BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2881–2887
doi:10.1093/bioinformatics/btm453

Gene expression

Moderated statistical tests for assessing differences in tag abundance

Mark D. Robinson[1,2] and Gordon K. Smyth[2,*]

[1]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010 and [2]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

Received on June 18, 2007; revised on July 16, 2007; accepted on August 27, 2007

Advance Access publication September 19, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: Digital gene expression (DGE) technologies measure gene expression by counting sequence tags. They are sensitive technologies for measuring gene expression on a genomic scale, without the need for prior knowledge of the genome sequence. As the cost of sequencing DNA decreases, the number of DGE datasets is expected to grow dramatically.

Various tests of differential expression have been proposed for replicated DGE data using binomial, Poisson, negative binomial or pseudo-likelihood (PL) models for the counts, but none of the these are usable when the number of replicates is very small.

Results: We develop tests using the negative binomial distribution to model overdispersion relative to the Poisson, and use conditional weighted likelihood to moderate the level of overdispersion across genes. Not only is our strategy applicable even with the smallest number of libraries, but it also proves to be more powerful than previous strategies when more libraries are available. The methodology is equally applicable to other counting technologies, such as proteomic spectral counts.

Availability: An R package can be accessed from http://bioinf.wehi.edu.au/resources/

Contact: smyth@wehi.edu.au

Supplementary information: http://bioinf.wehi.edu.au/resources/

1 INTRODUCTION

1.1 DGE technology

DGE technologies measure gene expression by generating sequence tags. A number of DGE technologies are now available, including serial analysis of gene expression (SAGE; Velculescu *et al.*, 1995), massively parallel signature sequencing (MPSS; Brenner *et al.*, 2000), sequencing by synthesis (SBS, Margulies *et al.*, 2005) and polony multiplex analysis of gene expression (PMAGE, Kim *et al.*, 2007). The affordability of DGE seems set for a breakthrough in the next few years. The same technologies that aim to produce a $1000 genome may also be applied to expression profiling studies, since the cost-limiting step until recently has been sequencing. In addition, there are a number of promising sequencing-based approaches that are commercially available for gene expression profiling or will be in the near future (Shaffer, 2007).

There are now several variations on the technique (Wang, 2007), but essentially a DGE system can quantify a snapshot of gene expression, without the necessity of either knowing the gene sequence or designing probe sequences, as is needed for microarrays. In the case of SAGE, messenger RNA (mRNA) is extracted from a sample of interest (e.g. a cancer tumour), reverse transcribed into cDNA, fragmented by an enzyme into small *tags* from a fixed location within the transcript. These tags are 10–20 bp in length, depending on the protocol. The tags are then sequenced, either by concatenating them and sequencing a stretch of them, or are sequenced in parallel. For each tag, a count of the number of times it was observed is recorded in a *library* and a larger count is indicative of higher expression. Where an mRNA database or a genome exists, tags can be mapped to a particular mRNA or location in the genome.

SAGE was initially used for determining transcripts expressed in pancreas (Velculescu *et al.*, 1995). Since then, SAGE and its variants have been successful in a number of applications including creating a database of gene expression in human cancers (Lal *et al.*, 1999), discovering prognostic factors in cancer (Aung *et al.*, 2006) and the creation of an atlas of mouse tissue expression (Siddiqui *et al.*, 2005).

Like microarrays, many sequencing-based techniques have applications beyond transcript profiling and we expect the approach developed here to have applications elsewhere. Examples include quantifying micro RNAs (known as miRAGE) (Cummins *et al.*, 2006), copy number analysis (Chen *et al.*, 2002), genome-wide DNA methylation analysis (Hu *et al.*, 2005) and serial analysis of chromatin occupancy (SACO) (Impey *et al.*, 2004).

We focus our attention on the problem of inferring differential expression between two sets of libraries (e.g. cancer versus normal), assuming minimal replication (at least one class has more than one sample). Traditional SAGE is laborious and expensive due to the cost of sequencing. Even with recent developments in high-throughput sequencing, typically most of the 'real estate' is given to sequencing more tags, as opposed to, sequencing more libraries (samples or replicates). So, there are rarely large numbers of libraries to compare. For this reason, it is essential that a statistical analysis method be stable in small samples.

*To whom correspondence should be addressed.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on April 6, 2015

Throughout this article, we refer to DGE data. However, our approach should be equally applicable to other count data of this type, such as peptide counts from mass spectrometry data Lu *et al.* (2007). We have optimized our calculations to the two class comparison problem, but the extension to many classes or accounting for covariates is very straightforward.

Many genome-wide statistical inference methods that share information over all genes, be it in an *ad hoc* way (Tusher *et al.*, 2001) or via hierarchical models (Smyth, 2004), have proven much more sensitive than standard methods. To the best of our knowledge, this is the first exploration of a moderated test statistic applied to the differential expression analysis of tag count data. The novelty in our method is that we share information over all tags in order to stabilize dispersion estimation in small samples.

2 PRELIMINARIES

2.1 Differential expression between multiple DGE libraries

Early methods for differential expression between multiple libraries involved pooling the libraries in each class and using the standard two-sample difference in proportions test or the Fisher exact test. As mentioned previously (Baggerly *et al.*, 2003, 2004, Lu *et al.*, 2005), this pooling inadequately deals with the within-class variability and more flexible models have been proposed. A later method (Ryu *et al.*, 2002) computed two-sample *t*-statistics on the proportions, thereby taking into account the library-to-library variability. However, *t*-statistics for very small samples when the data are genuinely non-normal can be problematic.

Natural choices for a statistical model for tag *counts* may be Poisson or Binomial. In practice, the mean-variance relationship of either the Poisson or Binomial distribution may not provide enough flexibility. More variation is typically observed than the model allows, known as overdispersion. Hence, more recent methods have explored the use of beta-binomial (Baggerly *et al.*, 2003) [and more generally, overdispersed logistic (Baggerly *et al.*, 2004)] and overdispersed log-linear (i.e. gamma-Poisson or negative binomial) models (Lu *et al.*, 2005). The simulation studies of (Lu *et al.*) suggest the negative binomial (NB) assumption can be reliable even with non-NB sampling situations and thus should provide a more flexible framework for real data. For this reason, we make comparisons of our model against that of Lu *et al.* (2005).

2.2 Statistical framework: negative binomial model

For ease of notation, we first consider a single tag. Let Y_{ij} denote the observed count for class i and library j for a particular tag. Here $j = 1, \ldots, n_i$ and for now, we assume just a two-group comparison so that $i = 1, 2$. A special feature of our analysis is that we require only one of n_1 or n_2 to be greater than 1. Strictly speaking, previous methods (Baggerly *et al.*, 2003, 2004; Lu *et al.*, 2005) may be able to operate in this setting. However, in the extreme case of 2 libraries versus 1, one-tag-at-a-time inference would require estimation of three parameters from three observations, which is a rather futile exercise.

Assuming an NB distribution for the tag counts Y_{ij}, we have:

$$Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi)$$

where ϕ is the dispersion. We choose the parameterization such that $E(Y_{ij}) = \mu_{ij}$ and $Var(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}\phi)$, making $\phi = 0$ the Poisson distribution.

Let λ_i be the true relative abundance of this tag in RNA of class i. Then $\mu_{ij} = m_{ij}\lambda_i$ where m_{ij} is the library size for sample j. To assess differences in relative abundance, the null hypothesis H_0: $\lambda_1 = \lambda_2$ is tested against the two-sided alternative, and this is repeated for each tag.

2.3 Dispersion estimation

Robinson and Smyth (2007) discuss a common dispersion model for SAGE data, which uses all tags to estimate a common dispersion (ϕ). The conditional likelihood for a single tag is formed by conditioning on the sum of counts for each class, a straightforward calculation since the sum of identically distributed NB random variables is also NB. The conditioning has the effect of removing the 'nuisance' λ parameter, and is a generalization of restricted maximum likelihood (REML). If the library sizes m_{ij} are equal within each class, the single-tag conditional log-likelihood for ϕ given $z_i = \sum_{j=1}^{n_i} Y_{ij}$ is:

$$l_g(\phi) = \sum_{j=1}^{2}\left[\sum_{j=i}^{n_i}\log\Gamma(y_{ij} + \phi^{-1}) + \log\Gamma(n_i\phi^{-1}) - \log\Gamma(z_i + n_i\phi^{-1}) - n_i\log\Gamma(\phi^{-1})\right]. \quad (1)$$

The common dispersion estimator maximizes the *common likelihood* $l_C(\phi) = \sum_{g=1}^{G} l_g(\phi)$ where G is the number of tags.

In the real situation of unequal library sizes, the counts are not identically distributed, and the conditioning argument does not hold exactly. Robinson and Smyth (2007) use a quantile adjustment to adjust the observed counts up or down depending on whether the corresponding library sizes are below or above the geometric mean (called qCML for quantile adjusted conditional maximum likelihood). This creates approximately identically distributed *pseudodata* that can be inserted into Equation (1), summed over all tags and maximized with respect to ϕ, resulting in a common estimate. With even as few as 100 tags, the qCML estimate is the least biased over a broad range of conditions among a panel of commonly used estimators (Robinson and Smyth, 2007).

2.4 Statistical testing

For testing the difference in expression between two conditions, we compare two statistical tests in what follows. As a default, we use the Wald test that was used in Lu *et al.* (2005). The Wald test simply divides $\hat{\lambda}_2 - \hat{\lambda}_1$ by its estimated standard error. Secondly, we use our previously developed exact test (Robinson and Smyth, 2007).

Briefly, the exact test works as follows. The same quantile adjustment that is used to adjust the tag counts to a common library size for estimation is used to construct the exact test. Using this *pseudodata*, we again use the fact that a sum of independent and identically distributed NB random variables is

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on April 6, 2015

also NB. By conditioning on the total pseudosum (an NB random variable), we can calculate the probability of observing counts as or more extreme than what we observed, resulting in an exact *P*-value.

3 MODERATED DISPERSION ESTIMATION VIA WEIGHTED LIKELIHOOD

3.1 Weighted conditional likelihood framework

The assumption of common dispersion, as in Robinson and Smyth (2007), offers a significant stabilization, compared with tag-wise estimation, especially in very small samples. However, it is not generally true that each tag has the same dispersion, suggesting that inference can be improved by more sophisticated and less drastic stabilization techniques. For microarray data, empirical Bayesian (EB) hierarchical models have been used to stabilize the variance estimates by sharing structure over all genes (Smyth, 2004). Such strategies are adaptive. If the variances are not very different, the EB model arrives at essentially a pooled estimate. However, if the variances are very different, the EB model shrinks a lesser amount. For our NB model, an EB solution is hampered by the fact that the NB falls outside the exponential family and no conjugate prior for ϕ exists. Bradlow *et al*. (2002) suggest a polynomial approximation in order to avoid the computational overhead of stochastic Markov Chain inference methods. Instead of enforcing a common dispersion on all tags, we propose instead to *squeeze* each tag-wise dispersion (denoted as ϕ_g, with an extra subscript to denote the tag) towards the common value (ϕ). We employ *weighted likelihood* and choose likelihood weights so as to approximate an EB solution.

We define the weighted conditional log-likelihood (WL) for ϕ_g to be a weighted combination of the individual and common likelihoods:

$$\mathrm{WL}(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g) \tag{2}$$

where α is the weight given to the common likelihood. This is a special case of weighted likelihood defined by Wang (2006).

The common likelihood plays the same role in the WL as the prior for ϕ_g would play in a Bayesian hierarchical model, with α the prior precision. If $\alpha = 0$ in (2), then we get tag-wise qCML estimates. At the other extreme, if α is chosen sufficiently large, the contributions from any individual log-likelihood is outweighed by the common likelihood and the result is a common dispersion. In between these two extremes lies an estimation scheme where the tag-wise estimates are somewhere between the individual and common estimates.

3.2 Selecting α as an approximate EB rule

We wish to select an appropriate α that will make the estimation adaptive. If evidence suggests that dispersions are not very different, α should be chosen high enough to encourage all tags to shrink strongly towards the common estimate. However, if there is evidence for variable dispersions, α should be selected to shrink a lesser amount.

To understand our strategy for selecting α, suppose that the qCML individual estimators $\hat{\phi}_g$ were normally distributed with means ϕ_g and known variances τ_g^2, and assume the hierarchical model:

$$\hat{\phi}_g|\phi_g \sim N(\phi_g, \tau_g^2), \quad \phi_g \sim N(\phi_0, \tau_0^2), \quad g = 1, \ldots, G.$$

The Bayes posterior mean estimator of ϕ_g would be:

$$\hat{\phi}_g^B = E(\phi_g|\hat{\phi}_g) = \frac{\hat{\phi}_g/\tau_g^2 + \phi_0/\tau_0^2}{1/\tau_g^2 + 1/\tau_0^2}.$$

In practice, the hyperparameters ϕ_0 and τ_0^2 are unknown but can be estimated from the marginal distribution of $\hat{\phi}_g$ to obtain an EB rule. Our strategy is to choose α so that WL coincides with this EB rule. Under this idealistic normal model, the maximum WL estimator is:

$$\hat{\phi}_g^{\mathrm{WL}} = \frac{\hat{\phi}_g/\tau_g^2 + \alpha \sum_{i=1}^G \hat{\phi}_i/\tau^2}{1/\tau_g^2 + \alpha \sum_{i=1}^G 1/\tau_i^2}.$$

This agrees with $\hat{\phi}_g^B$ if ϕ_0 equals the common dispersion estimator

$$\phi_0 = \hat{\phi}_0 = \frac{\sum_{g=1}^G \hat{\phi}_g/\tau_g^2}{\sum_{g=1}^G 1/\tau_g^2}$$

and

$$1/\alpha = \sum_{g=1}^G \tau_0^2/\tau_g^2. \tag{3}$$

It only remains to have an estimator for τ_0^2. Under the normal model, $(\hat{\phi}_g - \phi_0)^2/(\tau_g^2 + \tau_0^2) \sim \chi_1^2$, so a consistent estimator of τ_0 is obtained by solving

$$\sum_{g=1}^G \left[\frac{(\hat{\phi}_g - \hat{\phi}_0)^2}{\tau_g^2 + \tau_0^2} - 1 \right] = 0. \tag{4}$$

This rule for choosing α is not available to us directly, because the qCML estimators $\hat{\phi}_g$ are far from normally distributed, do not have known variances, and in fact can take values on the boundary of the sample space at zero or infinity with positive probability. To evade these difficulties, we take advantage of the fact that score statistics (log-likelihood derivatives) converge to normality more rapidly than do maximum-likelihood estimators. We also note that the estimating Equation (4) can be written in terms of the likelihood score $S_g(\phi) = \partial l_g(\phi)/\partial\phi$ and expected information $I_g(\phi) = E(J_g)$, $J_g = -\partial^2 l_g(\phi)/\partial\phi^2$, functions for ϕ_g. This allows us to state our estimation algorithm as follows.

(1) Find the common dispersion estimator $\hat{\phi}_0$ which maximizes l_C.

(2) Evaluate $S_g(\hat{\phi}_0)$ and $I_g(\hat{\phi}_0)$ for each tag.

(3) Estimate τ_0 by solving

$$\sum_{g=1}^G \left[\frac{S_g^2}{I_g(1 + I_g\tau_0^2)} - 1 \right] = 0.$$

If $\sum S_g^2/I_g < G$ then $\tau_0 = 0$.

(4) Set

$$1/\alpha = \tau_0^2 \sum_{g=1}^G I_g$$

(5) Obtain weighted likelihood estimators $\tilde{\phi}_g$ by maximizing $\mathrm{WL}(\phi_g)$.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on April 6, 2015

This algorithm agrees with (3) and (4) but is of more general application because it uses only quantities evaluated at $\hat{\phi}_0$ to estimate τ_0.

The expected informations I_g are difficult to compute directly, but can be well approximated using the observed informations J_g. For any given value of ϕ_g, I_g should be very nearly directly proportional to the total count $z_1 + z_2$. Hence, we compute the linear regression with intercept zero of J_g on the total pseudo-count (see Fig. 1), and use the fitted values to represent I_g.

The algorithm can actually be applied to any transformation of ϕ. We have found it convenient to implement the algorithm on the $\delta = \phi/(\phi + 1)$ scale because δ takes strictly bounded values.

3.3 Interpretation of the approach

The above algorithm has a nice statistical interpretation. If the dispersions are truly equal (all $\phi_g = \phi_0$), then $E(S_g^2) = I_g$ so that τ_0^2 will be estimated close to zero and hence α will be large. If, however, the dispersions are truly different, then $E(S_g)$ will be non-zero and S_g^2 will be greater than I_g on average, forcing τ_0^2 to be greater than zero and less weight is given to the common likelihood. The more dissimilar the dispersions are, the greater τ_0^2 will be estimated and the less shrinkage is done. The fact that $E(S_g^2) = I_g$ under the null hypothesis is an exact result which does not rely on asymptotic normality. This ensures that our algorithm has good qualitative behaviour even when the number of libraries is small.

4 RESULTS

4.1 Squeezing improves estimation of dispersion in the negative binomial model

For the NB model of tag counts, estimation of ϕ is a crucial step and can have an impact on the determined significance of differential expression. Note that ϕ does not have as direct an influence on statistical tests as the variance does in microarrays, since the variance in our data is also a function of the mean. Lu *et al.* (2005) use a PL model for the estimation of ϕ and estimate separately for each tag. Here, we calculate a common dispersion over all tags and shrink the tag-wise dispersions towards it, in a novel approximate EB strategy.

We first show that our approximate EB approach improves overall estimation of the dispersions in terms of mean squared error (MSE). We compare four estimation strategies: tag-wise qCML, WL using the approximate EB rule, common qCML (Robinson and Smyth, 2007) and tag-wise PL (Lu *et al.*, 2005), over three *true* situations. For all comparisons here, we fixed the library size at 50 000 and means at 10 ($\lambda = 0.0002$), sampled 1000 tags from NB and repeated the simulation 50 times. Overall MSE was calculated for each simulation. MSEs are calculated on the $\delta = \frac{\phi}{1+\phi}$ scale, since there is a non-zero probability of an infinite tag-wise qCML estimate. The first situation considered is a medium number of libraries ($n = 4$) and a fixed dispersion, shown in Figure 2A. This situation obviously favours a common dispersion estimate, and the common dispersion estimate has the lowest MSE. But, note that the approximate EB strategy does well here also, giving enough weight to the common model so as to squeeze the tag-wise dispersions almost entirely to the common value. The remaining two situations involve random dispersions, here taken from a gamma distribution. We picked gamma parameters for the simulation to match the empirical dispersion estimate distribution (using approximate EB estimates) on the Zhang dataset (Zhang *et al.*, 1997), and used a medium ($n = 4$) and large number of libraries ($n = 10$) (Fig. 2B and C, respectively). We can see that the approximate EB solution provides a significant advantage for estimating dispersions and adapts to the situation, showing the relevance of the evidence contained in the scores and informations.

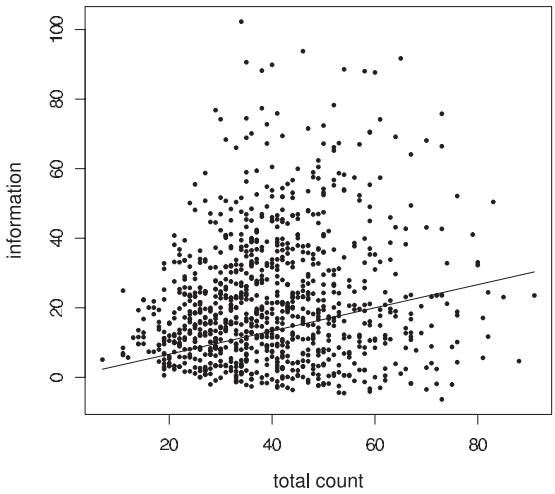

Fig. 1. An illustration of the expected information calculation. Total 1000 tags are sampled for $n = 4$ libraries under $\phi = 0.42$, $\lambda = 0.0002$, $m = 50\,000$. *X*-axis shows the total counts (z) and *Y*-axis shows the observed information J_g. The line through (0,0) predicts the expected information I_g as a function of z.

The weights adapt well to the situation. In the case of fixed dispersions, the weights are large enough to shrink estimates almost to the common value. In the face of more dispersed true dispersions, the weights decrease, as expected. In the case of random dispersion and a larger sample, again the weights decrease to automatically adjust for having more information contained in the tag-wise estimates, showing the approximate EB system seems to be achieving what is expected and therefore provides a suitable rule in practice.

4.2 Comparison of methods: simulated data

Improvements in dispersion estimation can improve our ability to separate the differentially expressed (DE) tags from non-DE. We repeat a subset of the simulation study of Lu *et al.* (2005) and consider an extended, more realistic study.

The simulation in Lu *et al.* (2005) considered sampling 10 000 tags under two conditions, with a fixed λ_1, libraries sizes sampled uniformly between 30 000 and 90 000, comparing 5 libraries of 1–5 libraries of the other. For 5000 tags, an implanted difference of $\lambda_2 = b.\lambda_1$ is used and the remaining tags have no difference ($\lambda_2 = \lambda_1$). When sampling from NB, they choose fixed dispersions at 0.17, 0.42 and 0.95. We repeated their performance analysis for $\lambda_1 = 0.0002$ and $b = 4$, which is directly comparable

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on April 6, 2015

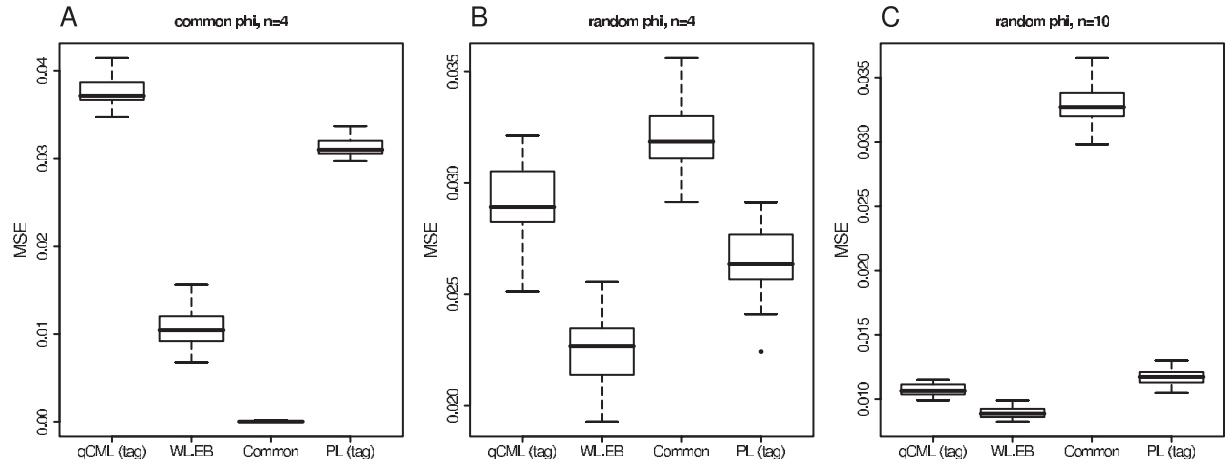

Fig. 2. Boxplots showing the distribution of MSEs over 50 simulations under three sampling conditions: (**A**) constant ϕ with $n=4$; (**B**) gamma distributed ϕ with shape 0.85 and scale 0.5 with $n=4$ (**C**) gamma distributed (same parameters) with $n=10$. Estimators are tag-wise qCML, moderated via WL, common qCML and tag-wise PL, respectively. Each simulation is comprised of sampling 1000 tags with mean 10. MSEs are calculated on the $\delta=\phi/(1+\phi)$ scale.

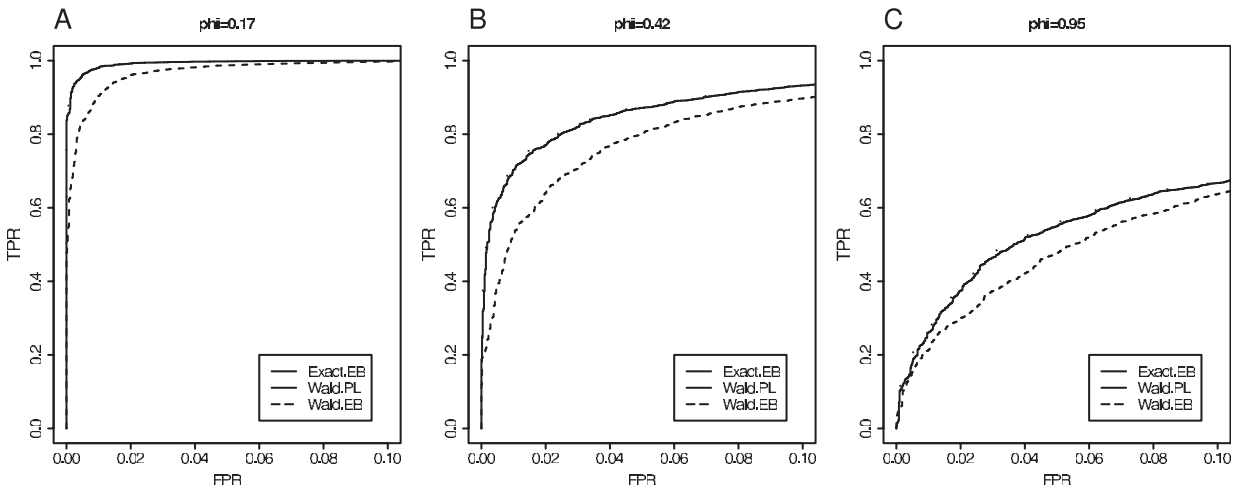

Fig. 3. ROC curves for three statistical tests for separating 5000 truly DE tags from 5000 non-DE tags, with 5 libraries of each condition. Exact.EB represents the small sample test of Robinson and Smyth (2007) with the moderated dispersion estimates of this article. Wald.PL uses the Wald test and the PL estimate of Lu *et al.*, (2005). Wald.EB uses the Wald test with our estimator. FPR: false positive rate (1-specificity), TPR: true positive rate (sensitivity). Here, $\lambda_1 = 0.0002$ and for the DE genes, $\lambda_2 = 0.0008$. The 10 000 tags are sampled under true dispersions: (**A**) $\phi = 0.17$; (**B**) $\phi = 0.42$ and (**C**) $\phi = 0.95$.

to Figure 2 of Lu *et al.* (2005). In Figure 3 in this article, we compare receiver-operating-characteristic (ROC) curves using the Wald test statistic they used, both for their PL estimate and our shrunken estimate. That is, we fit the generalized linear model of Lu *et al.* (2005) with our ϕ estimates to show that an improved estimator is beneficial. In all cases, an improved estimator for the dispersions improves the ability to separate the truly DE and non-DE tags. Here, there is very little difference between the exact and Wald tests.

Next, we extend their simulation study in a number of simple yet important ways. First, we make the problem more realistic by having non-fixed dispersions. Again, we set the random dispersions to the gamma-approximated empirical estimate distribution (shape = 0.85, scale = 0.5). Instead of a fixed λ, we use the empirical distribution of λ estimates for the Zhang dataset and assign the gamma-sampled dispersions to the λs at random. A subtle change we make to our simulation is that the multiplier of the implanted differences, b, does not always *increase* the true means, since larger counts lead to easier estimation problems. Instead, for the tags sampled with true differences, we use $\lambda_1/\sqrt{b}$ and $\lambda_1 \cdot \sqrt{b}$ as the true proportions. Finaly, we consider a more realistic 10% differentially expressed tags and compare small ($n_1 = n_2 = 2$) and moderate ($n_1 = n_2 = 5$) numbers of libraries.

Instead of ROC curves, we prefer false discovery (FD) plots since they highlight the performance at the top ranked tags. Figure 4 shows FD plots for four situations: small and medium true difference ($b = 4, 8$) and small and medium

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on April 6, 2015

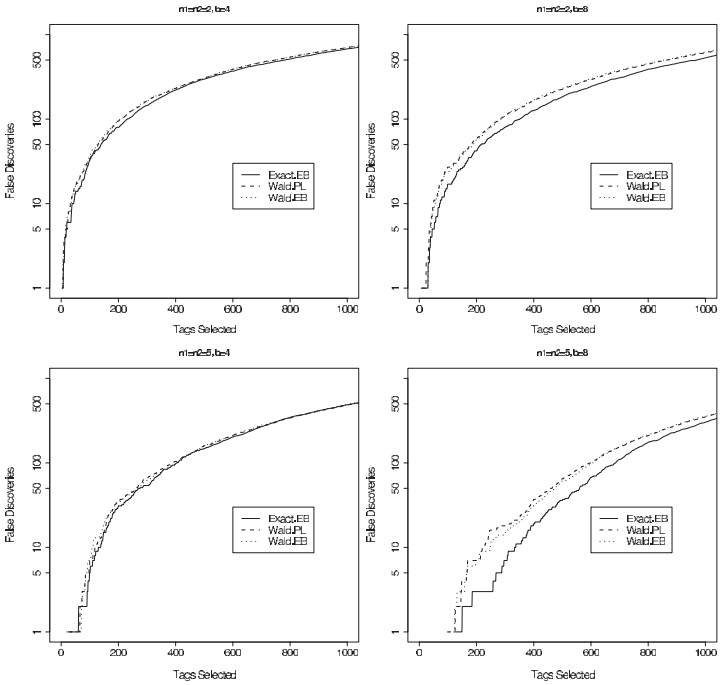

Fig. 4. False discovery plots for different numbers of libraries and different magnitudes of implanted differences. *x*-axis is the number of genes selected (in order from most DE to least DE) and of that many genes selected, the *y*-axis gives the number selected which are false detections. Note the *Y*-axis is presented on the log-scale.

number of libraries. Tags are ranked according to their test-statistic (Wald) or *P*-value (approximate exact test). Since there are 1000 truly DE tags, the top 1000 are selected (*x*-axis) and the number of false detections is plotted (*y*-axis) at each point. Fewer FDs is obviously preferred. We see that using the approximate EB estimate for dispersion estimation results in fewer FDs on average in all cases, regardless of the statistical test being used. If we were to validate the top 200 tags, e.g. the ratio of 'Exact.EB' to 'Wald.PL' FDs is 80/95, 42/59, 31/35 and 3/7 for the four cases presented in Figure 4, thus showing a consistent and practically meaningful improvement.

Of course, this only suggests the statistics are in a desirable order. Since the exact test does not rely on asymptotics for its distributional assumptions, it is best able to achieve a set false positive rate, allowing one to set reasonable cutoffs, presumably after adjusting for multiple testing. In a study of false positive rates for small NB samples, Robinson and Smyth (2007) demonstrate that the exact test is best able to achieve the nominal false positive rate and in fact, the Wald test has the highest false positive rate of all the asymptotic tests. Permutation tests (Tusher *et al.*, 2001) with such few libraries are unlikely to create a reasonable null distribution.

4.3 Application to SAGE data

We apply the method to the SAGE data from Zhang *et al.* (1997), since true biological replicates are available. Our comparison is just between two libraries of normal colon to two libraries of colon tumour. Figure 5A shows the both the dispersion and proportion (λ) estimates, assuming no difference between the sets of libraries. Here, we see that at low abundance, there are a small number of tags with large dispersion. Most of these are cases where one of two replicate counts is zero and the other is non-zero, and in some cases, the tags counts are all zero in one condition and have only one non-zero in the other. In these cases, the one-tag-at-a-time qCML estimate is $\phi = \infty$ ($\delta = 1$), so the squeezing towards the common estimator is essential. Note that it is not surprising that there is a decreasing trend of dispersions as the abundance increases, since the variance of the observations is a function of the mean.

Applying the exact test to the comparison of normal colon samples to colon tumours, we find that 49 genes are up-regulated and 115 genes are down-regulated in tumours, at a 5% false discovery rate (using a Benjamini–Hochberg correction). Figure 5B shows the analogous plot to an 'MA' plot for microarray data, where the *x*-axis is indicative of abundance and the *y*-axis show the magnitude of the change between the two conditions.

5 CONCLUSION

Estimation of dispersion for NB data is critical for assessing the significance of changes in the mean. For tag count data with even the most minimal amount of replication, we have introduced a weighted conditional likelihood estimator that squeezes individual tag-wise dispersions towards the common dispersion. The procedure can be thought of as using a data-dependent prior and finding the maximum a posteriori estimate, or simply as weighted likelihood. We choose the amount of shrinkage according to an approximate empirical Bayes rule. The EB

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on April 6, 2015

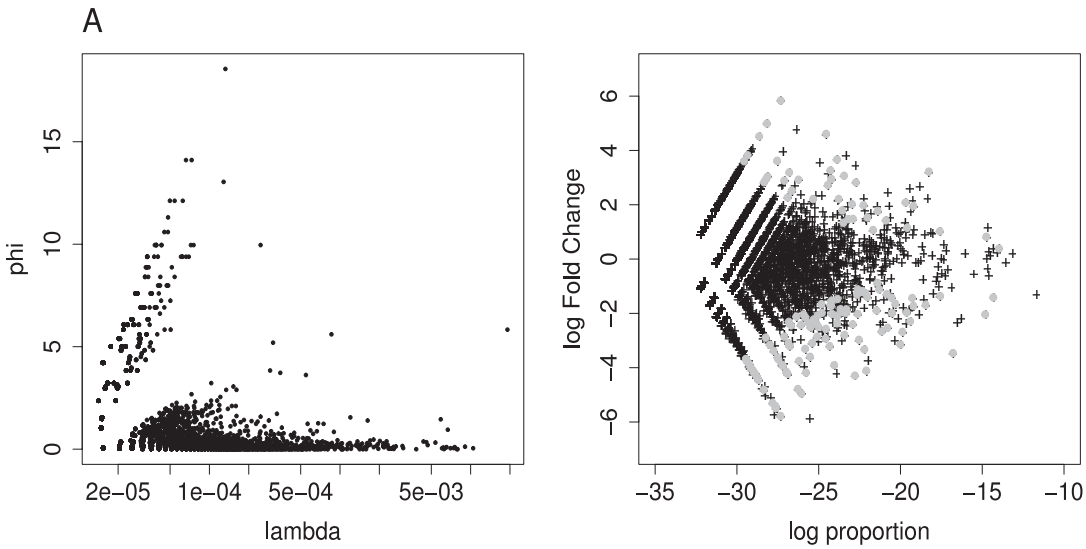

Fig. 5. Applying the method to the Zhang dataset. Panel A gives the 8647 estimates of abundance (*x*-axis) and dispersion (*y*-axis). Panel B gives an 'MA' plot with the significantly DE tags highlighted with grey circles. The tags with zero total count in one class are not shown.

rule works well in practice, adapting to the true similarity of the dispersions, according to sample size and the evidence contained in the first and second derivative of the conditional log-likelihood (scores and informations). The increased precision in estimating dispersion gives an increase in power for testing between experimental conditions. The exact test performs as well or better than the Wald test for testing differences between two experimental conditions. The exact test has the added advantage that it achieves close to its nominal error rates.

Our weighted likelihood shrinkage algorithm is of very general application, requiring only the log-likelihood function and its first two derivatives evaluated at a common parameter estimate. This approach may prove useful in a number of other genome-scale estimation and inference problems.

ACKNOWLEDGEMENTS

The authors wish to thank Terry Speed for valuable discussions. This research was supported by NHMRC Program Grant 406657.

Conflict of Interest: none declared.

REFERENCES

Aung,P.P. *et al.* (2006) Systematic search for gastric cancer-specific genes based on SAGE data: melanoma inhibitory activity and matrix metalloroteinase-10 are novel prognostic factors in patients with gastric cancer. *Oncogene*, **25**, 2546–2557.

Baggerly,K.A. *et al.* (2003) Differential expression in SAGE: accounting for normal between-library variation. *BMC Bioinformatics*, **19**, 1477.

Baggerly,K.A. *et al.* (2004) Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. *BMC Bioinformatics*, **5**, 144.

Bradlow,E.T. *et al.* (2002) Bayesian inference for the negative binomial distribution via polynomial expansions. *J. Comput. Graph. Stat.*, **11**, 189–202.

Brenner,S. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.

Chen,J. *et al.* (2002) High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary cDNAs. *Genes Chromosomes Cancer*, **33**, 252–261.

Cummins,J.M. *et al.* (2006) The colorectal microRNAome. *Proc. Natl Acad. Sci. USA*, **103**, 3687–92.

Hu,M. *et al.* (2005) Distinct epigenetic changes in the stromal cells of breast cancers. *Nat. Genet.*, **37**, 899–905.

Impey,S. *et al.* (2004) Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell*, **119**, 1041–1054.

Kim,J.B. *et al.* (2007) Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*, **316**, 1481–1484.

Lal,A. *et al.* (1999) A public database for gene expression in human cancers. *Cancer Res.*, **59**, 5403–5407.

Lu,J. *et al.* (2005) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, **6**, 165.

Lu,P. *et al.* (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, **25**, 117–124.

Margulies,M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Robinson,M.D. and Smyth,G.K. (2007) Small sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, published online August 29, 2007.

Ryu,B. *et al.* (2002) Relationships and Differentially Expressed Genes among Pancreatic Cancers Examined by Large-scale Serial Analysis of Gene Expression. *Cancer Res.*, **62**, 819–826.

Shaffer,C. (2007) Next-generation sequencing outpaces expectations. *Nat. Biotechnol.*, **25**, 149.

Siddiqui,A.S. *et al.* (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl Acad. Sci. USA*, **102**, 18485–18490.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mole. Biol.*, **1**, Article 3.

Tusher,V.G *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Velculescu,V.E. *et al.* (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wang,S.M. (2007) Understanding SAGE data. *Trends Genet.*, **23**, 42–50.

Wang,X. (2006) Approximating Bayesian inference by weighted likelihood. *Can. J. Stat.*, **34**, 279–298.

Zhang,L. *et al.* (1997) Gene Expression Profiles in Normal and Cancer Cells. *Science*, **276**, 1268–1272.

Downloaded from http://bioinformatics.oxfordjournals.org/ by guest on April 6, 2015