Lun, A.T.L. and Smyth, G.K. (2017). No counts, no variance: allowing for loss of degrees of freedom when assessing biological variability from RNA-seq data. *Statistical Applications in Genetics and Molecular Biology* 16(2), 83-94.

No counts, no variance: allowing for loss of degrees of freedom when assessing biological variability from RNA-seq data

Aaron T. L. Lun and Gordon K. Smyth
The Walter and Eliza Hall Institute of Medical Research

7 July 2017

Abstract

RNA sequencing (RNA-seq) is widely used to study gene expression changes associated with treatments or biological conditions. Many popular methods for detecting differential expression (DE) from RNA-seq data use generalized linear models (GLMs) fitted to the read counts across independent replicate samples for each gene. This article shows that the standard formula for the residual degrees of freedom (d.f.) in a linear model is overstated when the model contains fitted values that are exactly zero. Such fitted values occur whenever all the counts in a treatment group are zero as well as in more complex models such as those involving paired comparisons. This misspecification results in underestimation of the genewise variances and loss of type I error control. This article proposes a formula for the reduced residual d.f. that restores error control in simulated RNA-seq data and improves detection of DE genes in a real data analysis. The new approach is implemented in the quasi-likelihood framework of the edgeR software package. The results of this article also apply to RNA-seq analyses that apply linear models to log-transformed counts, such as those in the limma software package, and more generally to any count-based GLM where exactly zero fitted values are possible.

Keywords: RNA sequencing, differential expression, generalized linear models, quasi-likelihood

1 Introduction

Transcriptional profiling with RNA sequencing (RNA-seq) is widely used to study gene expression profiles associated with a particular biological condition. A common aim of RNA-seq studies is to detect differentially expressed (DE) genes between two or more conditions. To do so, the number of sequencing reads mapped to the exons of each gene is counted to quantify the expression of that gene (Liao et al., 2014). This is performed for multiple replicate libraries in each condition, where each replicate is prepared from an independent biological sample. Statistical analyses can then be performed with methods like `edgeR` (Robinson et al., 2010) to identify genes with significant differences in the read counts between conditions (Anders et al., 2013). These putative DE genes are

the basis for further investigation into the mechanisms driving the biological difference of interest.

The differential expression analysis must be able to handle discrete count data with extra-Poisson variability between biological replicates. To this end, the counts for each gene can be fitted to generalized linear models (GLMs) based on the negative binomial (NB) distribution (McCarthy et al., 2012). Overdispersion in the counts between replicates is modelled with the NB dispersion parameter. Further sophistication can be added with the quasi-likelihood (QL) framework (Lund et al., 2012), which introduces an additional QL dispersion parameter to model estimation uncertainty. The mean-variance relationship of the count y_{gi} for gene g in library i can then be written as

$$\text{var}(y_{gi}) = \sigma_g^2(\mu_{gi} + \mu_{gi}^2\phi_g)$$

where μ_{gi} is the mean, ϕ_g is the NB dispersion and σ_g^2 is the QL dispersion. Lund et al. (2012) used a global abundance-trend to estimate the NB dispersions ϕ_g while the QL dispersions σ_g^2 were allowed to be gene-specific. The small number of replicates means that there is often insufficient information to stably estimate the QL dispersion for each gene using only data from that gene. Imprecision is avoided by using empirical Bayes (EB) methods that share information between genes. These methods stabilize the QL dispersion estimates and improve power to detect differential expression (Phipson et al., 2016).

Fitting a GLM to RNA-seq data involves estimation of real-valued model coefficients from discrete non-negative counts. Consider an RNA-seq data set with n libraries. Assume that a stable estimate of the common or trended NB dispersion has been obtained by using an EB approach across all genes (McCarthy et al., 2012). For each gene, a GLM is fitted to the n counts using the estimated NB dispersion and a design matrix $\mathbf{X}$ that has p independent coefficients. The residual degrees of freedom (d.f.) is canonically defined as

$$d = n - p \, .$$

This represents the number of counts that need to be known in order to derive all n counts, given the estimates of the p coefficients from the fitted model. The total residual deviance of the fitted model for gene g is denoted as D_g, and is distributed as

$$D_g|\sigma_g^2 \sim \sigma_g^2\chi_d^2$$

where σ_g^2 is the QL dispersion for g. This distribution is based on analogous behaviour under normality (Lund et al., 2012). The deviance estimator of σ_g^2 is defined as

$$\hat{\sigma}_g^2 = D_g/d \, .$$

Thus, correct calculation of the residual d.f. is necessary for QL dispersion estimation in the EB framework described by Lund et al. (2012).

The appropriate residual d.f. may not be obvious when zeroes are present in the set of counts for a gene. Any library that has a fitted value of zero must also have a count of zero. This will not contribute any d.f. to the fit, because no additional observations need to be known to identify the count as zero. As such, the true residual d.f. will be lower than the canonical value d. D_g will also be lower than expected, as the unit deviance will be zero for libraries where the fitted value and count are identical. Thus, the above expression for $\hat{\sigma}_g^2$ will be

incorrect. The unit deviance will be exactly zero when the fitted value is zero. This occurs regardless of the true value of σ_g^2, showing that these observations contribute no information to the estimation of σ_g^2.

This paper demonstrates that the use of the canonical residual d.f. can result in underestimation of the QL dispersion and loss of type I error control in the presence of zero counts. A corrected definition of the residual d.f. is given for genes that have fitted values of zero for some libraries. The performance of this solution compares favourably to the canonical defininition on both simulated and real datasets.

2 Correctly specifying the residual d.f.

2.1 Definition of the reduced residual d.f.

Let $\hat{\mu}_{gi}$ denote the fitted value for library i in the GLM for this gene. Assume that there is a set of libraries Z_g for which $\hat{\mu}_{gi} = 0$ in each library $i \in Z_g$. For each of these libraries, the count y_{gi} must also be zero as all counts are non-negative and any positive count would result in $\hat{\mu}_{gi} > 0$. Knowing that $\hat{\mu}_{gi} = 0$ means that $y_{gi} = 0$, which can be determined without the need to know any other count. Thus, libraries in Z_g will not contribute any d.f. to the fit. The canonical definition for the residual d.f. fails to consider this subtlety. Consequently, the computed value for d may be larger than the true residual d.f. The latter is denoted as d_g and will be referred to as the "reduced residual d.f." in the following text.

The value of d_g can be determined by identifying and ignoring the libraries in Z_g. In practice, this is done by removing all libraries with fitted values below some arbitrarily small value like 10^{-4}. This ensures that libraries in Z_g are not overlooked due to numerical imprecision of the model fit. The effective number of libraries for gene g becomes $n - |Z_g|$. Similarly, a refined design matrix $\mathbf{X}_g$ can be constructed by removing all rows in $\mathbf{X}$ corresponding to $i \in Z_g$. When $\mathbf{X}_g$ is fitted to the counts for the remaining libraries, the reduced residual d.f. is defined as

$$d_g = n - |Z_g| - \text{rank}(\mathbf{X}_g)$$

where the column rank of $\mathbf{X}_g$ is simply the number of independent coefficients in $\mathbf{X}_g$. This can be obtained by performing a QR decomposition on $\mathbf{X}_g$ and counting the number of non-zero diagonal elements in the resulting R matrix. The above expression for d_g is appropriate as it explicitly ignores the libraries that do not contribute any d.f. to the model fit. This avoids any overestimation of the residual d.f. that might occur with the canonical definition. Of course, $d_g = d$ if Z_g is empty as the rank of $\mathbf{X}_g$ is equal to p.

2.2 Overview of the canonical QL framework

Lund et al. (2012) assume that $\sigma_g^2 \sim \sigma_0^2 \chi_{d_0}^{-2}$ across all genes g, where d_0 is the prior d.f. and σ_0^2 is a constant scaling factor. This means that $\hat{\sigma}_g^2 \sim \sigma_0^2 F(d, d_0)$. Both d_0 and σ_0^2 can be simultaneously estimated from the distribution of $\hat{\sigma}_g^2$ across all g (Smyth, 2004; Phipson et al., 2016). The shrunken dispersion in the

EB framework is defined as

$$\tilde{\sigma}_g^2 = \frac{d_0\sigma_0^2 + d\hat{\sigma}_g^2}{d_0 + d} ,$$

which effectively squeezes $\hat{\sigma}_g^2$ towards the estimated σ_0^2 for each gene.

The QL F-test is used to test a pre-specified null hypothesis by comparing nested designs. One or more columns are chosen and removed from $\mathbf{X}$ to obtain the null design matrix $\mathbf{X}_0$. The moderated F-statistic uses the shrunken dispersion and is defined as

$$F_g = \frac{\mathrm{LR}_g/t}{\tilde{\sigma}_g^2}$$

where t is the difference in the number of coefficients between designs and LR_g is the likelihood ratio, i.e., the difference in the total residual deviances between GLMs fitted with $\mathbf{X}_0$ and $\mathbf{X}$. Under the null hypothesis, F_g will be F-distributed on t and $d + d_0$ d.f.'s. This can be used to compute a p-value to reject or accept the null for each gene. In practice, a lower bound for the p-value is defined by using the likelihood ratio test after fitting a Poisson GLM to the counts for each gene. This reflects the fact that RNA-seq data should exhibit at least Poisson-level variance due to sequencing noise (Marioni et al., 2008).

2.3 Redefined statistics with the true residual d.f.

The canonical definitions are only correct for gene g when $d = d_g$. If these two values are not equal, all occurrences of d should be replaced with d_g. For simplicity, assume that all libraries $i \in Z_g$ have true means close to zero. This is reasonable in most replicated designs, where a fitted value of zero from a large true mean would require zero counts for all replicate observations (which is unlikely). The assumption means that the libraries in Z_g do not contribute to any sampled instance of D_g, i.e., the unit deviance from each library will always be zero. Thus, $D_g|\sigma_g^2 \sim \sigma_g^2\chi_{d_g}^2$ as only those libraries contributing d.f. are considered. The correct estimator of the QL dispersion for gene g is defined as

$$\hat{\sigma}_{g(2)}^2 = D_g/d_g .$$

This expression ensures that $E(\hat{\sigma}_{g(2)}^2|\sigma_g^2)$ is still equal to σ_g^2 when Z_g is not empty. EB shrinkage should be performed using the distribution of $\hat{\sigma}_{g(2)}^2$ across all genes, instead of $\hat{\sigma}_g^2$. This ensures that the correct scaling factor $\sigma_{0(2)}^2$ and prior d.f. $d_{0(2)}$ are estimated. The shrunken dispersion for each gene is similarly redefined as

$$\tilde{\sigma}_{g(2)}^2 = \frac{d_{0(2)}\sigma_{0(2)}^2 + d_g\hat{\sigma}_{g(2)}^2}{d_{0(2)} + d_g}$$

while the moderated F-statistic is redefined as

$$F_{g(2)} = \frac{\mathrm{LR}_g/t}{\tilde{\sigma}_{g(2)}^2} .$$

This is F-distributed on t and $d_g + d_{0(2)}$ d.f.'s, and can be used to compute a p-value as previously described.

3 Assessing performance on simulated data

3.1 The canonical method underestimates the dispersion

Consider a one-way layout with two replicate libraries in each of four conditions A_1, A_2, B_1 and B_2. NB-distributed counts for 10000 genes were generated using a dispersion of 0.05 and a mean of μ_c for each library in condition c. For the first 5000 genes, $\mu_{A_1} = \mu_{A_2} = 0$ and $\mu_{B_1} = \mu_{B_2} = 200$, while for the last 5000 genes, $\mu_{A_1} = \mu_{A_2} = 200$ and $\mu_{B_1} = \mu_{B_2} = 0$. This simulation design ensures that each library has a non-zero total number of reads for downstream normalization. (A constant value of 200 is only chosen for simplicity, and can be replaced with other non-zero values specific to each gene without affecting the results.) Methods in the `edgeR` package v3.16.5 were then used to estimate a common NB dispersion across all genes and to fit a NB GLM to the counts for each gene. QL dispersion estimates $\hat{\sigma}_g^2$ and $\hat{\sigma}_{g(2)}^2$ were computed for each gene as previously described, and the mean and standard error of each value was computed over all genes.

The expected value of the QL dispersion in this simulation is $\sigma_g^2 = 1$ for each gene. This is because counts are exactly NB-distributed such that no QL-based modification of the mean-variance relationship is required. Estimation of the mean $\hat{\sigma}_g^2$ across all genes yields a value of 0.508 whereas the mean $\hat{\sigma}_{g(2)}^2$ is 1.016, with negligible standard errors in both cases. The latter is closer to the expected value for σ_g^2, indicating that the reduced definition is correct. This difference in behaviour is due to the presence of zero counts such that $d_g = 2$. In contrast, $d = 4$ as one d.f. is provided from each condition in the experimental design. This drives the two-fold underestimation of the QL dispersion across all genes when d is used.

3.2 Assessing type I error control

3.2.1 Details of the simulation design

Consider a one-way layout with two replicates in each of four conditions for 10000 genes, similar to that described in Section 3.1. A set K was defined containing 5 to 100% of all genes. For half of all genes in K, zero counts were added by setting $\mu_{A_1} = \mu_{A_2} = 0$ and $\mu_{B_1} = \mu_{B_2} = 200$. For the other half, $\mu_{A_1} = \mu_{A_2} = 200$ and $\mu_{B_1} = \mu_{B_2} = 0$. Otherwise, for genes not in K, $\mu_c = 100$ for each library in all conditions. This setup ensures that library sizes are always non-zero for downstream normalization. Values of μ_c were also chosen such that the expected average count over all libraries was constant for all genes. This ensures that any changes in the observed error rate are not simply driven by changes in overall abundance when different proportions of genes with zero counts are added. NB-distributed counts were generated for each gene using the condition-specific means. The NB dispersion for each gene was sampled from an inverse chi-squared distribution on 20 d.f. to simulate variable dispersions.

The null hypothesis of $\mu_{B_1} = \mu_{B_2}$ was tested for most genes, using the QL framework in `edgeR` with the canonical and reduced residual d.f.'s. This hypothesis was chosen as it is true for all genes, meaning that the resulting p-value distribution can be used to assess type I error control. For genes where $\mu_{B_1} = \mu_{B_2} = 0$, the null hypothesis of $\mu_{A_1} = \mu_{A_2}$ was tested instead. This avoids a trivial result when the counts for all libraries in B_1 and B_2 are fixed

Table 1: Observed type I error rates at a range of specified thresholds in the four-condition simulation, using the canonical or reduced residual d.f. in the QL framework. Each simulation scenario has a different proportion of genes with zero counts. For each error rate, the mean was computed over 20 simulation iterations and is shown as a percentage (standard error in brackets).

Prop. (%)	**Method**	**Type I error threshold (%)**		
		0.1	*1*	*10*
5	Canonical	0.0615 (0.0071)	0.9070 (0.0203)	10.2590 (0.0690)
	Reduced	0.0970 (0.0098)	0.9750 (0.0232)	9.8505 (0.0764)
20	Canonical	0.0720 (0.0060)	1.1585 (0.0251)	12.4570 (0.0765)
	Reduced	0.1120 (0.0072)	1.0055 (0.0212)	10.0155 (0.0591)
50	Canonical	0.2245 (0.0099)	2.5935 (0.0373)	17.6345 (0.0931)
	Reduced	0.0975 (0.0064)	0.9650 (0.0183)	9.7920 (0.0769)
100	Canonical	1.2255 (0.0325)	7.0730 (0.0609)	28.0380 (0.1127)
	Reduced	0.1200 (0.0115)	1.0105 (0.0379)	10.0040 (0.0979)

at zero. The observed type I error rate was defined as the proportion of genes with p-values below a specified threshold. Multiple simulation iterations were performed and the mean error rate was calculated, along with its standard error. Mean estimates and standard errors for σ_0^2 or $\sigma_{0(2)}^2$ and d_0 or $d_{0(2)}$ were also computed across iterations.

3.2.2 The canonical method fails to control type I error

In all tested scenarios, the observed error rate for the reduced residual d.f. is closer to the threshold than that for the canonical d.f. (Table 1). This is consistent with the correctness of the reduced d.f. and its associated statistics when zero counts are present.

Loss of type I error control for the canonical method is observed in some scenarios. This is attributable to underestimation of the QL dispersion, which inflates the F-statistic and decreases the p-value for each gene in K. Moreover, this error is propagated to all genes through EB shrinkage. Smaller values of $\hat{\sigma}_g^2$ for $g \in K$ will drag down the estimated scaling factor σ_0^2. Indeed, increasing the size of K results in a drop in the estimated σ_0^2 in the simulations (Table 2). Thus, even the correct $\hat{\sigma}_g^2$ for genes not in K will be shrunk towards an inappropriately small estimate for σ_0^2. Subsequent underestimation of $\tilde{\sigma}_g^2$ inflates the moderated F-statistic for all genes in the analysis.

The canonical method is also conservative in some scenarios. This is driven by an increase in the variability of the dispersions when zero counts are present. Recall that the QL dispersions for $g \in K$ are consistently underestimated. This results in a population of dispersion estimates that is distinct from those for $g \notin K$, inflating the apparent variability of $\hat{\sigma}_g^2$. In the EB framework, this manifests as a decrease in the estimated prior d.f. when more genes are present with zero counts (Table 2). Smaller d_0 has a number of consequences, the most obvious of which is the decrease in the second d.f. of the QL F-test. This leads to larger p-values as the variance of F_g will be overestimated.

Table 2: EB statistics for the four-condition simulation with differing proportions of genes with zero counts in two conditions. The estimated scaling factor and prior d.f. were computed using the canonical and reduced definitions for the residual d.f. All values represent the mean over 20 simulation iterations. Standard errors are also shown in brackets.

Prop. (%)	**Scaling factor**		**Prior d.f.**	
	σ_0^2	$\sigma_{0(2)}^2$	d_0	$d_{0(2)}$
5	0.88 (0.00)	0.95 (0.00)	15.2 (0.3)	30.1 (1.1)
20	0.70 (0.00)	0.94 (0.00)	6.2 (0.1)	28.8 (1.2)
50	0.45 (0.00)	0.94 (0.00)	3.1 (0.0)	27.2 (2.0)
100	0.25 (0.00)	0.94 (0.00)	2.5 (0.0)	33.1 (6.1)

The overall effect of using the canonical d.f. on the type I error rate is not easy to predict. The outcome depends on whether the conservativeness from the reduced d_0 can offset the liberalness from the underestimation of σ_g^2 and σ_0^2. Liberalness will also be introduced by the use of the larger d instead of d_g to compute the second d.f. of the QL F-test. In addition, the shrunken dispersion will change as d_0 decreases, though the overall effect is unclear as $\tilde{\sigma}_g^2$ will increase for some genes and decrease for others. This unpredictability can be avoided by using the reduced residual d.f. to correctly estimate $d_{0(2)}$ and $\sigma_{0(2)}^2$.

3.3 Care is required when the reduced residual d.f. is zero

Some additional care is required when dealing with genes where $d_g = 0$. These genes provide no information on the variability of the counts and have undefined $\hat{\sigma}_{g(2)}^2$. This means that they cannot be used to estimate $\sigma_{0(2)}^2$ or $d_{0(2)}$. However, the null hypothesis can still be tested for these genes. As $d_g = 0$, the expression for the shrunken dispersion $\tilde{\sigma}_{g(2)}^2$ simply collapses to $\sigma_{0(2)}^2$. This means that $F_{g(2)}$ and a p-value can be computed.

To demonstrate, consider a one-way layout containing three conditions A, B_1 and B_2 for 10000 genes. Condition A contains two replicates whereas B_1 and B_2 contain one replicate each. For the first 50% of genes, $\mu_A = 0$ and $\mu_{B_1} = \mu_{B_2} = 200$. This simulates non-DE genes between B_1 and B_2 for which there are no residual d.f.'s. For the remaining genes, $\mu_A = 200$ and $\mu_{B_1} = \mu_{B_2} = 0$. This setup ensures that all library sizes are non-zero and that $\sigma_{0(2)}^2$ and $d_{0(2)}$ can be estimated from genes with defined QL dispersions. Counts for each library were sampled from a NB distribution with the specified mean. The QL framework was applied using the reduced residual d.f., and the null hypothesis $\mu_{B_1} = \mu_{B_2}$ was tested for the first 50% of genes. The observed type I error rate was computed at a nominal threshold of 1%. This was repeated for 20 simulation iterations, yielding a mean error rate of 0.91% with a standard error of 0.09. Similar results were obtained at thresholds of 0.1% ($0.14 \pm 0.03\%$) and 10% ($9.72 \pm 0.19\%$). Error rates close to the nominal threshold indicate that the QL framework remains valid when $d_g = 0$.

3.4 Effect on log-transformed counts in linear models

3.4.1 Overview

An alternative approach to analyzing count data involves fitting a linear model to log-CPM values (Law et al., 2014). Consider an example data set where all libraries have 10^6 reads. Any zeroes will be converted into a lower bound for the log-CPM, e.g., $\log_2(0 + 0.5) = -1$ where the 0.5 represents a continuity correction. If any fitted value of the linear model is equal to the lower bound, the corresponding log-CPM must also be equal to the lower bound, as no smaller log-CPM can exist. Thus, no d.f. will be provided by those libraries, which means that the canonical definition of the residual d.f. may be incorrect. This can be remedied for each gene by removing all libraries with fitted values equal to the lower bound. Linear modelling and variance estimation will then be performed using the reduced d.f. as described above for GLMs.

To demonstrate, NB-distributed counts for a four-condition design were simulated as described in Section 3.1. The voom function was applied to log-transform the counts and to compute precision weights from a fitted mean-variance trend. A linear model was fitted to the log-CPM values with the precision weights using `limma` v3.30.9, and an EB strategy was applied to robustly shrink the variances (Phipson et al., 2016). The mean variance estimate across all genes was recorded. The F-test was applied to compute p-values against the null hypothesis, i.e., $\mu_{B_1} = \mu_{B_2}$ for the first 5000 genes, and $\mu_{A_1} = \mu_{A_2}$ for the last 5000 genes (again, this distinction avoids a trivial result when means and counts are fixed at zero for some libraries). As the null is always true, the distribution of p-values across all genes should be uniform. The analysis was repeated after removing libraries with $\mu_c = 0$ from the dataset and design matrix prior to fitting, i.e., all libraries in A_1 and A_2 for the first 5000 genes, or in B_1 and B_2 for the last 5000 genes.

The mean sample variance across all genes was estimated as 0.065 and 0.132 using the canonical and reduced definitions, respectively. This fold difference is consistent with the differences in the residual d.f.'s, i.e., $d = 4$ and $d_g = 2$. Type I error control was subsequently lost with the canonical method (Figure 1), consistent with the liberalness observed in simulations for NB-based GLMs. Uniformity of the p-values was restored by removing the libraries in the offending conditions. This ensured that the reduced residual d.f. was used during modelling. In practice, removal of offending libraries is complicated by the presence of variable library sizes, resulting in a variable lower bound that is difficult to define for any given library. For simplicity, such cases will not be considered here.

4 Effect of d.f. specification on real data

4.1 Overview

To determine the relevance of the simulation results, analyses with the canonical and reduced residual d.f.'s were compared on a real RNA-seq dataset. Data was generated from a *Pax5* knock-out experiment in pro-B cells by Drs. Rhys Allan and Steve Nutt in the Molecular Immunology division of the Walter and Eliza Hall Institute of Medical Research. This dataset contains two replicate libraries

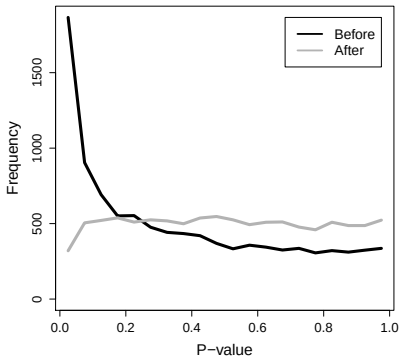


Figure 1: Histograms of p-values generated by voom and `limma` on simulated count data, before and after removal of libraries with true means of zero.

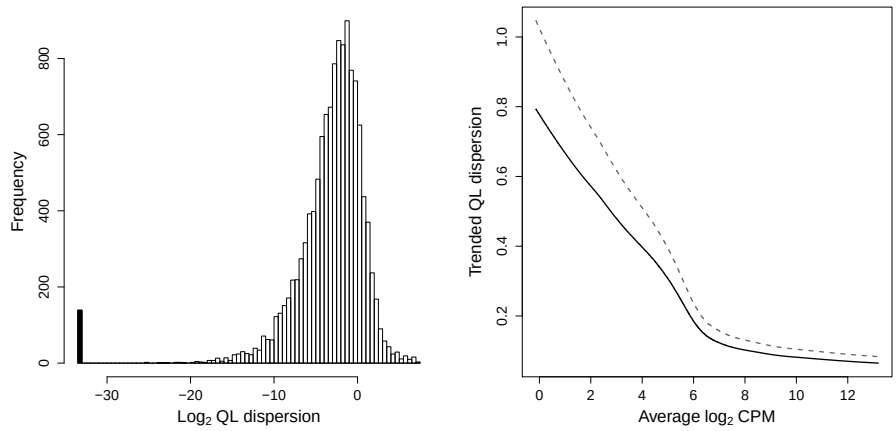
for the knock-out (KO) condition and one replicate for the wild-type (WT) condition.

4.2 Processing steps for real RNA-seq data

Paired-end sequencing was performed by the Beijing Genomics Institute on an Illumina HiSeq 2000. Each library consisted of approximately 9 million pairs of 90 bp reads. Reads were aligned to the mm10 build of the mouse genome using `subread` v1.4.6 (Liao et al., 2013) in paired-end mode with unique mapping and tie splitting by Hamming distance. Read pairs were summarized into gene counts using the featureCounts function (Liao et al., 2014) in the `Rsubread` package v1.18.1. The number of fragments mapped to exons was counted for each gene in the NCBI mouse build 38 annotation. Note that each read pair corresponds to a single cDNA fragment and is counted no more than once. Reads with MAPQ scores below 10 were ignored to avoid non-uniquely/poorly mapped reads. Approximately 64% of read pairs in each library were counted into genes.

Genes were filtered to remove those with a count sum across all libraries below 10. This removes lowly expressed genes that are unlikely to be DE and is roughly independent of the p-values when library sizes are similar. A differential expression analysis was performed using `edgeR` to compare the WT and KO conditions. Briefly, a trended NB dispersion was estimated and used for GLM fitting. Offsets were defined from the log-transformed effective library sizes, after using TMM normalization (Robinson and Oshlack, 2010) to remove composition biases. Raw QL dispersions were estimated with the canonical residual d.f., as previously described. A second mean-dependent trend was fitted to the raw QL estimates, and EB shrinkage was performed towards this trend (Lund et al., 2012). Trend fitting is necessary here to empirically model non-NB mean-variance relationships, but was not required for the simulations where counts were exactly NB-distributed for simplicity. Finally, the QL F-test was used to compute a p-value for each gene.

Genes were considered to be significantly DE if the false discovery rate (FDR) was $< 5\%$ after applying the Benjamini and Hochberg method to all p-values.



(a) (b)

Figure 2: QL dispersion estimates for the Pax5 dataset. (a) Histogram of log-transformed gene-wise estimates using the canonical d.f. A small prior was added to avoid undefined values. The black bar marks genes with a reduced residual d.f. of zero. (b) Loess-fitted trend in the estimates for all genes, using the average abundance as the covariate. Results are shown before (full) and after (dashed) removing genes with no residual d.f.

This was repeated using the reduced residual d.f. and its associated statistics.

4.3 Differences between definitions are present in real data

The QL framework was applied with the reduced and canonical residual d.f.'s to identify DE genes between WT and KO cells. For the canonical analysis, a cluster of low dispersion estimates was observed (Figure 2a). These correspond to 139 genes that have zero counts in the two KO samples. Here, there are no residual d.f. so the deviance is always zero. As $d = 1$, $\hat{\sigma}_g^2$ is incorrectly defined as zero for these genes. In comparison, $\hat{\sigma}_{g(2)}^2$ is undefined when $d_g = 0$ and should not participate in EB shrinkage. Removal of the corresponding genes increased the estimated prior d.f. from 2.71 to 5.17 and increased the estimated QL dispersions (Figure 2b). This is consistent with inflation of heteroskedasticity and underestimation of the QL dispersion when d is used instead of d_g in the simulations.

Hypothesis testing was also performed to identify DE genes using the reduced and canonical methods. At a FDR of 5%, the number of DE genes increased from 1926 to 2809 when d_g was used instead of d. Reduced detection with the canonical d.f. is consistent with the conservativeness observed in Table 1, where the drop in d_0 outweighs the liberalness caused by the underestimation of σ_g^2 and σ_0^2. Note that the 139 genes with undefined $\hat{\sigma}_g^2$ were identified as DE in both the reduced and canonical methods. This is not unexpected, given that the counts are zero in the KO samples and equal to or greater than 10 in the WT sample. However, it does illustrate that incorrect estimation of the dispersion for some genes will affect inferences for the entire data set. Almost 900 DE genes were

uniquely detected when the reduced method was used instead of the canonical method, despite the fact that all of them have correctly defined $d = d_g = 1$. This unique set includes genes such as *Tnfaip3* (adjusted p-value of 0.035), *Id2* (0.023) and *Nfatc2* (0.021) that have been implicated in B-cell differentiation (Chu et al., 2011; Becker-Herman et al., 2002; Peng et al., 2001). Such genes would not have been identified as putative downstream targets of Pax5 if the canonical method was used.

5 Differences are recapitulated in realistic simulations

The previous simulations were intentionally simplified to highlight the effects of misspecifying the residual d.f. To demonstrate that such effects were still present in realistic scenarios, additional simulations were performed based on the *Pax5* data. For each gene g, the trended NB dispersion ϕ_g and the average count were estimated from the *Pax5* counts using `edgeR` as described above. This recapitulates the distribution of abundances and the mean-dispersion relationship present in real data. A new value for the dispersion was obtained by sampling from $\phi_g \nu / \chi^2_\nu$, where $\nu = 5$ based on the prior d.f. estimated above. (This mimics the spread of gene-specific dispersions around the trend in real data.) The sampled dispersion and average count were set as the parameters of a NB distribution, from which one count was sampled for each library to obtain a gene-specific expression profile. This was repeated for each gene to generate a simulated data set.

Two experimental designs were considered in this simulation – a one-way layout containing four groups (A_1, A_2, B_1, B_2) of two replicates each, and a paired-samples design containing four pairs containing one sample from each of two groups (A_1, A_2). To introduce fitted values of zero to the simulation, half of all genes were allocated into the set K. For each gene in K, counts were set to zero for all samples in B_1 and B_2 in the one-way layout, or for all samples in two pairs in the paired-samples design. The QL framework was then applied to test the null hypothesis of equal expression in A_1 and A_2. This motivates the choice of samples to set to zero, as it ensures that the null hypothesis is still true for all genes. The observed type I error rate was estimated as the proportion of genes with p-values below the specified threshold. This was done separately for genes in and outside of K, to determine the effect of misspecified residual d.f. on each class of genes. The large size of K ensures that stable estimates of the observed type I error rates can be obtained.

The use of the reduced d.f. in the QL framework controlled the observed type I error rate close to or below the specified threshold in all simulation scenarios (Table 3). This is consistent with the proper specification of the residual d.f. in the presence of fitted values of zero. In contrast, use of the canonical d.f. yielded liberal results for all genes in K. This is attributable to underestimation of the QL dispersion when the residual d.f. is overstated. More subtle loss of type I error control was also observed for genes outside of K at some thresholds, caused by distortion of the EB statistics when the QL dispersions in K are incorrectly estimated. These results indicate that the advantages provided by the reduced d.f. are still present in realistic scenarios.

Table 3: Observed type I error rates at a range of thresholds in simulations based on real data, using the canonical or reduced d.f. in the QL framework. Simulations were performed using a one-way layout or a paired-samples design. Separate error rates are shown for genes in and outside K. Each value represents the mean error rate across 20 simulation iterations, with the standard error shown in brackets.

Design	**Method**	**Type I error threshold (%)**		
		0.1	*1*	*10*
One-way	Canonical (in K)	0.648 (0.019)	4.249 (0.062)	21.683 (0.114)
	Canonical (not in K)	0.085 (0.008)	1.202 (0.028)	12.302 (0.110)
	Reduced (in K)	0.122 (0.010)	0.953 (0.022)	9.450 (0.088)
	Reduced (not in K)	0.096 (0.009)	0.945 (0.029)	9.958 (0.098)
Paired	Canonical (in K)	1.998 (0.045)	9.715 (0.107)	31.449 (0.199)
	Canonical (not in K)	0.095 (0.007)	1.313 (0.023)	12.659 (0.080)
	Reduced (in K)	0.130 (0.012)	0.970 (0.041)	9.390 (0.084)
	Reduced (not in K)	0.114 (0.009)	0.938 (0.024)	9.437 (0.094)

The loss of residual d.f. in the paired-samples design warrants some further discussion. In the case where both observations in a pair are zero, the corresponding fitted values would obviously be zero (i.e., the coefficient for the blocking term for this pair in a log-link model would approach negative infinity). Neither observation would provide any residual d.f., which is not considered by the canonical definition. Another example involves pairs of control/treatment samples where the treatment sample has a count of zero in all pairs. In this scenario, the coefficient for the treatment effect would approach negative infinity, such that all treatment samples would have fitted values of zero. However, the coefficients for the pair-specific blocking terms would still be free to vary to fit the control sample in each pair. This means that there are no residual d.f. for dispersion estimation, regardless of the number of pairs. Both situations are handled properly by the reduced definition of the residual d.f., as demonstrated above.

6 Discussion

This article has shown that the standard formulation for the residual d.f. in a linear model is not correct when NB GLMs are used to model RNA-seq read count data and fitted values of zero are present. The incorrect formulation will result in underestimation of the QL dispersion and potential loss of type I error control. Such problems can be avoided by using a refined definition for the residual d.f. that accounts for the presence of zeroes. The new reduced d.f. formula restores error control in the QL framework for simulated count data. Similar behaviour is observed in linear models for log-CPMs and in a DE analysis of a real RNA-seq dataset. While this work focuses on RNA-seq data, similar conclusions can be drawn for analyses of read counts from other genomic technologies such as ChIP-seq (Lun and Smyth, 2016) or Hi-C (Lun and Smyth, 2015).

The results presented here have much wider implications for the application of GLMs in fields other than genomics and genetics. The reduced d.f. formula derived here will be a more appropriate definition of the residual d.f. for any GLM when the fitted values are non-negative but exactly zero fitted values are possible. This scenario can arise, for example, when conducting goodness of fit tests for Poisson or multinomial GLMs. The reduced d.f. will be appropriate for any count-based GLM when a quasi-dispersion parameter or parameters in the variance function need to be estimated (Wedderburn, 1974). Exactly zero fitted values can also arise when using Tweedie GLMs to model insurance claims (Smyth and Jørgensen, 2002) or rainfall data (Dunn and Smyth, 2005).

The reduced d.f. method described in this article has been implemented in the glmQLFit function of the `edgeR` package, available from the open-source Bioconductor project.

Acknowledgements

The authors would like to thank Dr. Yunshun Chen for assistance with incorporating the method into `edgeR` and Stephen Nutt and Rhys Allan for permission to use the pro-B cell RNA-seq data.

This work was funded by the National Health and Medical Research Council (Program Grant 1054618 to G.K.S., Fellowship to G.K.S.), by Victorian State Government Operational Infrastructure Support and by Australian Government NHMRC IRIIS.

References

Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson (2013): "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor," *Nat. Protoc.*, 8, 1765–1786.

Becker-Herman, S., F. Lantner, and I. Shachar (2002): "Id2 negatively regulates B cell differentiation in the spleen," *J. Immunol.*, 168, 5507–5513.

Chu, Y., J. C. Vahl, D. Kumar, K. Heger, A. Bertossi, E. Wojtowicz, V. Soberon, D. Schenten, B. Mack, M. Reutelshofer, R. Beyaert, K. Amann, G. van Loo, and M. Schmidt-Supprian (2011): "B cells lacking the tumor suppressor TNFAIP3/A20 display impaired differentiation and hyperactivation and cause inflammation and autoimmunity in aged mice," *Blood*, 117, 2227–2236.

Dunn, P. and G. Smyth (2005): "Series evaluation of Tweedie exponential dispersion model densities," *Statistics and Computing*, 15, 267–280.

Law, C. W., Y. Chen, W. Shi, and G. K. Smyth (2014): "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biol.*, 15, R29.

Liao, Y., G. K. Smyth, and W. Shi (2013): "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote," *Nucleic Acids Res.*, 41, e108.

Liao, Y., G. K. Smyth, and W. Shi (2014): "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, 30, 923–930.

Lun, A. T. L. and G. K. Smyth (2015): "diffHic: a bioconductor package to detect differential genomic interactions in Hi-C data," *BMC Bioinformatics*, 16, 258.

Lun, A. T. L. and G. K. Smyth (2016): "csaw: a bioconductor package for differential binding analysis of ChIP-seq data using sliding windows," *Nucleic Acids Research*, 44, e45.

Lund, S. P., D. Nettleton, D. J. McCarthy, and G. K. Smyth (2012): "Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates," *Stat. Appl. Genet. Mol. Biol.*, 11(5), Article Number 8.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008): "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res.*, 18, 1509–1517.

McCarthy, D. J., Y. Chen, and G. K. Smyth (2012): "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Res.*, 40, 4288–4297.

Peng, S. L., A. J. Gerth, A. M. Ranger, and L. H. Glimcher (2001): "NFATc1 and NFATc2 together control both T and B cell activation and differentiation," *Immunity*, 14, 13–20.

Phipson, B., S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth (2016): "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression," *Annals of Applied Statistics*, 10, 946–963.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010): "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 26, 139–140.

Robinson, M. D. and A. Oshlack (2010): "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biol.*, 11, R25.

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

Smyth, G. K. and B. Jørgensen (2002): "Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling," *Astin Bulletin*, 32, 143–157.

Wedderburn, R. W. M. (1974): "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method," *Biometrika*, 61, 439–447.