Small Sample Estimation of Negative Binomial Dispersion, with applications to SAGE data

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

[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050, Australia

[2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia

[*]Corresponding author: mrobinson@wehi.edu.au

Phone: +61 3 9345 2628, Fax: +61 3 9347 0852

June 8, 2007

Abstract

We derive a quantile-adjusted conditional maximum likelihood (qCML) estimator for the dispersion parameter of the negative binomial (NB) distribution and compare its performance, in terms of bias, to various other methods. Our estimation scheme outperforms all other methods in very small samples, typical of those from serial analysis of gene expression (SAGE) studies, the motivating data for this study. The impact of dispersion estimation on hypothesis testing is studied. We derive an "exact" test that outperforms the standard approximate asymptotic tests.

1 Introduction

Count data arise in numerous biological applications and can often be modeled by a Poisson distribution, where the mean and variance are the same. However, in situations where there is a positive correlation in the occurence of events, the observed variation is significantly greater than the mean and an extension to the Poisson model is more appropriate. A popular alternative is the negative binomial (NB) model, also known as the gamma-Poisson model, since the Poisson rate parameter is a mixture of gamma random variables with fixed coefficient of variation.

Serial analysis of gene expression (SAGE) is a celebrated molecular biology technique that affords the simultaneous measurement of the expressed messenger ribonucleic acid (mRNA) population at a given time or state of a biological organism [11]. Briefly, the procedure works as follows. Messenger RNA, an intermediate molecule between the DNA code and an eventual protein, is extracted from a cell and many short (e.g. 14 base pair) regions called *tags* are sequenced and recorded. Typically, there are upwards of 30,000 total tags sequenced for a single sample, called a *library*, and many tags occur multiple times. The tag counts are representative of the abundance the corresponding mRNA molecule. Unfortunately, the technique is expensive due to the current cost of sequencing and we often only have access to a very small number of libraries, thus making estimation and inference a challenge.

We explore the use of a NB model for each tag across multiple libraries and discuss the estimation of the dispersion parameter. The data on hand consist of counts for many seemingly unrelated tags, with each tag being observed through a small number of libraries. The major statistical inference task with SAGE data is to find mRNA transcripts which differ in expression between experimental conditions or between classes of patients. To do this, a statistical model such as the NB can be used to determine whether observed differences in tag counts can be attributable to chance or not. If there are no replicates in each class, a χ^2 test for 2-by-2 contingency tables or Fisher's exact test can be conducted for each tag [3]. Even with replication, previous analyses have ignored this and pooled the tag counts over libraries

[3]. This however overestimates the significance of each difference because it fails to allow for inter-library variation. Here, inter-library variation is due to a combination of biological (e.g. Zhang et al. [12] use two different patients for each tissue) and technical sources, however due to the current expense of SAGE, technical replication is rarely available. Models used to capture this additional variation include over-dispersed generalized linear models (GLMs), with the counts as either binomial [1] or Poisson [5], the latter proving to be more powerful in most situations.

We describe a new approach for estimating NB dispersion that uses simple distributional adjustments in combination with conditional maximum likelihood. Even though we typically only have a small number of observations (e.g. 2 libraries for a given experimental conditon), we do have counts for many tags and can leverage what little information that does give about the dispersion.

The paper is organized as follows. Section 2 describes the NB model and the generalizations we explore. Section 3 reviews the methods for estimating dispersion in NB models. Section 4 introduces our conditional maximum likelihood approach to estimation and Section 5 discusses hypothesis testing, including our new "exact" test. Performance comparisons are made in Section 6 and Section 7, and discussion follows in Section 8. Software for all calculations is available from the authors.

2 Negative Binomial Models

2.1 Genesis

Let Y be a NB random variable with mean μ and dispersion ϕ, denoted $Y \sim \mathrm{NB}(\mu, \phi)$. We choose the parameterization such that the probability mass function is

$$f(y;\mu,\phi) = P(Y=y) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)} \left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}} \left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y} \tag{1}$$

giving $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \mu + \phi\mu^2$. We focus here on *over*-dispersion where $\phi > 0$, though $\phi > -\mu^{-1}$ is in fact permitted by the model.

Note also that as $\phi \to 0$, the distribution reduces to the Poisson. Based on recent work in the analysis of SAGE data [5], the NB distribution is a robust alternative to the beta-binomial (i.e. overdispersed logistic regression) distribution and other models. In the SAGE context, ϕ accounts for the library-to-library variability.

2.2 Single tag with unequal library sizes

SAGE data motivates the following model for the counts for a single tag across n libraries. Consider $Y_1, \ldots, Y_n$ as independent and $NB(\mu_i = m_i\lambda, \phi)$ where m_i is the library size (i.e. total number of tags sequenced for library i) and λ represents the proportion of the library that is a particular tag. An important special case is $m_i \equiv m$, where the Y_i are identically distributed. This simple situation is integral to our proposed estimation mechanism. With an identically distributed sample, the MLE of λ will always be the sum of counts divided by sum of library sizes, independent of ϕ. If $m = 1$, the MLE of λ is the mean, as with the Poisson model. In the case of different m_i, the MLE of λ will depend on ϕ and ML estimation of the two parameters proceeds jointly.

2.3 Many tags with unequal library sizes: SAGE data

In a SAGE experiment, we simultaneously make observations on T tags over n libraries. Typically T is in excess of 5000-10000, depending on the diversity of the expressed mRNA and the sequencing time. We assume all tags are independent for the purposes of estimation and inference. This is not strictly true since some sets of genes are involved in related biological functions and are inherently coregulated in their expression levels. However, our primary concern is getting an unbiased estimate of the dispersion parameter and consistency of the estimator is unaffected by this correlation.

For tag t and library i, we denote the random variables as Y_{ti}, and model them as NB with

mean $m_i\lambda_t$ and dispersion ϕ. Notice that all tags are assumed have a *common dispersion*, ϕ. Since we are always dealing with very small samples (e.g. $n = 3$), estimating a different dispersion for each tag separately is unrealistic and, in any case, evidence for truly different ϕ is unclear.

3 Review of dispersion estimators

It is well known that MLEs in general tend to underestimate variance parameters because they fail to adjust for the fact that the mean is estimated from the same data.

Pseudo-likelihood (PL) models use a distribution-free goodness-of-fit statistic for estimating parameters in the variance function of a GLM [9]. In the one-tag and n replicate libraries case, $\hat{\phi}_{PL}$ is defined by:

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1 + \hat{\phi}_{PL}\hat{\mu}_i)} = n - 1. \tag{2}$$

where $\hat{\mu}_i = m_i\hat{\lambda}$ is the MLE of λ given ϕ. This is the dispersion estimator used by Lu et al. [5] in their analysis of SAGE data, though they estimate a separate dispersion for each tag.

Quasi-likelihood (QL) models estimate the dispersion in an identical fashion, but replace the Pearson statistic with a deviance statistic D [6]. For the NB model,

$$D = 2\sum_{i=1}^{n} \left\{ y_i \log\left[\frac{y_i}{\hat{\mu}_i}\right] - (y_i + \phi_{QL}^{-1}) \log\left[\frac{y_i + \phi_{QL}^{-1}}{\hat{\mu}_i + \phi_{QL}^{-1}}\right] \right\} = n - 1. \tag{3}$$

Thus, the PL and QL methods do allow for estimation of the mean in computing the residual degrees of freedom. Nelder and Lee [7] compare estimators for the NB and find PL almost always less efficient than QL and often less efficient than the MLE.

Adjusted likelihood methods are another means of reducing the bias introduced into MLEs. Cox and Reid's approximate conditional inference (CR) [2] is the most convenient form, and

adjusts the profile log-likelihood for ϕ by a term containing the observed information for λ:

$$l_{CR}(\phi) = l(\hat{\lambda}, \phi) - \frac{1}{2} \log |j_{\lambda\lambda}(\phi, \hat{\lambda})|. \tag{4}$$

Saha and Paul (2006) [8] derive a bias-corrected version of the MLE, though the smallest sample size they use for their performance comparisons is $n = 10$ and they do not present an estimator for the non-identically distributed case. Lord (2006) [4] fits the same model but considered sample sizes of 50, 100 and 1000, which are much higher than we can expect.

Each of the above estimators can be extended to the many-tag SAGE scenario simply by summing quantities over tags.

4 Conditional Dispersion Estimation

4.1 Conditional Maximum Likelihood

If all libraries are the same size (i.e. $m_i \equiv m$), the sum $Z = Y_1 + \ldots + Y_n \sim NB(nm\lambda, \phi n^{-1})$. It is trivial to show that the sample sum is a sufficient statistic for λ. Hence, we can form an exact conditional likelihood for ϕ that is independent of λ and estimation can be done univariately using conditional maximum likelihood (CML). In this setting, CML fits under the framework of residual maximum likelihood [10]. For NB, CML selects ϕ that maximizes:

$$l_{Y|Z=z}(\phi) = \left[\sum_{i=1}^{n} \log \Gamma(y_i + \phi^{-1}) \right] + \log \Gamma(n\phi^{-1}) - \log \Gamma(z + n\phi^{-1}) - n \log \Gamma(\phi^{-1}). \tag{5}$$

Where it exists, CML is an exact version of the CR adjusted profile likelihood. However, in the unequal m_i situation, the conditional likelihood cannot be written in closed form. As before, the many-tag SAGE setting would maximize with respect to ϕ after summing over the additional index t, having each tag contribute something to the inference.

4.2 Quantile Adjusted CML

When the library sizes are unequal, we devise a simple but effective approximate approach to equate the library sizes and create quantile-adjusted "pseudodata", allowing us to use the above CML machinery to achieve an accurate estimate of ϕ.

Let $m^* = (\prod_{i=1}^{n} m_i)^{\frac{1}{n}}$, the geometric mean of the library sizes. We adjust the observed data as if they were all sampled as identically distributed $NB(m^*\lambda, \phi)$, as follows:

1. Initialize ϕ (for example, the unadjusted CML estimate).
2. Given the current value of ϕ, estimate the rate λ.
3. Assuming each observation y_i was sampled from a NB distribution with mean $m_i\lambda$ and dispersion ϕ, calculate the observed percentiles,

$$p_i = P(Y < y_i; m_i\lambda, \phi) + \frac{1}{2}P(Y = y_i; m_i\lambda, \phi), i = 1, \ldots, n. \tag{6}$$

4. Using a linear interpolation of the quantile function, calculate pseudodata from a NB distribution with mean $m^*\lambda$ and dispersion ϕ, having quantiles p_i. Note that pseudodata will then be continuous and can be as negative as -0.5. The pseudodata for each tag is approximately identically distributed.
5. Calculate ϕ using CML on the pseudodata.
6. Repeat steps 2-5 until ϕ converges.

Effectively, this adjusts observed counts upwards for libraries with size below the geometric mean and downwards for counts from an above-average sized library. We call this quantile-adjusted conditional maximum likelihood (qCML). In the case of many samples (i.e. many tags), the above procedure creates pseudodata for each tag and the estimation of the common ϕ is done over all tags.

5 Hypothesis Testing

5.1 Asymptotic tests

We discuss the simple setting of a two sample comparison (e.g. cancer versus normal), that would be repeated for each tag in the SAGE context. For all tests, the common dispersion is estimated from a large number of tags and therefore is treated as fixed when applying these tests.

We wish to test whether the relative abundance under experimental condition A is the same as that in condition B, leading to a null hypothesis of $H_0 : \lambda_{tA} = \lambda_{tB}$, for each tag. There are no exact tests for testing such a hypothesis under the NB with unequal library sizes, but the standard approximate options based on asymptotics include the Wald test, the Score test or the likelihood ratio (LR) test. For the small sample sizes we expect from SAGE experiments, the large-sample approximations are questionable. For this reason, we develop a new test.

5.2 A small-sample test procedure

The quantile adjustment, discussed in the previous section, provides the opportunity for an exact test. If the quantile adjustment is applied under the null model of no difference, the pseudodata are then approximately identically distributed, leading to known distributions of the within-condition pseudodata totals for each tag. Also, the sum of the total tag pseudocount over all libraries has a known distribution. For the two-sample test discussed in Section 4, let Z_{tA} and Z_{tB} be the sum of pseudocounts for class A and B, respectively, over the number of libraries, n_A and n_B. Under the null, $Z_{tk} \sim \text{NB}(n_k m^* \lambda_t, \phi n_k^{-1}), k \in A, B$. One can construct an exact test similar to the Fisher exact test for contingency tables, but replacing the hypergeometric probabilities with NB. Conditioning on the total pseudo sum, $Z_{tA} + Z_{tB}$, also a NB random variable, we can calculate the probability of observing class totals as or more extreme than that observed, giving an exact method of assessing differential

expression. In other words, the two-sided P-value is defined as the sum of the probabilities of class totals that are no more likely than those observed.

The test is approximate in the sense that the quantile adjustment make the pseudodata approximately identically distributed, but the probabilities are otherwise exact. As with all of the asymptotic tests, the new test also has a many-class analog, which we do not consider here.

6 Comparison of Dispersion Estimators

6.1 General Considerations

Since the SAGE setting leads us to a common dispersion model, we are primarily concerned with having an estimator with minimal bias, to ensure that significance statements concerning λ are accurate. An analysis of significance testing is given in the next section.

Based on previous experience, we would expect to find that ML has smallest variance but largest bias. CML should have smaller bias but larger variance. CR should be closer to qCML and should perform well in larger samples. QL should outperform PL for non-normal models and especially in the small samples we consider here. In terms of bias, we expect qCML to be the best, followed by CR, QL, PL and MLE.

We have chosen four scenarios, with small and large μ, and small and large ϕ. For ϕ, we have chosen small at 0.25 and large at 1. The primary performance criteria is bias of the estimate of ϕ, since in real settings we have a large number of tags and the bias of ϕ will affect the signficance of test statistics in the search for differential tags. We focus here on the smallest samples possible where we can estimate 2 parameters ($n = 3$), and these are typical of SAGE studies.

We choose library sizes uniformly between 20000 and 80000 to simulate the effect of differing SAGE library sizes and select $\lambda = .0001$ and $\lambda = .0005$ as small and large, giving means between 2-8 and 10-40, respectively. These are meant to imitate typical SAGE counts

and the settings are similar to those used in the simulations in Lu et al [5].

Since there is non-zero probability of data situations in which the ML or CML estimate is infinite, we present results on the $\delta = \frac{\phi}{\phi+1}$ scale that is constrained to $(0, 1)$. It is important to note that biases observed on the δ scale are always more extreme when converted to the original ϕ scale.

6.2 Single tag with unequal library sizes

To achieve a baseline for comparison, we consider estimation for a single tag. Later, we consider estimation of a common dispersion using multiple tags. Figure 1 gives boxplots of 1000 independent samples of 1 tag over 3 libraries. These show the bias of each of the 6 estimators. In this situation, most methods underestimate the true dispersion. Here, CML acts as a control as it applies CML using incorrect assumptions and should overestimate the true value. Fortuitously, CML appears least biased and would be considered the top performer in this situation. Beyond CML, it appears the QL performs best in this situation.

6.3 Many tags with unequal library sizes: SAGE data

Here, we consider the actual SAGE setting, where many tags from unequally sized libraries are used to estimate the common dispersion. We consider a "mini-SAGE" setting with 100 tags being used to estimate the common dispersion. The library sizes and values for λ and ϕ are chosen as before and the results are shown in Figure 2. Here, we see that qCML is the only estimator that is reliable under the whole spectrum of λ and ϕ values. CML should overestimate the true value and clearly does, especially for large means and somewhat surprisingly, for small ϕ. PL also consistently underestimates the true value whereas ML grossly underestimates in all cases. QL overestimates the true value, especially for the more difficult small mean, large dispersion case. CR performs quite well in most situations, except the small mean, large dispersion state.

6.4 Other Settings

As the number of tags increases, the performance of the qCML is even more apparent (data not shown). We have tried 1000 tags and the results observed in Figure 3 are identical, yet more defined. qCML is unbiased in all four scenarios, ML and PL underestimate and QL and CML overestimate the true value. Similarly, CR performs very well, except in the small mean, large dispersion situation.

7 Comparison of Testing Procedures

7.1 Infinite Evidence Against the Null Hypothesis

There is a special case where some asymptotic tests seem to break down for the two-sample NB model and in fact, the same problem would exist for Poisson, binomial and beta-binomial models.

The following table illustrates this. For a two-class comparison, 2 libraries for each class, equal library sizes, suppose we observe zero total count in one class and varying levels of non-zero counts in the other class. In this case, the true value of λ for class 1 is 0, on the boundary of the parameter space. Assuming an arbitrary selection of $\phi = .5$, we calculate the various significance tests mentioned in the previous sections:

Class 1		Class 2		Wald		Score		LR		Exact
Y_{11}	Y_{12}	Y_{21}	Y_{22}	z	p	z	p	χ^2_1	p	p
0	0	6	8	1.23e-03	0.999	2.26	0.024	9.77	1.77e-03	1.17e-02
0	0	60	80	1.30e-03	0.999	2.75	0.006	25.69	4.01e-07	3.75e-06
0	0	600	800	1.37e-03	0.999	2.82	0.005	43.81	3.62e-11	4.31e-10
0	0	6000	8000	1.45e-03	0.999	2.83	0.005	62.20	3.10e-15	4.37e-14

There are two notable consequences. First of all, the Wald test tends to zero since the estimated standard error approaches infinity. Second, the score test gets more extreme, but

seems to approach an asymptote at approximately 2.83. As a result, the p-values never get very extreme. In a real SAGE setting, the tests would need to be adjusted for the large number of tests being conducted and the score test would lack power. Therefore, one can rule out both the Wald and score tests as reasonable approaches for testing. The LR and exact probabilities get more extreme with increasing evidence against the null hypothesis, but as we shall see in the next section, only the exact test holds its size.

7.2 Size of the Test: ϕ known

Assuming the dispersion is known, we look at how well each of the approximate tests achieves a 5% false positive rate given the 5% cutoff of their respective null distributions. Random samples under the null hypothesis of no difference are taken over 1000 tags, with low mean and high dispersion, as used previously, and sampled 30 times. For a fair comparison of the tests, the known value of ϕ is used. Figure 3 shows the 30 observed false positive rates (i.e. test statistics more extreme than the 5% cutoff) for a 2 versus 2 library comparison and 5 versus 5 library comparison.

Because of the discrete nature of count data, it is not surprising that the exact test is somewhat conservative, because an exact 5% cutoff can only rarely be achieved. For the larger samples, there becomes a smaller discreteness effect, as expected. Both the Wald test and the LR test give test statistics that are more extreme than their approximate χ^2_1 random variables, leading to high false discovery rates. The asymptotic normal score test appears to be the best for small samples, in terms of achieving the 5% nominal error rate, but fails with larger samples. The exact test performs reliably in both situations and is the only one to be correct or conservative.

7.3 Size of the Test: small-sample test with different estimators of ϕ

In the last section we considered the range of approximate tests with the dispersion known. Here we show the effect of the dispersion bias on the false positive rate of the test. Figure 4 shows the effect of dispersion estimation on the false positive rates. Again, we make 30 repeat samples of 1000 tags, a 2 versus 2 comparison with no difference and estimate the fraction of tags below a 5% probability. We chose the low mean, high dispersion case, as this was a situation where the estimators differed significantly (recall top left panel of Figure 2).

Not suprisingly, over estimating the dispersion (CML, QL, CR) results in a conservative test whereas under estimation (ML, PL) results in a liberal test and therefore more false discoveries than expected. qCML results in a slightly conservative test, due to the discreteness as mentioned previously.

7.4 Power Considerations

Lastly, we implant known differences and illustrate the power of the different tests.

We show the results as area under the receiver operating characteristic curve (AUC). Here, we sample 1000 tags from NB in the low mean, high dispersion setting with 10% having been sampled with a known difference between the classes ($\frac{\lambda_A}{\lambda_B} = 8$ or $\frac{\lambda_B}{\lambda_A} = 8$). The class sizes are set at $n_A = n_B = 2$, with library sizes at 10000 and 100000, for each class. Differences are embedded so that the true means of both classes are affected, one up and one down, picked randomly. Though eight-fold may appear to be quite a large difference, in the low mean and high dispersion setting, it is still a challenging statistical inference task. Generating 100 such datasets of 1000 tags, we compute the AUC for each combination of the 4 statistical tests and 6 estimators. Figure 5 shows boxplots of the AUCs.

It is evident that the larger effect on the ability to find the true differences is the statistical test used, not the estimate of dispersion. The Wald test suffers from its flaw in detecting dif-

ferences where one of the estimates of relative abundance is 0, thus introducing false negatives and lowering the achievable AUC. The results suggest that the dispersion estimate has only a small impact on the ordering. This is perhaps not suprising, as the dispersion estimate does not come into the significance equation directly, as it does in a Gaussian testing situation, for instance.

Another way to represent the ability of different tests and different estimators to rank the genes, is to record the mean rank of the truly different tags, with lower rankings being better. To focus on differences between the estimators, we compared the mean ranks of the truly different tags on a per-dataset basis. Figure 6 shows boxplots of mean ranks for the same 100 datasets sampled above. The vertical axis shows the difference between the mean rank for the qCML estimator and that for each of the other estimators, for each of the simulated data sets. In all cases the genes have been ranked using the exact test. The qCML estimator outperforms all the other estimators by 1-10 false discoveries. Although the differences are relatively small, they are remarkably consistent, so that qCML beats QL and QR on almost every dataset. Surprisingly, the ML and PL estimators, both of which are negatively biased, perform well. This suggests that it may be advantageous to have a slight negative bias in the dispersion estimator for the purpose of gene ranking.

8 Discussion

We have derived an estimator, qCML, for the dispersion parameter of the NB distribution and compared it against several other estimators using an extensive simulation study. While qCML does not outperform all estimators under all circumstances, it is the most reliable in terms of bias on a wide range of conditions and specifically performs best in the situation of many small samples with a common dispersion, the model which is applicable to SAGE data. We have deliberately focused on very small samples, due to the fact that DNA sequencing costs prevent a large number of replicates for SAGE experiments.

For a single tag with a small number of libraries, all estimators offer mediocre performance and there is no clear winner. As the number of tags used to estimate the common dispersion increases while holding the number of libraries at a small number, qCML is clearly the best estimator. With more libraries (e.g. $n = 10$), CR performs about as well as qCML.

The same quantile adjustment we use for dispersion estimation affords us a new *exact* statistical test, similar in flavour to the Fisher exact test for contingency tables. The exact test is compared to the standard Wald, LR and Score tests in terms of achieving their nominal false discovery rates and on power. The exact test is the only test to achieve its nominal false discovery rate, due in part to the fact that sample sizes are not large enough for the asymptotics to achieve a reasonable approximation. Bias in the estimation of the dispersion has significant impact on the expected false positive rates but does not seems to adversely affect power to the same degree, since the ranking is only slightly affected by the dispersion estimate. Other than the Wald test, the remaining tests have similar power to detect differences.

The quantile adjustment approach can be applied more generally, not simply to SAGE data. For example, qCML can be applied to any NB regression, and probably outperforms the standard methods such as PL in small sample situations. For SAGE data, we had picked the geometric mean of the library sizes in order to adjust the observed data so that CML machinery can be put to use. For general GLMs, one must create pseudodata to equate the fitted means analogously and the iterative procedure is straightforward. This pseudodata could be used only for the over-dispersion estimation or for both estimation and testing. More experimentation would be required in the GLM setting.

Acknowledgements

Thanks are due to Alicia Oshlack for valuable discussions and to the editor and an associate editor for suggestions on making the manuscript clearer. This work is supported by NHMRC

Program Grant 406657.

References

[1] KA Baggerley, L Deng, JS Morris, and CM Aldaz. Overdispersed logistic regression for sage: Modelling multiple groups and covariates. *BMC Bioinformatics*, 5:144, 2004.

[2] DR Cox and N Reid. Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc B*, 49(1):1–39, 1987.

[3] AJ Kal, AJ van Zonneveld, V Benes, M van den Berg, MG Koerkamp, K Albermann, N Strack, JM Ruijtter, A Richter, B Dujon, and W Ansorge. Dynamics of gene expression revealted by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol Bio Cell*, 10:1859–1872, 1999.

[4] D Lord. Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention*, 38:751–766, 2006.

[5] J Lu, JK Tomfohr, and Kepler TB. Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6:165, 2005.

[6] JA Nelder. Quasi-likelihood and pseudo-likelihood are not the same thing. *Journal of Applied Statistics*, 27(8):1007–1011, 2000.

[7] JA Nelder and Lee Y. Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. *J. R. Statist. Soc B*, 54(1):273–284, 1992.

[8] K Saha and S Paul. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61:179–185, 2005.

[9] GK Smyth. Pearson's goodness of fit statistic as a score test statistic. *Science and Statistics: A Festschrift for Terry Speed*, pages 115–126, 2003.

[10] GK Smyth and Verbyla AP. A conditional likelihood approach to residual maximum likelihood in generalized linear models. *Journal of the Royal Statistical Society Series B*, 58(3):565–572, 1996.

[11] VE Velculescu, L Zhang, B Vogelstein, and KW Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

[12] L Zhang, W Zhou, VE Velculescu, SE Kern, RH Hruban, SR Stanley R. Hamilton, B Vogelstein, and KW Kinzler. Gene expression profiles in normal and cancer cells. *Science*, 276:1268–1272, 1997.

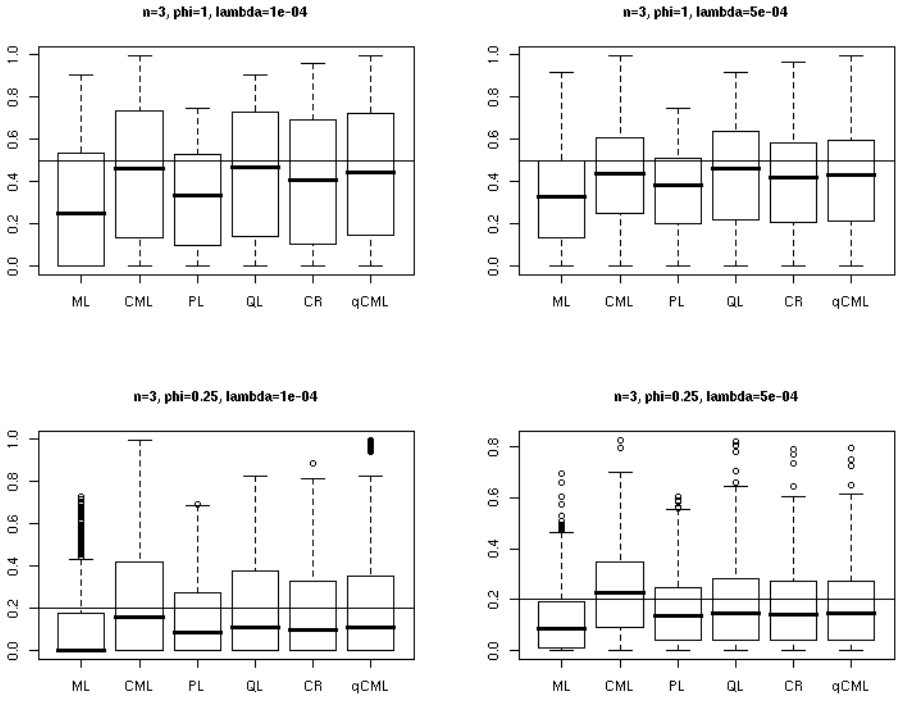

Figure 1: Estimates of ϕ for 1000 samples of a single "tag" (i.e. each observation from a different library size) of size $n = 3$ on the $\delta = \frac{\phi}{\phi+1}$ scale. The horizontal line indicates the correct value.

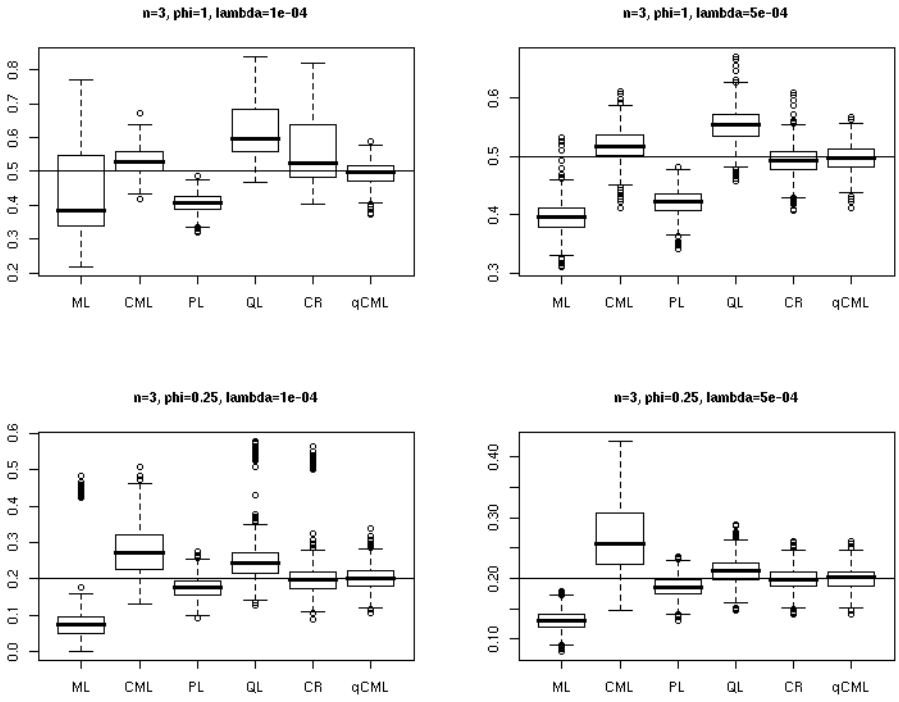

Figure 2: Estimates of common ϕ for 1000 samples with 100 tags (i.e. each dispersion is estimated from 100 randomly sampled tags) of with $n = 3$ libraries. Results are presented on the $\delta = \frac{\phi}{\phi+1}$ scale. The horizontal line indicates the correct value.

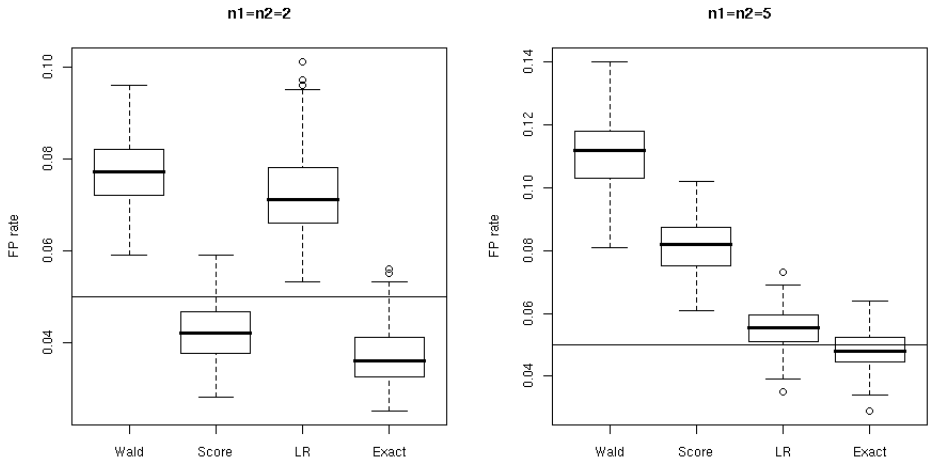

Figure 3: Achieved false positive rates sampled under the null hypothesis (i.e. no differential expression) for four statistical tests, in either a 2 versus 2 comparison or a 5 versus 5 comparison.

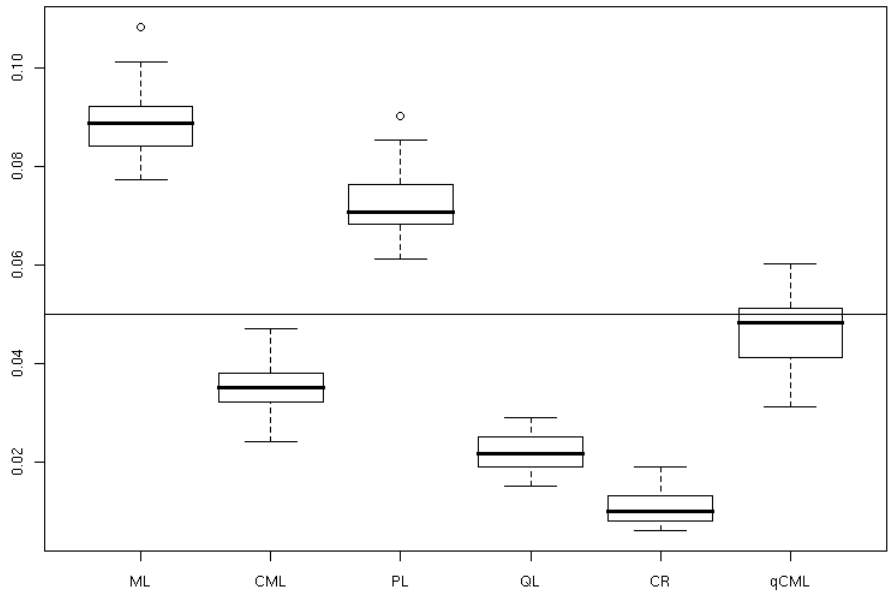

Figure 4: Achieved false positive rates sampled under the null hypothesis (i.e. no differential expression) for the exact test using different estimators.

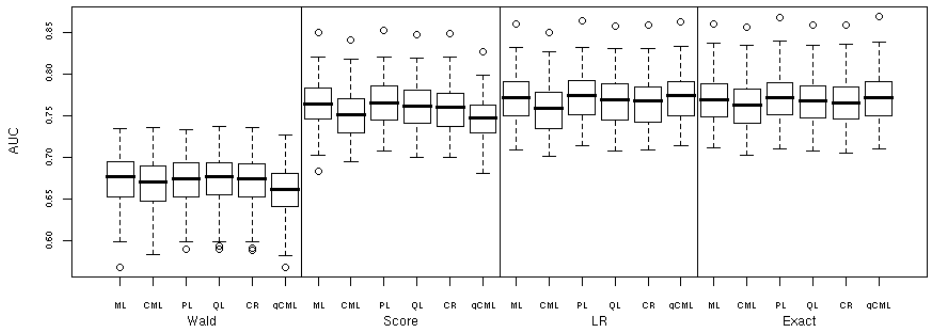

Figure 5: Achieved AUCs for 4 statistical tests and 6 estimators. The tests and estimators are presented in the same order as all previous figures.

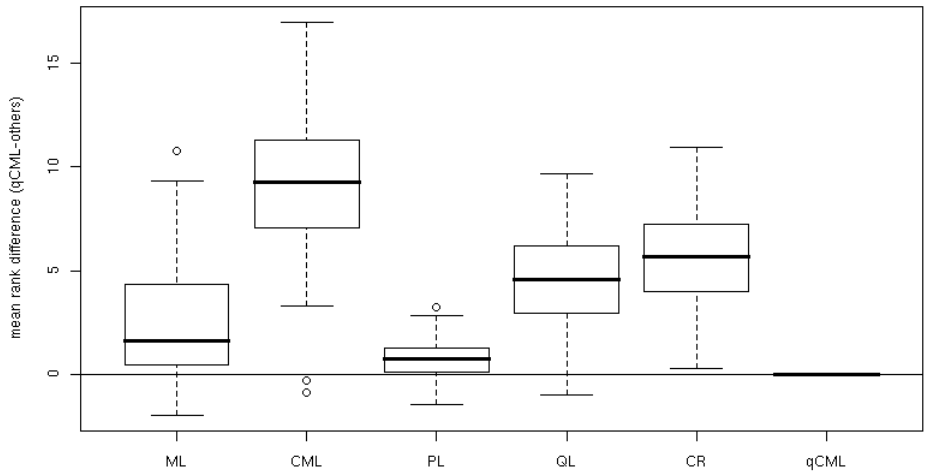

Figure 6: Difference in mean ranks of the truly differential tags between the qCML estimator and all other estimators on a per-dataset basis, using the exact test.