BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 5 2013, pages 605–613
doi:10.1093/bioinformatics/bts713

Gene expression

Advance Access publication January 9, 2013

SIBER: systematic identification of bimodally expressed genes using RNAseq data

Pan Tong[1,2], Yong Chen[3], Xiao Su[3] and Kevin R. Coombes[1,*]

[1]Department of Bioinformatics and Computational Biology, The University of Texas M. D. Anderson Cancer Center, [2]Biomathematics and Biostatistics, Graduate School of Biomedical Sciences, The University of Texas Health Science Center at Houston and [3]Division of Biostatistics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Identification of bimodally expressed genes is an important task, as genes with bimodal expression play important roles in cell differentiation, signalling and disease progression. Several useful algorithms have been developed to identify bimodal genes from microarray data. Currently, no method can deal with data from next-generation sequencing, which is emerging as a replacement technology for microarrays.

Results: We present SIBER (systematic identification of bimodally expressed genes using RNAseq data) for effectively identifying bimodally expressed genes from next-generation RNAseq data. We evaluate several candidate methods for modelling RNAseq count data and compare their performance in identifying bimodal genes through both simulation and real data analysis. We show that the lognormal mixture model performs best in terms of power and robustness under various scenarios. We also compare our method with alternative approaches, including profile analysis using clustering and kurtosis (PACK) and cancer outlier profile analysis (COPA). Our method is robust, powerful, invariant to shifting and scaling, has no blind spots and has a sample-size-free interpretation.

Availability: The R package SIBER is available at the website http://bioinformatics.mdanderson.org/main/OOMPA:Overview.

Contact: kcoombes@mdanderson.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 15, 2012; revised on December 5, 2012; accepted on December 16, 2012

1 INTRODUCTION

Gene expression at the messenger RNA level frequently operates in two modes: baseline and underexpression or overexpression. Although the underlying mechanisms driving these two modes of expression (bimodal expression hereafter) may be different, [e.g. *cis*/*trans* regulation, DNA copy number change, microRNA regulation, DNA methylation, transcription factor activity and so on (Biggar and Crabtree, 2001; Chen and Widom, 2005; Louis and Becskei, 2002)], the impact of bimodal expression is critical to cell functioning. For instance, tissue-specific expression for certain genes is needed for cell differentiation; the maintenance and regulation of bimodal expression is essential to cell signalling; the change of expression mode, especially for oncogenes and tumour suppressors, can lead to uncontrolled cell proliferation and malignant cancer.

Because of its importance, systematic identification of bimodally expressed genes has received great attention, especially in cancer research. Various methods have been proposed during the past decade. Current methods can be put into two categories: non-parametric and normal mixture models. For example, cancer outlier profile analysis (COPA) is a nonparametric method to search explicitly for genes with 'outlier' expression patterns (Tomlins *et al.*, 2005). Specifically, the expression of each gene is standardized by its median and scaled with its median absolute deviation, and then ranked by a pre-specified quantile (i.e. 75, 90 or 95%) of the transformed data. They applied COPA to 132 datasets studying a variety of cancers and discovered that the fusion of ERG and ETV1 was overexpressed in 57% of prostate cancer patients (Tomlins *et al.*, 2005).

Methods in the second category postulate a mixture of normal distributions for gene expression levels. For example, the profile analysis using clustering and kurtosis (PACK) proposed by Teschendorff *et al.* (2006) first filters unimodal genes with BIC and then ranks bimodal genes using negative kurtosis. Ranking with positive kurtosis is also possible but could be dominated by outliers. This method was later applied to identify bimodally expressed genes with expression status linked to breast cancer prognosis (Teschendorff *et al.*, 2007). Another method, proposed by Ertel and Tozeren (2008), is based on the likelihood ratio test (LRT) and uses a χ^2 distribution with six degrees of freedom as the null distribution to obtain *P*-values of LRT statistics. This method was applied to identify bimodal genes from diverse tissue types in mouse. More recently, Wang *et al.* (2009) proposed a new definition of a bimodality index (BI). This method calculates the BI for each gene from the estimated parameters of a two-component normal mixture model and then ranks the genes by BI. One attractive feature of this method is that BI has an intuitive interpretation, as it is derived from a sample size calculation. The BI method was later applied to show that melanoma antigen family A defined a subset of triple negative breast cancers that might benefit from immune augmentation (Karn *et al.*, 2012).

Despite their success, all of these methods (with the exception of COPA) are designed for microarray data under the assumption of normal distributions. With next-generation sequencing (NGS) emerging as a replacement technology for microarrays,

*To whom correspondence should be addressed.

it is important to develop valid methods for identifying bimodally expressed genes from NGS data. RNAseq data, unlike microarray data, consists of read counts mapped to each gene. In practice, some investigators may treat the RNAseq data the same as microarray data after some transformation (e.g. log or Box–Cox), but the validity of such an approach has never been investigated. Special attention may be required because of the discrete nature of RNAseq data. Current methods for analysing RNAseq data for differential expression rely on generalized Poisson or negative binomial distributions rather than normal distributions. In this article, we propose a class of generalized bimodality indices in the framework of mixture models, including mixtures of negative binomial, generalized Poisson or log normal distributions. We also formally investigate the performance of naïve methods using COPA and PACK, which treat RNAseq data as equivalent to microarray data after some transformation.

The remainder of this article is organized as follows: in Section 2, we describe several discrete models for dealing with RNAseq data and define the generalized BI that can be used for various mixture modelling. We also discuss RNAseq library normalization in the context of mixture modelling. We present simulation results in Section 3.1, where three different models are evaluated and compared with alternative approaches, including COPA and PACK. In Section 3.2, we analyse data from The Cancer Genome Atlas (TCGA) that further confirms the effectiveness of the generalized BI in identifying bimodal genes from RNAseq data.

2 METHODS

We propose a two-step procedure to identify bimodally expressed genes. The first step is to fit a two-component mixture model. Specifically, three candidate mixture models are considered. Two of these models explicitly account for the discrete nature of the RNAseq data, whereas the third model treats the data as continuous after some transformation. The second step is to calculate the BI corresponding to the assumed mixture distribution.

2.1 Mixture models for RNAseq count data

We model the observed raw counts using a two-component mixture model, as in Wang *et al.* (2009). Denote the raw count for gene g in sample s by $C_{g,s}$ and the true expression by $\mu_{g,c(s)}$ depending on the component (or cluster) membership $c(s)$ that sample s belongs to. Here, $c(s) = k$ (for $k = 1, 2$) means that sample s belongs to component k with mean expression $\mu_{g,k}$. To avoid model non-identifiability, we require $\mu_{g,1} \leq \mu_{g,2}$. As each gene is studied separately, we may suppress the index g for simplicity of notations. We consider three different mixture models.

Negative binomial mixture: Our first model is motivated by the negative binomial (NB) distribution, which is widely used to model RNAseq data in differential gene expression analysis (Anders and Huber, 2010; Di *et al.*, 2011; Hardcastle and Kelly, 2010; Robinson and Smyth, 2007). Of note, we prefer NB rather than the Poisson distribution to account for the overdispersion observed in RNAseq data. Specifically, the probability of observing count C_s can be formulated as:

$$\Pr(C_s) = \pi f_{\mathrm{NB}}(C_s; \mu_1, \phi) + (1-\pi) f_{\mathrm{NB}}(C_s; \mu_2, \phi) \quad (1)$$

where $f_{\mathrm{NB}}(\cdot; \mu, \phi)$ is the probability mass function for the NB distribution with mean μ and dispersion ϕ (variance $= \mu + \phi\mu^2$):

$$f_{\mathrm{NB}}(y; \mu, \phi) = \frac{\Gamma(\frac{1}{\phi} + y)}{\Gamma(y+1)\Gamma(\frac{1}{\phi})} \left(\frac{1}{\phi\mu + 1}\right)^{\frac{1}{\phi}} \left(1 - \frac{1}{\phi\mu + 1}\right)^{y},$$

and μ_1 and μ_2 are the true expression levels for the two components. The parameter ϕ affects the within-group variability. Note that we assume equal dispersion in the two distributions (we do not assume equal variance because, for NB distribution, the variance depends on the mean. Assuming equal variance would impose an undesirable constraint on the component means), similar to the tagwise dispersion mode in EdgeR (Robinson and Smyth, 2007; Robinson *et al.*, 2010). When the dispersion parameter $\phi = 0$, Equation (1) reduces to a mixture of Poisson distributions. The parameters $(\pi, \mu_1, \mu_2, \phi)$ can be estimated by maximizing the likelihood function:

$$\mathrm{L}(\pi, \mu_1, \mu_2, \phi | C_s) = \prod_{s=1}^{n} \{\pi f_{\mathrm{NB}}(C_s; \mu_1, \phi) + (1-\pi) f_{\mathrm{NB}}(C_s; \mu_2, \phi)\}$$

The expectation-maximization (EM) algorithm or direct optimization can be used for this purpose.

Generalized Poisson mixture: The generalized Poisson (GP) distribution is another model used to describe RNAseq count data (Srivastava and Chen, 2010). Under the two-component mixture framework, it can be formulated similarly as:

$$\Pr(C_s) = \pi f_{\mathrm{GP}}(C_s; \mu_1, \phi) + (1-\pi) f_{\mathrm{GP}}(C_s; \mu_2, \phi) \quad (2)$$

where $f_{\mathrm{GP}}(\cdot; \mu, \phi)$ is the probability density function for the Generalized Poisson distribution with mean μ and dispersion ϕ (variance $= \phi\mu$)

$$f_{\mathrm{GP}}(y; \mu, \phi) = \frac{\mu}{\sqrt{\phi}} \{\frac{\mu}{\sqrt{\phi}} + (1 - \frac{1}{\sqrt{\phi}})y\}^{y-1} exp\{-\frac{\mu}{\sqrt{\phi}} - (1 - \frac{1}{\sqrt{\phi}})y\}/y!,$$

and μ_1 and μ_2 are the true expression levels for the two components. For similar reasons to the NB model, we assume equal dispersion between the two components. Note that the variance of the GP distribution is a linear function of its mean, whereas the variance of the NB distribution is a quadratic function of the mean. When $\phi = 1$, the GP distribution reduces to Poisson. As a result, a mixture of Poisson distribution is automatically included in the GP model.

Normal mixture with Box–Cox transformation: Instead of accounting for the discrete nature of the RNAseq data as in models (1) and (2), we could treat the data as normal after some transformation. A wide class of transformations were proposed by Box and Cox (1964), known as the Box–Cox transformation or power transformation,

$$y_i^{(\lambda)} = \begin{cases} (y_i^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

As suggested by the data empirically, the optimal choice of λ for RNAseq data considered in this article is $\lambda = 0$, which corresponds to the log-transformation (details in Section 3.2.1). This leads to our third model as a mixture of lognormal (LN) distributions,

$$\Pr(C_s) = \pi f_{\mathrm{LN}}(C_s; \mu_1, \sigma^2) + (1-\pi) f_{\mathrm{LN}}(C_s; \mu_2, \sigma^2) \quad (3)$$

where $f_{\mathrm{LN}}(\cdot; \mu, \sigma^2)$ is the probability density function for LN distribution with mean μ and variance σ^2 at the log scale. The variances of the two log-transformed distributions are assumed to be equal, similar to the mixture of normals considered in Wang *et al.* (2009). We note that the log transformation has been used previously to analyse RNAseq experiments (Cloonan *et al.*, 2008; Lee *et al.*, 2011; McIntyre *et al.*, 2011).

2.2 Generalized bimodality index

The BI (Wang *et al.*, 2009) is defined as:

$$\mathrm{BI} = \sqrt{\pi(1-\pi)} \frac{|\mu_1 - \mu_2|}{\sigma} \quad (4)$$

where $\delta = |\mu_1 - \mu_2|/\sigma$ is the effect size that measures the distance between the two-components. The coefficient $\sqrt{\pi(1-\pi)}$ is maximized at $\pi = 0.5$; hence, it penalizes unbalanced allocation into the two components. A limitation of the original BI is that it is defined based on a

normal mixture with equal variance. It does not apply to normal mixtures with unequal variance or to genes whose expressions do not follow normal distributions (e.g. discrete distributions as in RNAseq data). To deal with these situations, here we generalize the original BI.

The definition of BI in Wang *et al.* (2009) was motivated by sample size considerations. For a normal mixture with unequal variances, similar calculations tell us:

$$\mathrm{BI}^2 = \frac{\pi(1-\pi)(\mu_1-\mu_2)^2}{(1-\pi)\sigma_1^2+\pi\sigma_2^2} = \frac{(Z_{\alpha/2}+Z_\beta)^2}{\mathrm{N}} \tag{5}$$

Hence, the generalized BI can be calculated by:

$$\mathrm{BI} = \sqrt{\pi(1-\pi)}\frac{|\mu_1-\mu_2|}{\sqrt{(1-\pi)\sigma_1^2+\pi\sigma_2^2}} \tag{6}$$

Formula (6) is similar to formula (4) except the effect size is modified as: $\delta = |\mu_1-\mu_2|/\sqrt{(1-\pi)\sigma_1^2+\pi\sigma_2^2}$. Note that when $\sigma_1=\sigma_2$, the generalized BI formula reduces to (4).

For other mixture models, deriving the exact BI formula directly from sample size considerations is a tedious task, and the resulting BI may be complicated. In general, there is no closed form for BI. However, we can instead obtain a formula for BI under large sample approximation. It turns out that, even for a mixture of discrete distributions, we can obtain the same BI formula as in (6) by using the central limit theorem. Details are provided in Supplementary Section S2.

The generalized BI for normal mixtures with unequal variance has a sample size interpretation, as indicated in formula (5) where α and β are the type I and type II error and N is the sample size. Z_α determines the quantile for a standard normal distribution that has right tail probability being α. In a typical microarray or RNAseq experiment, the sample size N is predetermined, and BI can be computed for each gene. Different values of BI then represent different type I and type II error. To reliably detect useful bimodal genes, BI needs to exceed a certain threshold that is determined by N.

2.3 Adjusting for library size and gene length effect

Similar to microarray data, RNAseq data require proper normalization to make meaningful comparisons between samples. A common practice is to scale the raw counts by both the gene length L_g for gene g and the total reads T_s in sample s, giving the so-called RPKM value (Mortazavi *et al.*, 2008). For this reason, we introduce a normalization term, $d_{g,s}$, into our mixture models. This term accounts for technical effects, including lane, flow-cell and library preparation effects. In the case of RPKM, $d_{g,s} = L_g T_s$. Directly scaling the count data by $d_{g,s}$ would transform the data onto continuous scale that cannot be modelled by NB or GP distributions. Instead, we incorporate $d_{g,s}$ through the expected count as:

$$\mathrm{E}[C_{g,s}|d_{g,s}, c(s)] = d_{g,s}\mu_{g,c(s)}$$

Hence, we only need to replace the component distribution $f(C_s; \mu_{c(s)}, \phi)$ in Section 2.1 with $f(C_s; d_{g,s}\mu_{c(s)}, \phi)$. The rest of the inference remains the same.

As pointed out by Bullard *et al.* (2010), RPKM normalization performs poorly when there are highly differentially expressed genes. More robust normalization methods such as TMM (Robinson and Oshlack, 2010) and the method used in DESeq or DEXseq can be applied (Anders and Huber, 2010; Anders *et al.*, 2012). Inclusion of such normalization methods to our models is similar, only adding a scaling factor to the component means.

3 RESULTS

We let NB, GP and LN denote the three models described earlier in the text. For each model, let $\mathrm{BI}_{\mathrm{NB}}$, $\mathrm{BI}_{\mathrm{GB}}$ and $\mathrm{BI}_{\mathrm{LN}}$ be the generalized BI computed with respect to that model. The fundamental question to be addressed is which model yields more robust and more reliable identification of bimodal genes from RNAseq data. To evaluate the performance of different models, and to compare with alternative methods, we first conduct simulation studies. We generate artificial RNAseq data from one of the three models. Regardless of which model is used to generate the data, we compute BI using all three models. This procedure allows us to evaluate the performance of BI under misspecified models; this step is important because the true underlying model for RNAseq data is often unknown in practice.

In the second scenario, we look at TCGA data, where both microarray and RNAseq data are available for the same set of breast cancer samples. We establish the 'true' bimodal status for a subset of genes by applying the existing methods to the microarray data, then manually confirming the results by visually inspecting the distributions. For this subset of genes, the misclassification rates (of genes as bimodal or unimodal) are expected to be low. We then compute BI from the RNAseq data using all three models. Because there is a good correspondence between microarray and RNAseq data (Marioni *et al.*, 2008), we can evaluate the performance of the BI models by constructing receiver operating characteristic (ROC) curves that test their ability to correctly match the microarray-based gene classifications.

3.1 Simulation study

In this subsection, we consider RNAseq data generated from one of the three mixture models, which will be referred to as NB, GP and LN datasets, respectively.

3.1.1 NB, GP and LN datasets For each of the NB, GP and LN datasets, we simulate both bimodal and unimodal genes, which are generated from two-component and one-component mixture models, respectively. To cover a spectrum of settings in practice, we allow the mixture proportion parameter π, the effect size δ and the sample size to vary. As we know the true status of the generated gene data, we can construct ROC curves that evaluate the ability of the BI models to correctly predict the true status. The performance of $\mathrm{BI}_{\mathrm{NB}}$, $\mathrm{BI}_{\mathrm{GP}}$ and $\mathrm{BI}_{\mathrm{LN}}$ will be evaluated using the area under the corresponding ROC curves (AUC).

Bimodal genes: For the bimodally expressed genes, we choose different combinations of parameters to represent a wide range of bimodal shapes. Specifically, π takes values between 0.1 and 0.5 with a step of 0.1 ($\pi = 0.6, \ldots, 0.9$ are omitted by symmetry). In practice, $\pi = 0.1$ or 0.2 leads to an unbalanced mixture, whereas $\pi = 0.3, 0.4$ or 0.5 leads to more balanced mixture distribution. We also use a range of effect sizes, $\delta = 2.5, 3, 3.5, 4$. To simulate genes that have different expression levels, we set $\mu_1 = 5$ for LN model (corresponding mean at exponential scale is 244.7), $\mu_1 = 100, 1000, 5000, 10000$ for NB model and $\mu_1 = 100, 1000, 2000, 4000$ for GP model. For LN, we set $\sigma = 1$ because of the equal variance assumption (corresponding variance at exponential scale is 34.5). We assume equal dispersion between the two groups for both NB and GP models. As a result, we set the dispersion parameter $\phi = 0.1$ for NB model and $\phi = 0.5\mu_1$ for GP model. This implies that in both NB and GP

models, the variance is a quadratic function of the mean, as typically seen in RNAseq experiments (see mean-variance relationship in Supplementary Figure S15). We use Equation (6) to solve for $\mu_2 = \mu_1 + \delta\sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}$. All possible combinations of π, μ_1 and δ are considered in each model, which results in 20 ($5 \times 1 \times 4$) settings in LN datasets and 80 ($5 \times 4 \times 4$) settings in NB and GP datasets. The parameters were chosen to mimic real data. For each setting, we simulate 100 genes, which lead to 2000 bimodal genes in LN datasets and 8000 genes in NB or GP datasets. We choose four different sample size settings ($N = 50$, 100, 200 and 300) for each dataset. Data generated from the LN model is continuous, but is rounded to the nearest integer.

Unimodal genes: The unimodal genes are simulated from a one-component model. To match the parameter settings for the bimodal genes, we generate unimodal genes from the larger component corresponding to the mixture proportion >0.5. Equal numbers of unimodal and bimodal genes are simulated and combined to form LN, NB and GP datasets.

3.1.2 Effect of the mixture proportion To examine the effect of the parameter π that describes the proportion of samples in the smaller group, we prepared box-and-whisker plots of BI as a function of π (Fig. 1). For each of the three datasets (for $N = 300$), we fit three mixture models to obtain parameter estimates and calculate BI. The distributions when the data come from a unimodal distribution are included at $\pi = 0$. Ideally, we expect BI to increase as π increases from 0 to 0.5. We see this behaviour when we analyse each dataset using the same model that was used to simulate it. The NB model, however, exhibits different behaviour when the model is misspecified; BI values for $\pi = 0.1$ are lower than in the unimodal case and peak when $\pi = 0.2$. The GP model behaves well on the GP and NB datasets, but behaves poorly when the true model is LN. The LN model, by contrast, has similar performance regardless of which model was used to simulate the data.

3.1.3 Performance evaluation metrics A useful measure to evaluate the performance of BI is the ROC curve. Figure 2 and Supplementary Figure S1 plot the ROC curves for the three BIs, namely, BI_{NB}, BI_{GP} and BI_{LN}, in NB, GP and LN datasets, respectively, for different sample sizes. Specifically, Supplementary Figure S1 shows the performance when the assumed distribution is correct, whereas Figure 2 shows the performance under misspecified models. Note that the evaluation using AUC is limited, as it puts the performance under different false-positive (FP) rates on an equal footing. Hence, to explicitly control FP, we also evaluate the performance by looking at the power under predefined FP rates in Tables 1–3.

3.1.4 Performance under correctly specified models Supplementary Figure S1 shows that when the model is correctly specified, the three methods perform similarly in terms of AUC. Even when the sample size $N = 50$, the ROC curve is still satisfactory. For FP rate and power under the correctly specified model, details are given in column BI_{NB} in Table 1, column BI_{GP} in Table 2 and column BI_{LN} in Table 3. Note that the largest power in each row is bolded. We see that under the correctly specified model, BI_{NB}, BI_{GP} and BI_{LN} all perform reasonably well. In all three models, increased sample size improves power by reducing the cut-off of BI when controlling the same type I error. For the same sample size and type I error, the LN model has slightly larger power (in average 1.9% larger) than the NB model. Both the LN and the NB model perform significantly better than the GP model (the power of NB is 8.9% larger than GP) when FP < 0.1. Nevertheless, the GP model has larger power at larger FP, which makes its AUC better than the NB model and almost matches that of the LN model (Supplementary Figure S2). There is a minor decline of AUC in the NB and GP models. This happens because the BI formula for the NB and GP models relies on large sample approximation and hence loses some efficiency. Note that the FP rate and power under different sample sizes for the LN model agrees with the results in Wang *et al.* (2009).

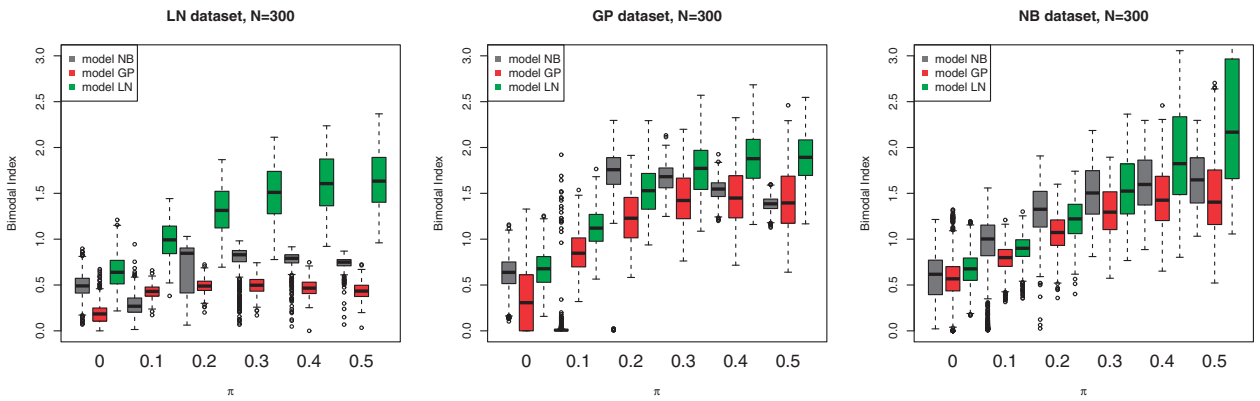

Fig. 1. BI as a function of the size (π) of the smaller group. For each of the three models, we simulated datasets with a range of different distances ($\delta = 2.5, 3, 3.5, 4$) and applied all three models to compute BI. Boxplots for $\pi = 0$ give the distribution of BI when the data are simulated from a unimodal distribution. Performance under the correctly specified model is similar for all three methods, with equal splits ($\pi = 0.5$) yielding the largest BI values. The NB model (grey) performs extremely poorly under misspecified models, with BI values for $\pi = 0.1$ clearly less than the unimodal BI values and peak BI when $\pi = 0.2$. The GP model (red) performs poorly on data simulated from the LN model. The LN model (green) performs consistently regardless of how the data are simulated

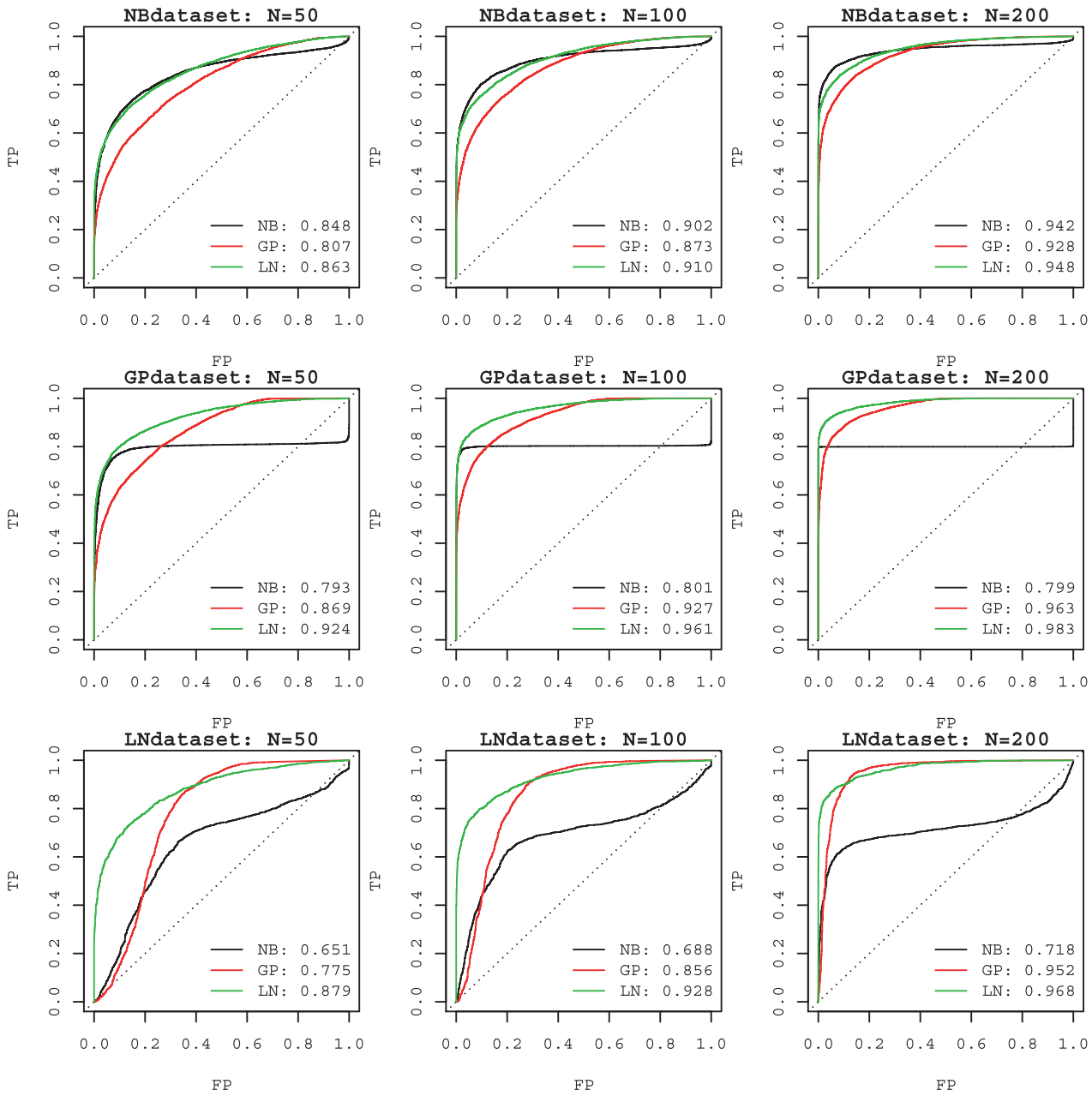

Fig. 2. Robustness of NB, GP and LN models. ROC curves for the three mixture models fitted on NB, GP and LN datasets are compared under sample size $N = 50$, 100, and 200 ($N = 300$ is omitted because of space limitations). Various bimodal shapes as characterized by different distances ($\delta = 2.5, 3, 3.5, 4$) and component size ($\pi = 0.1, 0.2, 0.3, 0.4, 0.5$) are simulated to mimic real data. The LN model is most robust and provides satisfactory performance even when the model is misspecified

3.1.5 Performance under misspecified models In practice, the true underlying model is often unknown. It is important to investigate the robustness of the proposed BIs under a misspecified model. Here, we compare the performance of the three models and study the effect of model misspecification. Figure 2 shows the ROC curves under misspecified models. We see that the performance varies, suggesting different robustness among the three models.

For the NB datasets, where the true model generating the simulated data is the NB model, Table 1 shows the smallest BI needed to achieve a given type I error and corresponding power to detect bimodal genes for the three mixture models with $N = 50$, 100, 200 or 300. BI_{LN} provides competitive performance compared with the true model, whereas BI_{GP} performs much worse. For the GP datasets, BI_{LN} even dominates BI_{GP} at all FP rate and sample sizes listed in Table 2 (however, not at all possible FP rates).

For the LN datasets, the power of BI_{NB} and BI_{GP} stemming from misspecified models is much lower than that of BI_{LN} under the same FP rate (Table 3). More importantly, when the sample

Table 1. Performance on NB datasets

FP	N	BI_{NB}		BI_{GP}		BI_{LN}	
		BI cut-off	Power	BI cut-off	Power	BI cut-off	Power
0.01	50	1.605	0.395	1.540	0.264	1.608	**0.438**
	100	1.347	**0.590**	1.312	0.384	1.325	0.589
	200	1.144	**0.772**	1.123	0.549	1.145	0.707
	300	1.054	**0.842**	1.006	0.666	1.042	0.771
0.05	50	1.410	0.582	1.344	0.418	1.394	**0.584**
	100	1.199	**0.732**	1.138	0.556	1.181	0.698
	200	1.030	**0.856**	0.966	0.706	1.026	0.798
	300	0.946	**0.901**	0.880	0.788	0.948	0.851
0.10	50	1.301	**0.678**	1.226	0.524	1.287	0.659
	100	1.111	**0.802**	1.039	0.652	1.100	0.757
	200	0.958	**0.891**	0.886	0.786	0.958	0.850
	300	0.882	**0.924**	0.814	0.854	0.894	0.890

Table 2. Performance on GP datasets

FP	N	BI_{NB}		BI_{GP}		BI_{LN}	
		BI cut-off	Power	BI cut-off	Power	BI cut-off	Power
0.01	50	1.521	0.510	1.515	0.336	1.604	**0.586**
	100	1.275	**0.759**	1.277	0.520	1.356	**0.759**
	200	1.093	0.799	1.154	0.632	1.157	**0.869**
	300	0.992	0.799	1.031	0.742	1.068	**0.915**
0.05	50	1.323	0.714	1.296	0.521	1.406	**0.732**
	100	1.126	0.796	1.102	0.688	1.205	**0.846**
	200	0.969	0.799	0.929	0.829	1.047	**0.920**
	300	0.898	0.799	0.839	0.877	0.970	**0.952**
0.10	50	1.223	0.770	1.183	0.628	1.298	**0.799**
	100	1.042	0.800	1.001	0.775	1.122	**0.887**
	200	0.903	0.799	0.841	0.883	0.979	**0.945**
	300	0.844	0.799	0.764	0.918	0.914	**0.967**

size is small or moderate, the power of BI_{NB} and BI_{GP} is even smaller than the FP rate or half of the power achieved by BI_{LN} at best. When the sample size is relatively large, i.e. $N = 200$, the performance of BI_{GP} improves and almost matches BI_{LN} (AUC: 0.95 versus 0.97, Fig. 2). However, increasing sample size only improves the power of BI_{NB} at low FP rate while decreasing the power at high FP rate. Overall, the AUC of BI_{NB} only increases slightly with sample size. The reason is that the fitted NB model fails to detect most bimodal genes with $\pi = 0.1$ in the GP and LN datasets (Fig. 1). These results suggest that BI_{NB} and BI_{GP} are highly sensitive to model misspecification. Hence, from the spectrum of settings considered, BI_{LN} outperforms the other two methods in terms of power under the correctly specified model as well as robustness under a misspecified model.

3.1.6 Difficulty in identifying the true model In general, it is desirable to identify the true underlying model (e.g. NB, GP, LN or other models). However, this task is extremely challenging (and perhaps impossible) in practice. For example, when BIC is used as the criterion for model selection, the BICs from NB, GP and LN models are almost indistinguishable for all three simulated datasets (Supplementary Figure S4). Compared with misspecified models, the true model does not show a clear advantage in terms of BIC. In this sense, each of the three models provides similar fits for the data, despite the fact that they have different performance in terms of identifying bimodal genes. This finding suggests that robustness of BI is important because of the practical difficulty in identifying the true model.

Table 3. Performance on LN datasets

FP	N	BI_{NB}		BI_{GP}		BI_{LN}	
		BI cut-off	Power	BI cut-off	Power	BI cut-off	Power
0.01	50	1.218	0.007	1.100	0.005	1.561	**0.410**
	100	0.976	0.052	0.818	0.008	1.304	**0.613**
	200	0.820	0.322	0.574	0.136	1.109	**0.790**
	300	0.763	0.482	0.459	0.520	1.034	**0.856**
0.05	50	1.014	0.086	0.840	0.049	1.367	**0.592**
	100	0.862	0.263	0.628	0.143	1.160	**0.750**
	200	0.737	0.568	0.409	0.734	1.005	**0.868**
	300	0.689	0.628	0.360	0.871	0.932	**0.914**
0.10	50	0.931	0.197	0.700	0.152	1.265	**0.688**
	100	0.797	0.436	0.511	0.430	1.094	**0.800**
	200	0.693	0.634	0.354	0.886	0.954	**0.899**
	300	0.653	0.650	0.308	**0.956**	0.878	0.938

3.1.7 Robustness to outlier data In practice, microarray and RNAseq data often contain outliers because of various technical artifacts such as library preparation and amplification bias as well as biological variations that make the expression (RNAseq data after log transformation) deviate from the assumed normal distribution. Ignoring these outliers might lead to FP calls. Therefore, in addition to examine the robustness to model misspecification, we also examine the robustness to outlier data. We consider two kinds of outlier data, namely, data of heavy tailed distribution such as t distributions and data containing extreme values. The detailed summary of our investigation is in Supplementary Section S3. Extensive simulation studies suggest that BI is robust to both heavy tailed distributions and extreme values (Supplementary Figures S5 and S6).

3.1.8 Comparison with alternative approaches Although there are no existing methods specifically designed to identify bimodal genes in RNAseq data, it is still meaningful to compare the performance of BI with naïve methods that treat the RNAseq data as similar to microarray data after some transformation ['log(data+1)']. To this end, we compare BI_{LN} with PACK and COPA (full details are provided in Supplementary Section S4). When there are no outliers, Supplementary Figure S9 shows the performance of PACK is better than BI_{LN} or COPA in most cases. However, PACK has difficulty detecting bimodal genes with 20–80% or 30–70% split, as the kurtosis values in these cases are near zero (Supplementary Figure S10). The reason

PACK still achieves a good ROC curve is mostly attributable to the model selection step. When the data contains outliers, the performance of BI and COPA is more robust than PACK (Supplementary Figures S13 and S14). The reason is that model selection by BIC would flag most unimodal genes with outliers as bimodal candidates, which make it difficult for PACK to classify them correctly. In fact, BIC would claim that ~40% of the genes are bimodal candidates in the breast cancer data in the section. Supplementary Figure S11 shows that COPA fails to detect bimodal genes with 50–50% or 10–90% split at the chosen 10% quantile. We have to mention that COPA has a tuning parameter, which is the quantile used to rank the genes. If this parameter changes, it is possible to identify a different set of bimodal genes (Supplementary Figure S12). However, the downside of using different quantiles is that it is difficult to obtain a consensus ranking of the genes as well as evaluate the FP rate. Based on our simulation studies, we recommend the use of BI in practice for its ability of detecting a wide variety of bimodal genes, having no blind spots and being robust to outliers.

3.2 Real data analysis

We applied our methods to the TCGA Breast Cancer Dataset (BRCA) that contains 341 breast cancer samples for which both microarray and RNAseq data are available. The microarray data can serve as a reference to the RNAseq data in detecting bimodal genes.

3.2.1 LN model fits best for RNAseq data To examine which of the three models is most appropriate for real RNAseq data (and to identify the optimal λ in the Box–Cox power transformation), we need to identify reliable bimodal and unimodal genes in this dataset with high fidelity. For this purpose, we use the microarray data to guide our search. As genes with null expression are usually beyond the detection limit of microarray technology that may mislead the training set, we only looked at genes with mean expression >1.5 (Supplementary Figure S16). We then selected 142 candidate unimodal genes with BI <0.5, which ensures that there is no apparent bimodality. For candidate bimodal genes, we used BI >1.2 as minimum requirement and found 181 candidates. All these genes passed manual examination. The complete curated gene list is provided in Supplementary Table S3.

Figure 3a shows an example gene where we used profile likelihood to identify the optimal transformation indexed by λ. Figure 3b shows the histogram of optimal λ for all genes. Figure 3c shows that the optimal λ for the candidate unimodal genes is concentrated at 0, suggesting that a log-transformation is optimal. Figure 3d shows that the LN model recovers almost all bimodal and unimodal genes in the curated dataset, whereas the performance of the NB and GP models is limited. This suggests that the LN model (with log transformation of the normalized counts) provides a better fit of real RNAseq data for the purpose of identifying bimodal genes.

3.2.2 Bimodal genes identified using RNAseq data Figure 4a shows the distribution of the mixture parameters (π and δ) in the BRCA RNAseq data after fitting the LN model. The red curve is the contour where FDR $=0.01$ (BI $=1.093$); genes identified as bimodal by the LN model are above this curve and circled in

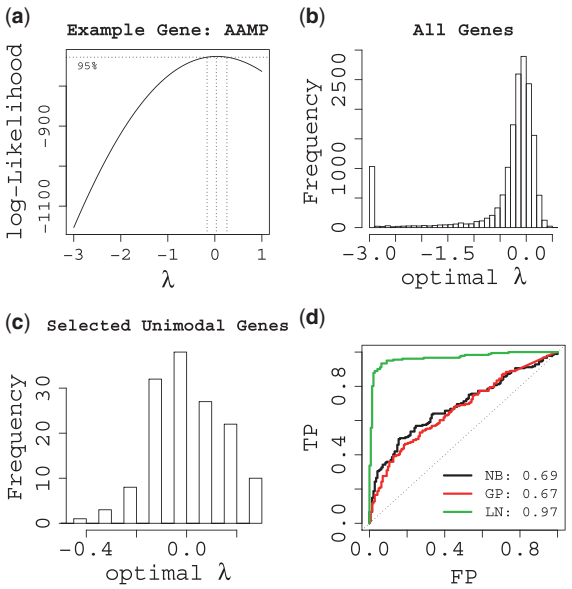

Fig. 3. RNAseq data best fit by LN model. (**a**) An example showing log transformation ($\lambda = 0$) is identified as optimal by profile likelihood. Vertical dash line indicates 95% confidence interval for optimal λ. (**b**) Histogram of optimal λ for all genes in RNAseq data. λ is concentrated at 0, suggesting log transformation is optimal for the majority of genes. (**c**) Histogram of optimal λ for the unimodal genes from curated dataset. Log transformation is optimal for all these curated unimodal genes. (**d**) ROC curve for LN, NB and GP models fitted on RNAseq data for manually curated unimodal and bimodal genes. The performance of LN model dominates NB and GP models, suggesting the data are fitted best by LN model

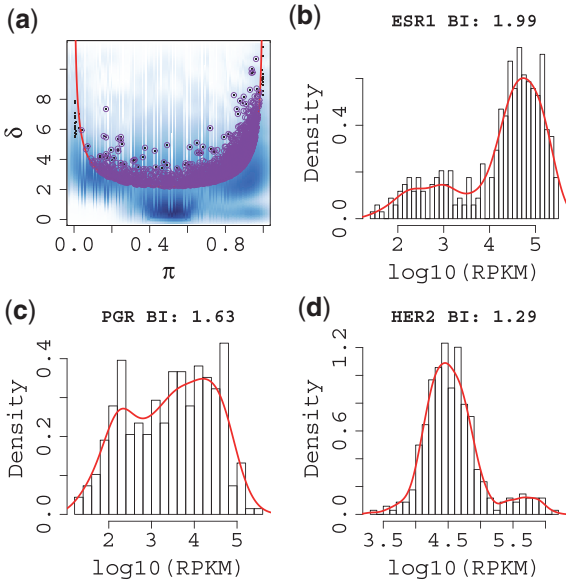

Fig. 4. Example bimodal genes. (**a**) Genes identified by LN model in BRCA data. Genes with bimodal expression are circled in purple under FDR $=0.01$ (corresponding BI $=1.093$). π is the size of first component; δ defines the distance between the two components as in (6). (**b–d**) Know breast cancer bimodal genes for ESR1, PGR and HER2

purple. We present the distributions of log-transformed count data for three genes known to be bimodally expressed in breast cancer (Fig. 4b–d). Supplementary Table S1 shows the number of genes identified at different BI cut-off, with FDR obtained through simulation. A complete list of BI values for all genes is listed in Supplementary Table S2.

4 DISCUSSION

In this article, we have proposed and compared three mixture models for systematic identification of bimodally expressed genes. All models are designed to deal explicitly with the discrete nature of RNAseq data. Our method follows the ideas in Wang *et al.* (2009) by first modelling the data through a two-component mixture coupled with computing the BI from the mixture fit that prioritizes bimodally expressed genes, which have a large distance between the two modes and sufficient samples in each mode. We extended the previous definition of BI, so that it can be applied to a mixture of unequal variance and mixture of discrete distributions.

Our definition of BI was motivated by sample size calculations. The resulting formula is intuitive and easily interpretable: BI rewards balanced allocation of the samples into two components (as it is largest when $\pi = 0.5$) and large distance between the two modes as defined by δ. The derivation through sample size calculations paved the road for computing BI under other mixture distributions or a mixture with unequal variance. Here, we proved that the generalized BI applies to mixture of discrete distributions, such as negative binomial and generalized Poisson. In such cases, the BI formula is an approximation to the exact formula, which does not have a closed form. Our simulation shows there is only minimum power loss in this approximation. For other mixtures, such as a mixture of t-distributions, the BI formula still applies. It is also possible to derive a non-parametric version of BI as done in (Al-watban and Yang, 2012). In all cases, the resulting BI formula is invariant under shifting and scaling, which is an appealing feature.

Among the three models proposed, the LN model performs best in both simulation and real data analysis. Under the correctly specified model, all three models perform similarly. However, both the NB and GP models are highly sensitive to model misspecification. In comparison, the LN model is robust. Our analysis of BRCA RNAseq data further demonstrated the superior performance of the LN model. Applying the method of Box–Cox showed that a log transformation is optimal for the majority of genes. For manually curated unimodal genes, the optimal transformation was always log transformation. In terms of recovering unimodal and bimodal genes that had been manually selected based on available microarray data, the LN model provided much better performance (AUC: 0.964 versus 0.686 and 0.666). At first, it may be a little surprising that such a simple transformation performs so well. In practice, however, all microarray data are already log transformed before any formal analysis. Inherently, this might be because of the nature of mRNA abundance level in the cell.

It remains an open question whether the lognormal model also performs well in other tasks such as differential expression analysis for RNAseq data. Comparisons with date of methods for differential expression in RNAseq have focused on the negative binomial and generalized Poisson models (Bullard *et al.*, 2010; Kvam *et al.*, 2012). The nature of the two tasks differs; identifying bimodally expressed genes is unsupervised, whereas differential expression analysis needs to know the treatment condition and hence is supervised. So, the results obtained here for bimodality do not generalize directly to differential expression analysis. The role of normalization complicates matters; in our LN analysis, we still normalized the count data using the TMM mode. Further study is required.

For parameter estimation, both the EM algorithm and Markov-chain Monte Carlo (MCMC) methods can be applied. From previous work, the results from EM and MCMC are known to be similar (Wang *et al.*, 2009). The advantage of the EM algorithm is its computational efficiency. This is extremely useful when there are thousands of genes to be evaluated. In comparison, MCMC usually takes a much longer time to converge. However, MCMC provides a posterior distribution rather than just a point estimate for the parameters. Because similar results were obtained by both estimation methods, we focused here on the EM implementation. For the LN mixture or normal mixture, the MCLUST package can be used with core functions written in Fortran (Fraley and Raftery, 2002). Our SIBER package also provides parallel computing capability to further boost the computation speed.

Although several approaches have been proposed for microarray data, it is not easy to apply them to RNAseq data. For example, LRT relies on simulation to obtain the approximate χ^2 distribution. The theoretical null distribution is not available, as the parameter is located on the boundary under the null hypothesis, which violates the regularity conditions of the LRT. The most serious problem with BIC is oversensitivity (Wang *et al.*, 2009); it gives too many candidate bimodal genes, with a high-FP rate. Moreover, BIC is not invariant under shifting, scaling or changes in sample size. This makes it impossible to rank bimodal genes based on BIC. After log transformation, both PACK and COPA can be applied. PACK performs well when the data satisfy the normality assumption. The performance of PACK suffers from both heavy tail distribution and extreme values. COPA instead is robust. However, at a given quantile, COPA can detect a limited set of bimodal genes and thus have blind spots. Therefore, it is necessary to fine tune this parameter when using COPA. Another comparison of these methods and correlation analysis with survival time can be found in Hellwig *et al.* (2010). Our approach compares favourably with existing approaches, and it is the first method that can be applied successfully to both microarray and RNAseq data.

Funding: National Institutes of Health/National Cancer Institute [U24 CA143883 (UT-MD Anderson TCGA Genome Data Analysis Center) and P30 CA016672]; Startup fund (in part) and PRIME award from the University of Texas School of Public Health (to Y.C.).

Conflict of Interest: none declared.

REFERENCES

Al-watban,A. and Yang,Z.R. (2012) Bimodal Gene Prediction Via Gap Maximisation. In: *Proceedings of 2012 International Conference on*

Bioinformatics and Computational Biology. Las Vegas, NV, USA from July 16–19.
Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-Seq data. Genome Res., **22**, 2008–2017.
Biggar,S. and Crabtree,G. (2001) Cell signaling can direct either binary or graded transcriptional responses. *EMBO J.*, **20**, 3167.
Box,G. and Cox,D. (1964) An analysis of transformations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **26**, 211–252.
Bullard,J. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
Chen,L. and Widom,J. (2005) Mechanism of transcriptional silencing in yeast. *Cell*, **120**, 37–48.
Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
Di,Y. *et al.* (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.*, **10**, 24.
Ertel,A. and Tozeren,A. (2008) Switch-like genes populate cell communication pathways and are enriched for extracellular proteins. *BMC Genomics*, **9**, 3.
Fraley,C. and Raftery,A. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.
Hardcastle,T. and Kelly,K. (2010) BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
Hellwig,B. *et al.* (2010) Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics*, **11**, 276.
Karn,T. *et al.* (2012) Melanoma antigen family A identified by the bimodality index defines a subset of triple negative breast cancers as candidates for immune response augmentation. *Eur. J. Cancer.*, **84**, 12–23.
Kvam,V.M. *et al.* (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.*, **99**, 248–256.
Lee,S. *et al.* (2011) Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res.*, **39**, e9.
Louis,M. and Becskei,A. (2002) Binary and graded responses in gene networks. *Sci STKE*, **2002**, pe33.
Marioni,J. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
McIntyre,L. *et al.* (2011) RNA-seq: technical variability and sampling. *BMC Genomics*, **12**, 293.
Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
Robinson,M. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
Robinson,M. and Smyth,G. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
Robinson,M. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
Srivastava,S. and Chen,L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.*, **38**, e170.
Teschendorff,A. *et al.* (2006) PACK: profile analysis using clustering and kurtosis to find molecular classifiers in cancer. *Bioinformatics*, **22**, 2269–2275.
Teschendorff,A. *et al.* (2007) An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol.*, **8**, R157.
Tomlins,S. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
Wang,J. *et al.* (2009) The bimodality Index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Inform.*, **7**, 199–216.