets from two popular human genome arrays. A major advantage of the proposed approach is its capability to detect unreliable probes



(a) HG-U95A/Av2

(b) HG-U133A

Fig. 4. Preprocessing performance for spike-in data. ROC curves for the various methods that were used to estimate the signal log-ratio: RPA, PECA, RMA, FARMS, and MAS for the two spike-in data sets (Affymetrix HG-U95Av2 and HG-U133A). For each curve, the results from the investigated spike-in samples within the data sets were pooled. The axes have been truncated to focus on the most relevant area. When comparing the curves, the one closest to the upper left corner shows the best performance.

independently of physical models or external, constantly updated information such as genomic sequence data. Probe reliability information can be useful in many applications, including evaluation of the end results of gene expression analysis, and recognition of potentially unknown probe-level error sources. It can be used to quantify the uncertainty in the measurements and in designing the probes, and is also utilized by our model to provide robust estimates of differential gene expression. A better understanding of the various probe-level error sources could advance probe design and contribute to reducing probe-related noise in the future generations of gene expression arrays.

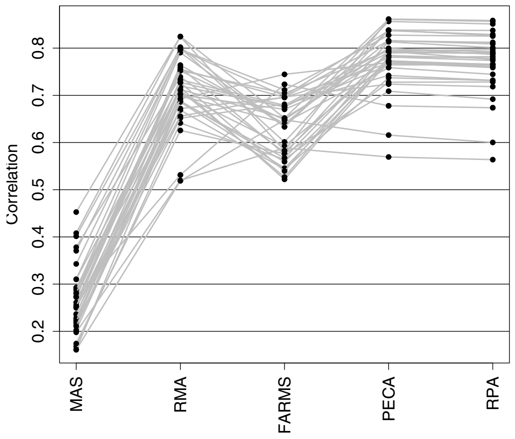


Fig. 5. Reproducibility of signal estimates in real data sets between the technical replicates, i.e., the best match probe sets between the HG-U95Av2 and HG-U133A platforms. The consistency was measured by the Pearson correlation between the pairs of arrays, to which the same sample was hybridized.

Acknowledgment

The work was supported by the Academy of Finland, grants 207467 (LL, SK) and 203632 (LE, TA). SK was additionally funded by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and LE by the Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi) and the Academy of Finland, grant 127575.

References

[1] C. Benedict, M. Geisler, J. Trygg, N. Huner, and V. Hurry, "Consensus by Democracy. Using Meta-Analyses of Microarray and Genomic Data to Model the Cold Acclimation Signaling Pathway in Arabidopsis," *Plant Physiology*, vol. 141, no. 4, pp. 1219–1232, 2006.

[2] Z. Hu, C. Fan, D. S. Oh, J. Marron, X. He, B. F. Qaqish, C. Livasy, L. A. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M. G. Ewend, L. R. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. R. Orrico, D. Dreher, J. P. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J. F. Quackenbush, M. J. Ellis, O. I. Olopade, P. S. Bernard, and C. M. Perou, "The molecular portraits of breast tumors are conserved across microarray platforms," *BMC Genomics*, vol. 7, p. 96, 2006.

[3] S. Katz, R. A. Irizarry, X. Lin, M. Tripputi, and M. W. Porter, "A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database," *BMC Bioinformatics*, vol. 7, p. 464, 2006.

[4] S. Yoon, Y. Yang, J. Choi, and J. Seong, "Large scale data mining approach for gene-specific standardization of microarray gene expression data," *Bioinformatics*, vol. 22, pp. 2898–2904, 2006.

[5] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnology*, vol. 14, pp. 1675–80, 1996.

[6] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms," *Nature Methods*, vol. 2, pp. 345–350, 2005.

[7] L. Gautier, M. Moller, L. Friis-Hansen, and S. Knudsen, "Alternative mapping of probes to genes for Affymetrix chips," *BMC Bioinformatics*, vol. 5, p. 111, 2004.

[8] K.-B. Hwang, S. W. Kong, S. A. Greenberg, and P. J. Park, "Combining gene expression data from different generations of oligonucleotide arrays," *BMC Bioinformatics*, vol. 5, p. 159, 2004.

[9] B. H. Mecham, D. Z. Wetmore, Z. Szallasi, Y. Sadovsky, I. Kohane, and T. J. Mariani, "Increased measurement accuracy for sequence-verified microarray probes," *Physiological Genomics*, vol. 18, pp. 308–315, 2004.

[10] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data," *Nucleic Acids Research*, vol. 33, pp. e175–, 2005.

[11] J. Zhang, R. P. Finney, R. J. Clifford, L. K. Derr, and K. H. Buetow, "Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach," *Genomics*, vol. 85, pp. 297–308, 2005.

[12] MAQC Consortium, "The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, pp. 1151–1161, 2006.

[13] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proc. Nat'l Academy of Sciences, USA*, vol. 99, pp. 14 031–14 036, 2002.

[14] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *Proc. Nat'l Academy of Sciences, USA*, vol. 98, pp. 31–36, 2001.

[15] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Research*, vol. 31, p. e15, 2003.

[16] Z. Wu and R. Irizarry, "Stochastic models inspired by hybridization theory for short oligonucleotide arrays," in *Proc. 8th Conf. Research in Computational Molecular Biology (RECOMB'04)*. New York: ACM Press, 2004, pp. 98–106.

[17] S. Hochreiter, D.-A. Clevert, and K. Obermayer, "A new summarization method for affymetrix probe level data," *Bioinformatics*, vol. 22, pp. 943–949, 2006.

[18] M. Milo, A. Fazeli, M. Niranjan, and N. Lawrence, "A probabilistic model for the extraction of expression levels from oligonucleotide arrays," *Biochemical Society Transactions*, vol. 31, pp. 1510–1512, 2003.

[19] A.-M. K. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green, "BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data," *Biostatistics*, vol. 6, pp. 349–373, 2005.

[20] X. Li, Z. He, and J. Zhou, "Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation," *Nucleic Acids Research*, vol. 33, pp. 6114–6123, 2005.

[21] R. Mei, E. Hubbell, S. Bekiranov, M. Mittmann, F. C. Christians, M.-M. Shen, G. Lu, J. Fang, W.-M. Liu, T. Ryder, P. Kaplan, D. Kulp, and T. A. Webster, "Probe selection for high-density oligonucleotide arrays," *Proc. Nat'l Academy of Sciences, USA*, vol. 100, pp. 11 237–11 242, 2003.

[22] F. Naef and M. O. Magnasco, "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays," *Physical Review E*, vol. 68, 2003.

[23] C. Wu, R. Carta, and L. Zhang, "Sequence dependence of cross-hybridization on short oligo microarrays," *Nucleic Acids Research*, vol. 33, p. e84, 2005.

[24] A. L. Oberg, D. W. Mahoney, K. V. Ballman, and T. M. Therneau, "Joint estimation of calibration and expression for high-density oligonucleotide arrays," *Bioinformatics*, vol. 22, pp. 2381–2387, 2006.

[25] L. Zhang, L. Wang, A. Ravindranathan, and M. Miles, "A new algorithm for analysis of oligonucleotide arrays: Application to expression profiling in mouse brain regions," *J. Molecular Biology*, vol. 317, pp. 225–235, 2002.

[26] L. L. Elo, L. Lahti, H. Skottman, M. Kyläniemi, R. Lahesmaa, and T. Aittokallio, "Integrating probe-level expression changes across generations of Affymetrix arrays," *Nucleic Acids Research*, vol. 33, p. e193, 2005.

[27] M. E. Ross, X. Zhou, G. Song, S. Shurtleff, K. Girtman, W. Williams, H.-C. Liu, R. Mahfouz, S. Raimondi, N. Lenny, A. Patel, and J. Downing, "Classification of pediatric acute lymphoblastic leukemia by gene expression profiling," *Blood*, vol. 102, pp. 2951–2959, 2003.

[28] A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch, "Large-scale analysis of the human and mouse transcriptomes," *Proc. Nat'l Academy of Sciences, USA*, vol. 99, pp. 4465–4470, 2002.

[29] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch, "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proc. Nat'l Academy of Sciences, USA*, vol. 101, pp. 6062–6067, 2004.

[30] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, *et al.*, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, pp. 133–143, 2002.

[31] H. Yu, F. Wang, K. Tu, L. Xie, Y. Li, and Y. Li, "Transcript-level annotation of affymetrix probesets improves the interpretation of gene expression data," *BMC Bioinformatics*, vol. 8, p. e194, 2007.

[32] H. Auer, S. Lyianarachchi, D. Newsom, M. I. Klisovic, G. Marcucci, and K. Kornacker, "Chipping away at the chip bias: RNA degradation in microarray analysis," *Nature Genetics*, vol. 35, pp. 292–293, 2003.

[33] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy–analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, pp. 307–315, 2004.

[34] C. J. Burden, Y. Pittelkow, and S. R. Wilson, "Adsorption models of hybridization and post-hybridization behavior on oligonucleotide microarrays," *J. Physics: Condens. Matter*, vol. 18, pp. 5545–5565, 2006.

[35] A. Sequeira, F. Meng, B. Rollins, R. Myers, E. Jones, S. Watson, H. Akil, A. Schatzberg, J. Barchas, W. Bunney, and V. M.P., "Coding SNPs included in exon arrays for the study of psychiatric disorders," *Molecular Psychiatry*, vol. 13, pp. 363–365, 2008.

[36] E. Sliwerska, F. Meng, T. Speed, E. Jones, W. Bunney, H. Akil, S. Watson, and M. Burmeister, "SNPs on chips: the hidden genetic code in expression arrays," *Biological Psychiatry*, vol. 61, pp. 13–16, 2007.

[37] I. Lee, A. A. Dombkowski, and B. D. Athey, "Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray," *Nucleic Acids Research*, vol. 32, pp. 681–690, 2004.

[38] L. L. Elo, M. Katajamaa, R. Lund, M. Oresic, R. Lahesmaa, and T. Aittokallio, "Improving identification of differentially expressed genes by integrative analysis of Affymetrix and Illumina arrays," *OMICS: A Journal of Integrative Biology*, vol. 10, pp. 369–380, 2006.

[39] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis (2nd edition)*. Boca Raton, FL: Chapman & Hall/CRC, 2003.

[40] D. Goldfarb, "A family of variable-metric methods derived by variational means," *Mathematics of Computation*, vol. 24, pp. 23–26, 1970.

[41] M. McGee and Z. Chen, "New spiked-in probe sets for the Affymetrix HG-U133A Latin Square experiment," *COBRA Preprint Series*, Article 5, 2006.

[42] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. L. C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, 2004.

[43] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 36, suppl. 1, pp. D13–21, 2008.

[44] W. J. Kent, "BLAT—The BLAST-Like Alignment Tool," *Genome Research*, vol. 12, pp. 656–664, 2002.

[45] E. C. Rouchka, A. W. Phatak, and A. V. Singh, "Effect of single nucleotide polymorphisms on affymetrix match-mismatch probe pairs," *Bioinformation*, vol. 2, pp. 405–411, 2008.

[46] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–264, 2003.

[47] Y. Xing, K. Kapur, and W. H. Wong, "Probe selection and expression index computation of affymetrix exon arrays," *PLoS ONE*, vol. 1, p. e88, 2006.

[48] J. Yan and T. G. Marr, "Computational analysis of 3-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat," *Genome Research*, vol. 15, pp. 369–375, 2005.

[49] A. Reverter, W. Barris, S. McWilliam, K. Byrne, Y. Wang, S. Tan, N. Hudson, and B. Dalrymple, "Validation of alternative methods of data normalization in gene co-expression studies," *Bioinformatics*, vol. 21, pp. 1112–1120, 2005.