The Annals of Applied Statistics
2016, Vol. 10, No. 2, 946–963
DOI: 10.1214/16-AOAS920

ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION[1]

BY BELINDA PHIPSON*, STANLEY LEE[†,‡], IAN J. MAJEWSKI[†,‡], WARREN S. ALEXANDER[†,‡] AND GORDON K. SMYTH[†,‡]

Murdoch Childrens Research Institute, The Walter and Eliza Hall Institute of Medical Research*[†] *and The University of Melbourne*[‡]

One of the most common analysis tasks in genomic research is to identify genes that are differentially expressed (DE) between experimental conditions. Empirical Bayes (EB) statistical tests using moderated genewise variances have been very effective for this purpose, especially when the number of biological replicate samples is small. The EB procedures can, however, be heavily influenced by a small number of genes with very large or very small variances. This article improves the differential expression tests by robustifying the hyperparameter estimation procedure. The robust procedure has the effect of decreasing the informativeness of the prior distribution for outlier genes while increasing its informativeness for other genes. This effect has the double benefit of reducing the chance that hypervariable genes will be spuriously identified as DE while increasing statistical power for the main body of genes. The robust EB algorithm is fast and numerically stable. The procedure allows exact small-sample null distributions for the test statistics and reduces exactly to the original EB procedure when no outlier genes are present. Simulations show that the robustified tests have similar performance to the original tests in the absence of outlier genes but have greater power and robustness when outliers are present. The article includes case studies for which the robust method correctly identifies and downweights genes associated with hidden covariates and detects more genes likely to be scientifically relevant to the experimental conditions. The new procedure is implemented in the limma software package freely available from the Bioconductor repository.

1. Introduction. Modern genomic technologies such as microarrays and RNA sequencing have made it routine for biological researchers to measure gene expression on a genome-wide scale. Researchers are able to measure the expression level of every gene in the genome in any set of cells chosen for study under specified treatment conditions. This article focuses on one of the most common

Received October 2014; revised December 2015.

[1]Supported in part by the University of Melbourne (Ph.D. scholarship to BP), by the National Health and Medical Research Council (Fellowship 1058892, Program Grant 1054618 and the IRIISS) and by a Victorian State Government OIS grant.

Key words and phrases. Empirical Bayes, outliers, robustness, gene expression, microarrays.

analysis tasks, which is to identify genes that are differentially expressed (DE) between experimental conditions.

Gene expression experiments pose statistical challenges because the data are of extremely high dimension, while the number of independent replicates of each treatment condition is often very small. Simply applying univariate statistical methods to each gene in succession can produce imprecise results because of the small sample sizes. Substantial gains in performance can be achieved by leveraging information from the entire dataset when making inference about each individual gene.

Empirical Bayes (EB) is a statistical technique that is able to borrow information in this way [Efron and Morris (1973), Morris (1983), Casella (1985)]. EB has been applied very successfully in gene expression analyses to moderate the genewise variance estimators [Baldi and Long (2001), Wright and Simon (2003), Smyth (2004)]. These articles assume a conjugate gamma prior for the genewise variances and produce posterior variance estimates that are a compromise between a global variance estimate and individual genewise variance estimates. The posterior variance estimators can be substituted in place of the classical estimators into linear model t-statistics and F-statistics. Wright and Simon (2003) and Smyth (2004) derived exact small-sample distributions for the resulting moderated test statistics. They showed that the EB statistics follow classical t and F distributions under the null hypothesis but with augmented degrees of freedom. The additional degrees of freedom of the EB statistics relative to classical statistics represent the information that is indirectly borrowed from other genes when making inference about each individual gene.

EB assumes a Bayesian hierarchical model for the genewise variances, but, instead of basing the prior distribution on prior knowledge as a Bayesian procedure would do, the prior distribution is estimated from the marginal distribution of the observed data. Smyth (2004) developed closed-form estimators for the parameters of the prior distribution from the marginal distribution of the residual sample variances. This procedure is implemented in the limma software package [Ritchie et al. (2015)] and the resulting EB tests have been shown to offer improved statistical power and false discovery rate (FDR) control relative to the ordinary genewise t-tests, especially when the sample sizes are small [Kooperberg et al. (2005), Murie et al. (2009), Ji and Liu (2010), Jeanmougin et al. (2010)]. The limma software has been used successfully in thousands of published biological studies using data from a variety of genomic technologies, especially studies using expression microarrays and RNA-seq.

This article improves the limma EB differential expression tests by robustifying the hyperparameter estimation procedure. As in the original method, we fit genewise linear models to the log-expression values and extract residual variances, but now we give special attention to residual variances that are exceptionally large or exceptionally small. Genes corresponding to extreme variances will be considered "outliers." Following terminology used in the genomics literature, we refer

to outlier genes with large variances as "hypervariable genes." We show that, for certain genomic datasets, a small number of outlier genes can have an undesirable influence on the hyperparameter estimators, decreasing the effectiveness of the EB differential expression procedures. (Figure 1 plots residual standard deviations for three datasets. In each study there are special biological factors that cause a subset of the genes to have larger than expected residual variances.) We show that in such cases the effectiveness of EB can be restored by a robust approach that isolates the outlier genes. We develop a robust estimation scheme with a positive breakdown point for the hyperparameters and incorporate this into the differential expression procedure. The robust EB procedure has the effect of decreasing the informativeness of the prior distribution for hypervariable genes while increasing its informativeness for other genes. This effect has the double benefit of reducing the chance that hypervariable genes will be spuriously identified as DE while increasing statistical power for the main body of genes.

Our robust EB approach uses more diffuse prior distributions for the variances of hypervariable genes. A conjugate prior is still used for each gene and this allows us to preserve a key feature of the original EB differential expression procedures, which is the ability to derive exact small-sample null distributions for the test statistics. Our robust EB approach is fast and numerically stable without difficult convergence issues. It reduces exactly to the original EB procedure when there are no hypervariable genes. Simulation studies show that the robustified tests for differential expression have similar performance to the original tests in the absence of outlier genes but have greater power and robustness when they are present.

To the best of our knowledge, there has been no previous work on robust EB for variances. Most robust EB schemes in other contexts have been based on heavy-tailed prior distributions. We have avoided such an approach because crucial advantages of the original DE procedures would thereby be lost, in particular, the posterior mean variance estimators would no longer be available in closed form and the test statistics would no longer yield exact p-values. Efron and Morris (1972) proposed limited translation rules when estimating means of standard normal distributions. This proposal originated the idea of limited learning for extreme cases, but with the aim of limiting the bias in extreme cases rather than improving estimation of the hyperparameters. Gaver and O'Muircheartaigh (1987) analyzed a Poisson process using a heavy-tailed (log-Student) prior distribution for the Poisson mean. This approach achieves insensitivity to outliers but loses efficiency as well as being less mathematically tractable. Liao, McMurry and Berg (2014) estimated log-fold expression changes and achieved robustness with respect to misspecified working priors by conditioning on the rank of each estimated log-fold change rather than on the actual observation. Again, this is less mathematically tractable than our approach.

Our robustified EB procedure has been implemented in the limma software package and can be invoked by the option `robust=TRUE` in calls to the `eBayes` or `treat` [McCarthy and Smyth (2009)] functions. Invoking the option requires

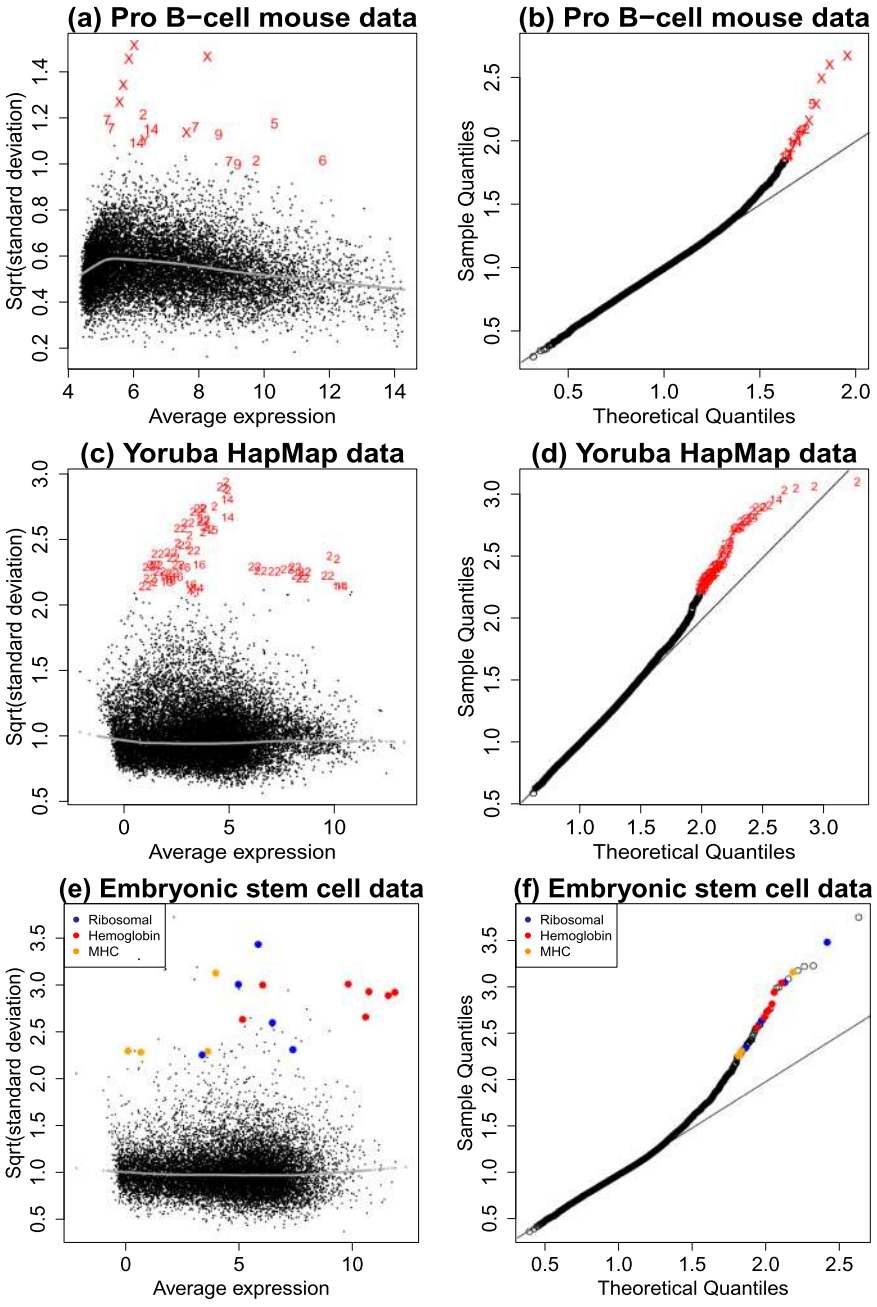

FIG. 1. *Hypervariable genes for the three case study datasets. Left panels plot genewise residual standard deviations vs average* $\log 2$*-expression. Right panels show probability plots of the standard deviations after transforming to equivalent normal deviates. Hypervariable genes are marked by chromosome number in panels* (a)–(d) *and by gene function in panels* (e)–(f). *For each dataset the hypervariable genes have clear biological interpretations.*

no other changes to the analysis pipelines from a user point of view. All downstream functions recognize and work with robust EB results as appropriate.

The following sections review the EB approach to differential expression, then derive the robust estimators and modified differential expression scheme. We evaluate the performance of the robustified procedure using simulations, then present three case studies in which the EB approach identifies genetic instabilities specific to each study. We give a detailed analysis of a microarray study of PRC2 function in pro-B cells for which a gender effect produces hypervariable genes. The EB procedure is effective at downweighting sex-linked genes associated with the unwanted covariate in favor of genes of more scientific interest. Finally, we discuss possible generalizations of our robust EB approach to other contexts.

2. Linear models and moderated *t*-statistics. Consider a genomic experiment in which the expression levels of G genes are measured for n RNA samples. We follow the notation and linear model formulation introduced by Smyth (2004). Write y_{gi} for the log-expression level of gene g in sample i. The log-expression values satisfy genewise linear models

$$E(y_g) = X\beta_g,$$

where y_g is the column vector $(y_{g1}, \ldots, y_{gn})^T$, X is an $n \times p$ design matrix of full column rank representing the experimental design and β_g is an unknown coefficient vector that parametrizes the average expression levels in each experimental condition. For each gene, the y_{gi} are assumed independent with

$$\text{var}(y_{gi}) = \sigma_g^2 / w_{gi},$$

where σ_g^2 is an unknown variance and the w_{gi} are known weights. The least squares coefficient estimator is

$$\hat{\beta}_g = (X^T W_g X)^{-1} X^T W_g y_g,$$

where W_g is the diagonal matrix with elements $w_{g1}, \ldots, w_{gn}$. The residual sample variances are

$$s_g^2 = (y_g - \hat{\mu}_g)^T (y_g - \hat{\mu}_g)/d_g,$$

where $\hat{\mu}_g = X\hat{\beta}_g$ and d_g is the residual degrees of freedom. Usually $d_g = n - p$, but genes with missing y values or zero weights may have smaller values for d_g. Conditional on σ_g^2, $d_g s_g^2/\sigma_g^2$ is assumed to follow a chi-square distribution with d_g degrees of freedom, an assumption we write as

$$s_g^2 | \sigma_g^2 \sim \sigma_g^2 \chi_{d_g}^2 / d_g.$$

We assume a conjugate prior distribution for σ_g^2 in order to stabilize the genewise estimators. The σ_g^2 are assumed to be sampled from a scaled inverse chi-square prior distribution with degrees of freedom d_0 and location s_0^2,

$$\sigma_g^2 \sim s_0^2 d_0 / \chi_{d_0}^2.$$

It follows that the posterior distribution of σ_g^2 given s_g^2 is scaled inverse chi-square,

$$\sigma_g^2 | s_g^2 \sim \frac{d_0 s_0^2 + d_g s_g^2}{\chi^2_{d_0+d_g}},$$

and the posterior expectation of $1/\sigma_g^2$ given s_g^2 is $1/\tilde{s}_g^2$ with

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}.$$

The $\tilde{s}_g^2$ are the EB moderated variance estimators. The moderated t-statistic for a given coefficient β_{ig} is

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_i}},$$

where v_i is the ith diagonal element of $(X^T W_g X)^{-1}$. If the null hypothesis $\beta_{gj} = 0$ is true, then $\tilde{t}_{gj}$ follows a t-distribution on $d_g + d_0$ degrees of freedom [Smyth (2004)]. In general, any ordinary genewise t or F-statistic derived from the linear model can be converted into an EB moderated statistic by substituting $\tilde{s}_g^2$ for s_g^2, in which case the denominator degrees of freedom for the null distribution increase from d_g to $d_0 + d_g$.

3. Robust hyperparameter estimation. Under the above hierarchical model, s_g^2 follows a scaled F-distribution on d_g and d_0 degrees of freedom,

$$s_g^2 \sim s_0^2 F_{d_g, d_0}.$$

The log-variances $\log s_g^2$ follow Fisher's z-distribution, which is roughly symmetric and has finite moments of all orders. The limma package estimates the hyperparameters s_0^2 and d_0 by matching the theoretical mean and variance of the z-distribution to the observed sample mean and variance of the z_g. The empirical estimates s_0^2 and d_0 are then substituted into the above formulas to obtain $\tilde{t}_{gj}$ and to conduct genewise statistical tests for differential expression.

As the observed variance of the $\log s_g^2$ increases, the estimated value of d_0 decreases, meaning that less information is borrowed from the prior to form the moderated t-statistics. In the examples shown in Figure 1, the variance of $\log s_g^2$ would be much reduced if a small number of the most variable genes were excluded. We therefore seek to replace limma's moment estimation scheme for the hyperparameters with a robust version.

Our approach is to apply moment estimation to the Winsorized sample variances. The idea of Winsorizing is to reset a specified proportion of the most extreme sample variances to less extreme values [Tukey (1962)]. Let p_u and p_l be the maximum proportion of outliers allowed in the upper and lower tails of the s_g^2 respectively. Typical values are $p_l = 0.05$ and $p_u = 0.1$, although any values

strictly between 0 and 0.5 are permissable. Let q_l and q_u be the corresponding quantiles of the empirical distribution of s_g^2, so that p_l of the variances are less than or equal to q_l and p_u are greater than or equal to q_u. The empirical Winsorizing transformation is defined by

$$\text{win}(s_g^2) = \begin{cases} q_l & \text{if } s_g^2 \le q_l, \\ s_g^2 & \text{otherwise}, \\ q_u & \text{if } s_g^2 \ge q_u. \end{cases}$$

Write $z_g = \log \text{win}(s_g^2)$ for log-transformed Winsorized variances, and let $\bar{z}$ and s_z^2 be the mean and variance of the observed values of z_g.

Define the Winsorized F-distribution as follows. If $f \sim F_{d_g,d_0}$, then the Winsorized random variable is

$$\text{win}(f) = \begin{cases} q_l & \text{if } f \le q_l, \\ f & \text{otherwise}, \\ q_u & \text{if } f \ge q_u, \end{cases}$$

where now q_l and q_u are the lower tail p_l and upper tail p_u quantiles of the F_{d_g,d_0} distribution.

Write $\nu(d_g, d_0)$ and $\phi(d_g, d_0)$ for the expected value and variance of $\log \text{win}(f)$. An efficient and accurate algorithm for computing ν and ϕ using Gaussian quadrature is described in the Supplementary Methods [Phipson et al. (2016)].

Assuming that the d_g are all equal, the hyperparameter d_0 is estimated by equating $s_z^2 = \phi(d_g, d_0)$ and solving for d_0 using a modified Newton algorithm [Brent (1973)]. Having estimated d_0, the logarithm of the parameter s_0^2 is estimated by $\bar{z} - \nu(d_g, \hat{d}_0)$.

If the d_g are not all equal, then the s_g^2 are transformed to equivalent random variables with equal d_g before applying the above algorithm. Details of this transformation are given in the Supplementary Methods [Phipson et al. (2016)].

4. Gene-specific prior degrees of freedom. Having estimated d_0 and s_0 robustly, we can identify genes that are outliers in that their variances are too large to have reasonably arisen from the estimated prior. The question naturally arises as to how to handle such outliers. One reasonable approach would be to ignore the prior information for such genes, on the basis that the prior appears to be inappropriate. This approach would assign $d_0 = 0$ for such genes, meaning that ordinary t-tests would be used for these genes instead of EB moderated statistics. On the other hand, the prior should in principle still have some limited relevance even for the outliers. Our approach is to assign gene-specific prior degrees of freedom, d_{0g}, whereby $d_{0g} = d_0$ for nonoutlier genes, but outlier genes are assigned smaller values depending on how extreme the outlier is. In effect, we assume that each hypervariable gene g has a true variance σ_g^2 sampled from $s_0^2 d_{0g}/\chi_{d_{0g}}^2$ with $0 < d_{0g} < d_0$.

We start by identifying a lower bound for the d_{0g} from the largest observed s_g^2 value. Specifically, we find d_{outlier} such that the maximum s_g^2 is equal to the median

of the $s_0^2 F_{d_g,d_{\text{outlier}}}$ distribution. A fast stable numerical algorithm for finding d_{outlier} is given in the Supplementary Methods [Phipson et al. (2016)].

Next we evaluate the posterior probability that each gene is a hypervariable outlier. Let p_g be the p-value for testing whether gene g is a outlier, defined by $p_g = P(f > s_g^2/s_0^2)$ where $f \sim F_{d_g,d_0}$. Let π_0 be the prior probability that gene g is not an outlier and let r_g be the marginal probability of observing a residual variance more extreme than s_g^2. The posterior probability, given s_g^2, that case g is not an outlier is $\pi_g = p_g\pi_0/r_g$. Assuming that most genes are not outliers, we conservatively set $\pi_0 = 1$. The marginal probability r_q can be estimated empirically from the rank of s_g^2 among all the observed values of s_g^2, that is, by $r_g = (r - 0.5)/G$ where r is the rank of s_g^2. Substituting these values in the above formula yields a conservative estimate $\pi_g = p_g/r_g$.

The initial estimate of π_g is not necessarily monotonic in s_g^2 or p_g. We ensure that π_g is a nondecreasing function of p_g in the following manner. First the cases are ordered in increasing order of p_g. Then the cumulative mean $\bar{\pi}_g = (1/g)\sum_{i=1}^{g}\pi_i$ is computed for each g. Let g_m be the first value of g for which $\bar{\pi}_g$ achieves its minimum. All π_g for $g = 1, \ldots, g_m$ are set to the minimum value of $\bar{\pi}_g$ to allow for the possibility that π_g might be small for a group of cases but not for the most extreme case. Finally, a cumulative maximum filter is applied to the π_g, after which the π_g are nondecreasing. In practice, this process is nearly identical to using isotonic regression [Barlow et al. (1972)] to enforce monotonicity.

Finally, the genewise prior degrees of freedom are defined by

$$d_{0g} = \pi_g d_0 + (1 - \pi_g)d_{\text{outlier}}.$$

This process ensures that $d_{0g} = d_0$ for most genes, but any gene that is a clear outlier with a very small p_g value will be assigned a much lower value.

5. Covariate dependent priors. In gene expression experiments, the variance of the log-expression values often depends partly on the magnitude of the expression level [Sartor et al. (2006), Law et al. (2014)]. It is therefore helpful to extend the EB principle to permit the prior variance s_0^2 to depend on the average log-expression A_g of each gene. This extension generalizes the prior distribution for σ_g^2 to be gene-specific:

$$\sigma_g^2 \sim s_{0g}^2 \chi_{d_0}^2/d_0,$$

where s_{0g}^2 varies smoothly with A_g. In other words, the prior distribution depends on the covariate A_g. Such a strategy is implemented in the limma package [Ritchie et al. (2015)].

Our strategy for robust EB with a variance trend is as follows. First we fit a robust lowess trend [Cleveland (1979)] to $\log s_g^2$ as a function of A_g. We detrend the $\log s_g^2$ by subtracting the fitted lowess curve, then apply the robust EB algorithm described above to the detrended variances. The final genewise prior values s_{0g}^2 are

the product of the unlogged lowess trend and the s_0^2 estimated from the detrended variances.

6. Software implementation. The robust hyperparameter estimation strategy described above is implemented in the limma function `fitFDistRobustly`. The tail proportions for Winsorizing are user settable with defaults $p_l = 0.05$ and $p_u = 0.1$. This function is called by `squeezeVar`, which computes EB moderated variances. `squeezeVar` in turn is called by user-level functions including `eBayes` and `treat` in the limma package, and `estimateDisp` and `glmQLFit` in the edgeR package [Robinson, McCarthy and Smyth (2010)]. These functions integrate the robust EB strategy into analysis pipelines for gene expression microarrays and RNA-seq. `glmQLFit` is also called in analysis pipelines of the csaw and diffHic packages for the analysis of ChIP-seq and Hi-C sequencing data [Lun and Smyth (2015a, 2016)].

7. Evaluation using simulated data. Simulations were used to evaluate the performance of the robust hyperparameter estimators. Expression values were generated for 10,000 genes and 6 RNA samples. The RNA samples were assumed to belong to two groups, with three in each group, leading to a linear model with $d_g = 4$ residual degrees of freedom. Genewise variances and expression values were generated according to the hierarchical model of Section 2 with $s_0 = 0.2$ and with $d_0 = 2, 4$ or 10. Both the standard and robust hyperparameter estimators were found to be accurate in the absence of outliers [Figures 2(a), (c)]. When 250 hypervariable genes were included, however, the robust estimators were considerably more accurate than the standard [Figures 2(b), (d)]. The hypervariable genes were simulated to have $d_{0g} = 0.5$. Next we checked type I error rates for the EB t-tests in the absence of outliers. Both standard and robust tests were found to control the error rate correctly (Table 1). In all simulations, the tail proportions p_l and p_u were at their default values.

Finally, we evaluated statistical power and FDR control. Each simulated dataset now included 500 DE genes, with log-fold changes generated from a $N(0, 4)$ distribution. In the absence of hypervariable genes, the standard and robust EB tests were indistinguishable in terms of false discoveries or power [Figures 3(a), (c)]. In the presence of 250 outliers, the robust tests consistently gave fewer false discoveries [Figure 3(b)] and higher power [Figure 3(d)]. As expected, the improvement achieved by the robustified tests was greater for larger values of d_0. For simplicity of interpretation, no genes were both DE and hypervariable in these simulations.

8. Case studies.

8.1. *Loss of polycomb repressor complex* 2 *function in pro B cells.* Polycomb group proteins are transcriptional repressors that play a central role in the establishment and maintenance of gene expression patterns during development. Suz12

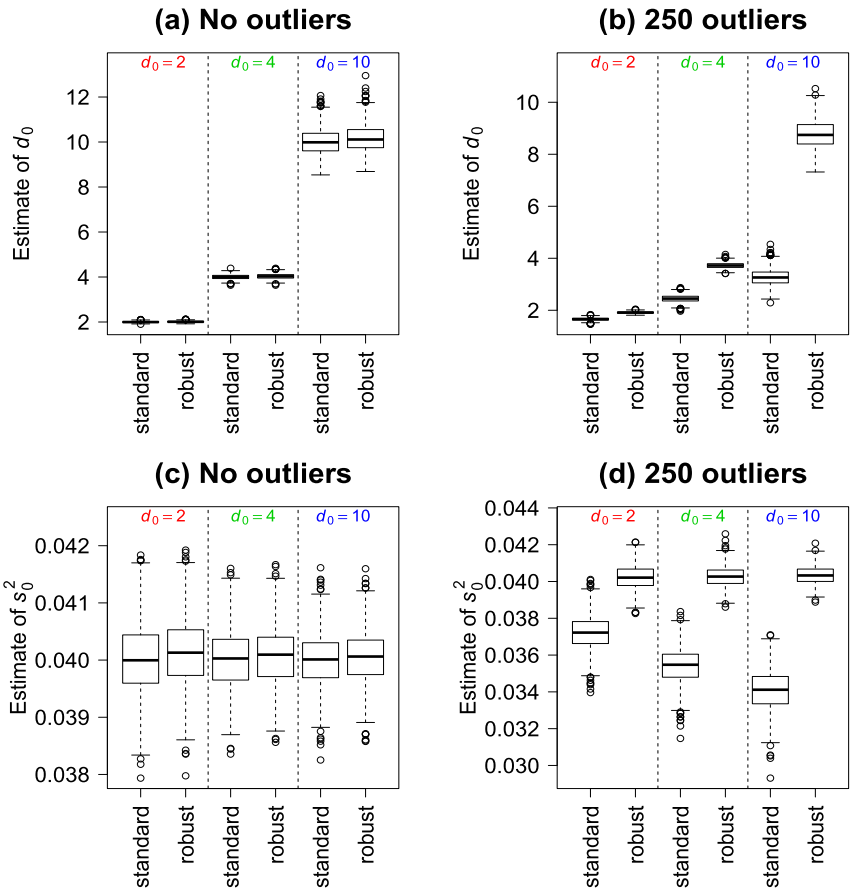

FIG. 2. *Boxplots of standard and robust hyperparameter estimates from* 1000 *simulated datasets. True values are* 0.4 *for* s_0^2 *and either* 2, 4 *or* 10 *for* d_0. *Panels* (a) *and* (c) *show estimates when no outliers are present. Panels* (b) *and* (d) *show estimates when the data includes* 250 *hypervariable genes.*

is a core component of Polycomb Repressive Complex 2 (PRC2). Majewski et al. (2008, 2010) studied mice with a mutation in the *Suz12* gene that results in loss of function of the Suz12 protein and hence PRC2. They profiled gene expression in hematopoietic stem cells from these mice. Here we describe a gene expression study of a different hematopoietic cell type from the same *Suz12* mutant mice strain. This study profiles gene expression in pro-B cells, an early progenitor immune cell intermediate in a series of development stages between hematopoietic stem cells and mature B-cells.

Our interest is to study development, so cells were isolated from 16-day embryonic mice. For this study, RNA was extracted from foetal pro B cells that were isolated from the liver of four wild-type mice and four Suz12 mutant mice. RNA was

TABLE 1
Type I error rates for standard and robust EB t-tests. Datasets were simulated with different d_0 values but with no DE genes and or outliers. The table gives the mean error rate over all genes in 1000 *simulated datasets for various p-value cutoffs*

		Nominal error rate			
d_0	**Method**	**0.001**	**0.01**	**0.05**	**0.1**
2	Standard	0.000996	0.00998	0.05001	0.09994
	Robust	0.000996	0.00998	0.04996	0.09983
4	Standard	0.001008	0.01002	0.05008	0.10012
	Robust	0.001013	0.01004	0.05006	0.10005
10	Standard	0.001017	0.01005	0.05018	0.10008
	Robust	0.001033	0.01010	0.05021	0.10005

hybridized at the Australian Genome Research Facility to Illumina Mouse Whole-Genome-6 version 2 BeadChips, a microarray platform containing about 48,000 60-mer DNA sequences probing most genes in the genome. Intensities were background corrected, quantile normalized and transformed to the $\log_2$-scale using the neqc function [Shi, Oshlack and Smyth (2010)]. One of the *Suz12* mutant samples was discarded because it clustered with the wild type instead of the *Suz12* samples, leaving four wild type and three *Suz12* mutant samples. Probes were filtered from further analysis if they failed to achieve a detection p-value of less than 0.01 in at least two of the remaining samples, leaving 14,084 probes for analysis.

Linear modeling was applied to the normalized log-expression values, resulting in a residual sample variance on 5 degrees of freedom for each probe. Figure 1(a) shows the residual standard deviation plotted against the average log intensity for each probe. The gray curve shows the estimated trend for the prior variance. The robust algorithm identified a number of outlier variances [Figure 1(b)]. The non-robust estimate of the prior degrees of freedom was 11.9. The robust algorithm estimated prior degrees of freedom 14.1 for most genes, but with prior degrees of freedom as low as 0.5 for the outlier variances [Figure 4(a)].

Further examination showed that many of the probes identified as outliers corresponded to genes known to have sex-linked expression, including many on the X or Y chromosomes [Figure 1(a), (b)]. The most outlying variances corresponded to Y chromosome genes *Erdr1* and *Eif2s3y* up-regulated in males, and X chromosome gene *Xist*, known to be up-regulated in females. Other outlier genes were ribosomal genes *Rn18s* and *Rpl7a*, suggesting ribosomal RNA retention in one or more samples, and hemoglobin genes *Hbb-y* and *Hbb-b1*, suggesting red blood or bone marrow content in some tissue samples. None of these genes should be related to the *Suz12* mutation.

Differential expression between the *Suz12* mutants and the wild-type mice was assessed using EB moderated t-statistics. P-values were adjusted to control the

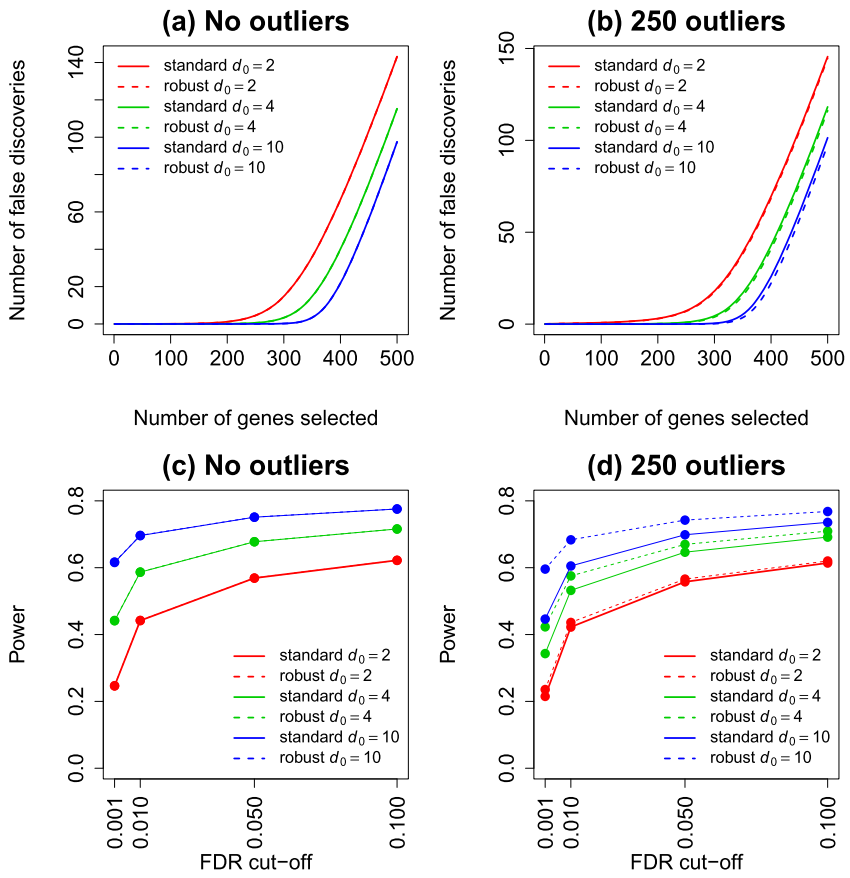

FIG. 3. *Detection of differential expression. Panels* (a)–(b) *show the number of false discoveries among the* 500 *top-ranked genes. Panels* (c)–(d) *show power, the proportion of truly DE genes selected as significant at various FDR cutoffs. Results are averaged over* 1000 *simulations. Simulations* (b) *and* (d) *include* 250 *hypervariable genes. Results are shown for both standard and robust EB tests and for three values of the prior degrees of freedom.*

false discovery rate at less than 5% [Benjamini and Hochberg (1995)]. The standard and robust procedures found 251 down-regulated and 35 up-regulated probes in common [Figure 4(b)]. However, 22 and 16 down-regulated genes were found only by the robust or standard procedures respectively. The nonrobust unique genes tended to be sex linked or hemoglobin related (*Xist*, *Apoa2*, *Hbb-b1*, etc.), whereas the robust unique genes were related to programmed cell death (*Bcl2l1*), cell cycle (*Ccne2*) or chromatin remodeling (*Myst2*). For up-regulated genes, 8 and 10 unique probes were found by the robust and standard procedures respectively. The nonrobust unique genes tended to be Y chromosome sex-linked genes (*Ddx2y*, *Erdr1* etc.), whereas the robust unique genes appeared related to the PRC2 process of interest.

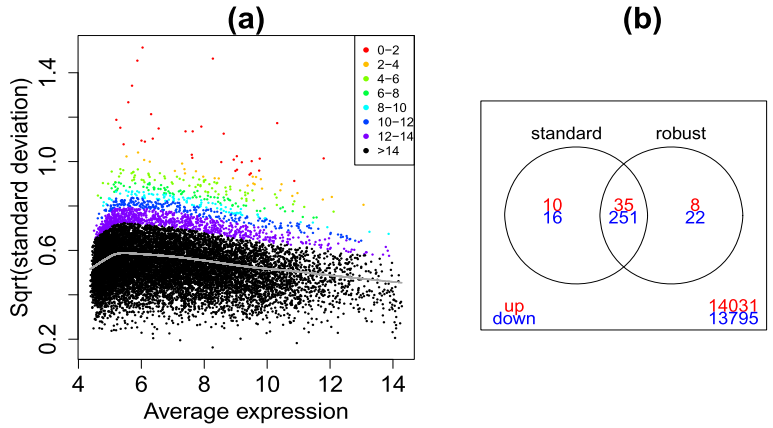

FIG. 4. (a) *Square-root standard deviation plotted against average log intensity for each probe. The gray line shows the trended estimate of the prior variance. Points are colored by the estimated prior degrees of freedom: probes with larger sample variances have smaller df.* (b) *Venn diagram showing overlap of the numbers of significant genes for Suz*12 *versus wild type using the standard and robust methods.*

Investigation after the analysis confirmed that two of the *Suz12* mutant embryos were in fact female, whereas all the other mice were male. This sex imbalance was an unwanted complication in the experiment, difficult to avoid without sex typing of the embryo mice at the time of tissue collection. The results show that the robust EB method was successful in identifying and downweighting genes that are associated with the hidden covariate. The robust procedure results in more statistical power to detect other genes that are more likely to be of scientific significance.

8.2. *RNA-seq profiles of Yoruba HapMap individuals.* As part of the International HapMap project, RNA-Seq profiles were made of cell lines derived from B lymphocytes from 69 different Yoruba individuals from Ibadan, Nigeria [Pickrell et al. (2010a, 2010b)]. Genewise read counts were obtained from the tweeDEseqCountData package version 1.8.0 from Bioconductor and were transformed to log 2-counts per million with precision weights using voom [Law et al. (2014)]. The analysis compares males to females. Figure 1(c) shows the genewise standard deviations and Figure 1(d) gives a probability plot of the standard deviations against the fitted F_{d,d_0} distribution. The hypervariable genes were identified as B cell receptor segments on chromosomes 2 and 22. B cells contain a random selection of these segments in order to generate a repertoire of antigen binding sites. The robust analysis reveals that the cell lines were clonal, each cell line apparently derived from a single B cell or from a very small number of original cells. It would be appropriate to remove the receptor segments from the RNA-seq analysis, and the robust EB procedure effectively achieves that.

8.3. *Embryonic stem cells*. Sheikh et al. (2015) used RNA-seq to profile embryonic stem cells. Genewise read counts were transformed using voom. Figure 1(e) shows the genewise standard deviations and Figure 1(f) gives a probability plot of the standard deviations against the fitted F_{d,d_0} distribution. The hypervariable genes are predominately associated with the ribosome or with hemoglobin, suggesting inconsistencies in cell purification and RNA processing. Other hypervariable genes are located in the major histocompatibility complex, known to be one of the most variable parts of the genome.

9. Discussion. In recent years we have routinely checked for hypervariable genes in gene expression studies. We have found that many studies harbor a subset of outlier genes. In many cases, the identities of the hypervariable genes suggest a mechanism for their variability. We have analyzed studies, for example, where the hypervariable genes are enriched for sex-linked genes, for ribosomal genes, for mitochondrial genes or for B cell receptor segments. In other cases, hypervariable genes may be associated with a particular cell type suggesting inconsistent cell population proportions in the different biological samples. In some cases, the reasons why some genes are highly variable between individuals are unknown. The phenomenon is common enough to be viewed as an unavoidable part of cutting edge genomic research rather than a result of flawed experimental procedures. In most cases the studies are overall of high quality.

Hypovariable genes can also arise, although usually for technical rather than biological reasons. Quantile normalization [Bolstad et al. (2003)] of expression data can occasionally produce expression values that are numerically identical for all samples for a given gene when the number of samples is small. Sequencing data can also give rise to variances that are zero or very small because of the discreteness of read counts.

This article describes a robustified version of EB differential expression analysis. This procedure protects against hyper and hypovariable genes in the sense that it allows nonoutlier genes to share information among themselves as if the outlier genes were not present. In many cases, this results in a gain in statistical power for the nonoutlier genes. Hypervariable genes are not removed from the analysis but instead borrow less information from the ensemble and are assigned test statistics closer to ordinary t-statistics. Hypovariable genes, on the other hand, are moderated as for nonoutlier genes—this increases their posterior variances closer to typical values.

A key feature of our procedure is that a conjugate Bayesian model is used for each gene, enabling closed-form posterior estimators and exact small-sample p-values. Robustness is achieved by fitting the prior distribution to nonoutlier genes and by assigning lower degrees of freedom to the hypervariable outliers. We have proposed a practical algorithm for assigning prior degrees of freedom to outlier genes. In practice, the list of DE genes is not sensitive to the exact values of the prior degrees of freedom, provided that $d_{0g} = d_0$ for nonoutliers and d_{0g} is

substantially smaller for clear outliers. For many datasets the list of DE genes is unchanged for a range of reasonable values for d_{outlier}.

The default values for the Winsorizing tail proportions p_l and p_u work well in our practical experience. Users, however, may choose to increase the default values for datasets where high proportions of outlier genes are expected.

Simulations show that the robustified EB procedure estimates the hyperparameters equally as accurately as the original method in the absence of outliers. When outliers are present, however, the robustified EB procedure was able to simultaneously increase power and decrease the false discovery rate when assessing differential expression.

The robust EB method developed here has been applied not only to microarray data, but also to data from RNA-seq [Good-Jacobson et al. (2014)], ChIP-seq [Lun and Smyth (2015b, 2016)] and Hi-C [Lun and Smyth (2015a)] technologies. With microarrays, the linear models are fitted to normalized log-intensities. With RNA-seq, the number of sequence reads overlapping each gene can be counted and the EB models can be applied to the log-counts-per million [Law et al. (2014)]. Alternatively, the robust EB method can be applied to count data by way of quasi-generalized linear models. Lun, Chen and Smyth (2016) analyzed RNA-seq read counts used quasi-negative-binomial generalized linear models. The limma robust EB variance estimation method was applied to the genewise residual deviances, leading to the construction of EB quasi-F-tests. The robust EB method has also been used to estimate the prior degrees of freedom for the weighted likelihood approach used in the edgeR package, again using residual deviances [Chen, Lun and Smyth (2014)].

In this article we view genes as outliers rather than individual expression values as outliers. An alternative robustifying approach would be to replace least squares estimation of the genewise linear models with robust regression [Gottardo et al. (2006), Zhou, Lindsay and Robinson (2014)], and the limma package has included an M-estimation option for this purpose for over a decade. In principle, the two approaches are complementary and both can be used simultaneously, that is, one could apply the robust EB procedure of this article to variances estimated by robust regression. The robust regression approach assumes that the expression values for a gene contain one or two outliers that are "errors," whereas the other values are "correct." Our experience suggests that this scenario is relatively rare for gene expression data. The outlier-gene approach of this article allows a more general context in which a hypervariable gene may produce an arbitrary number of inconsistent expression values that cannot be meaningfully categorized into correct and incorrect. The robust regression is only applicable to experimental designs with at least three expression values per experimental group, whereas the outlier-gene approach can be usefully applied to any experimental design, even down to studies with a single residual degree of freedom.

The robust EB strategy of this article could, in principle, be applied in other EB contexts. The basic idea is to estimate hyperparameters robustly, then to test

for outlier cases, and finally to assign a more diffuse prior to outlier cases. This approach may be particularly attractive for use with conjugate Bayesian models and seems different from previous robust EB strategies.

It is interesting to contrast our approach with the large literature on robust Bayesian analysis [Berger (1984, 1990), Insua and Ruggeri (2000)]. Robust Bayesian analysis considers a class of possible prior distributions and tries to limit or at least quantify the range of posterior conclusions as the prior ranges over the class. The issues that concern us in this article are different in a number of important ways. The issues that we address are specific to empirical Bayes and do not arise in true Bayesian frameworks for which hyperparameters do not need to be estimated. Our aim is not to limit the influence of the prior but to increase it for the majority of genes. Our method applies only to large-scale data with many cases (probes, genes or genomic regions), whereas robust Bayesian analysis is typically concerned with individual cases.

SUPPLEMENTARY MATERIAL

Supplementary Methods: Details of computational algorithms (DOI: 10.1214/16-AOAS920SUPP; .pdf). We provide further details of the various numerical algorithms described in this article.

REFERENCES

BALDI, P. and LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17** 509–519.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, London. MR0326887

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300. MR1325392

BERGER, J. O. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses* (J. Kadane, ed.). *Stud. Bayesian Econometrics* **4** 63–144. North-Holland, Amsterdam. With comments and with a reply by the author. MR0785367

BERGER, J. O. (1990). Robust Bayesian analysis: Sensitivity to the prior. *J. Statist. Plann. Inference* **25** 303–328. MR1064429

BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M. and SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.

BRENT, R. P. (1973). *Algorithms for Minimization Without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

CASELLA, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39** 83–87. MR0789118

CHEN, Y., LUN, A. T. L. and SMYTH, G. K. (2014). Differential expression analysis of complex RNA-seq experiments using edgeR. In *Statistical Analysis of Next Generation Sequence Data* (S. Datta and D. S. Nettleton, eds.) 51–74. Springer, New York.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836. MR0556476

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139. MR0323015

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

GAVER, D. P. and O'MUIRCHEARTAIGH, I. G. (1987). Robust empirical Bayes analyses of event rates. *Technometrics* **29** 1–15. MR0876882

GOOD-JACOBSON, K. L., CHEN, Y., VOSS, A. K., SMYTH, G. K., THOMAS, T. and TARLINTON, D. (2014). Regulation of germinal center responses and B-cell memory by the chromatin modifier MOZ. *Proc. Natl. Acad. Sci. USA* **111** 9585–9590.

GOTTARDO, R., RAFTERY, A. E., YEUNG, K. Y. and BUMGARNER, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62** 10–18. MR2226551

INSUA, D. R. and RUGGERI, F., eds. (2000). *Robust Bayesian Analysis*. *Lecture Notes in Statistics* **152**. Springer, New York. MR1795206

JEANMOUGIN, M., DE REYNIES, A., MARISA, L., PACCARD, C., NUEL, G. and GUEDJ, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE* **5** e12336.

JI, H. and LIU, X. S. (2010). Analyzing 'omics data using hierarchical models. *Nature Biotechnology* **28** 337.

KOOPERBERG, C., ARAGAKI, A., STRAND, A. D. and OLSON, J. M. (2005). Significance testing for small microarray experiments. *Stat. Med.* **24** 2281–2298. MR2151706

LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15** R29.

LIAO, J. G., MCMURRY, T. and BERG, A. (2014). Prior robust empirical Bayes inference for large-scale data by conditioning on rank with application to microarray data. *Biostatistics* **15** 60–73.

LUN, A. T. L., CHEN, Y. and SMYTH, G. K. (2016). It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Methods in Molecular Biology* **1418** 391–416.

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

LUN, A. T. L. and SMYTH, G. K. (2015b). From reads to regions: A Bioconductor workflow to detect differential binding in ChIP-seq data. *F1000Research* **4** 1080.

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

MAJEWSKI, I. J., BLEWITT, M. E., DE GRAAF, C. A., MCMANUS, E. J., BAHLO, M., HILTON, A. A., HYLAND, C. D., SMYTH, G. K., CORBIN, J. E., METCALF, D. et al. (2008). Polycomb repressive complex 2 (PRC2) restricts hematopoietic stem cell activity. *PLOS Biology* **6** e93.

MAJEWSKI, I. J., RITCHIE, M. E., PHIPSON, B., CORBIN, J., PAKUSCH, M., EBERT, A., BUSSLINGER, M., KOSEKI, H., HU, Y., SMYTH, G. K. et al. (2010). Opposing roles of polycomb repressive complexes in hematopoietic stem and progenitor cells. *Blood* **116** 731–739.

MCCARTHY, D. J. and SMYTH, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25** 765–771.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. With discussion. MR0696849

MURIE, C., WOODY, O., LEE, A. Y. and NADON, R. (2009). Comparison of small *n* statistical tests of differential expression applied to microarrays. *BMC Bioinformatics* **10** 45.

PHIPSON, B., LEE, S., MAJEWSKI, I. J., ALEXANDER, W. S. and SMYTH, G. K. (2016). Supplement to "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression." DOI:10.1214/16-AOAS920SUPP.

PICKRELL, J. K., MARIONI, J. C., PAI, A. A., DEGNER, J. F., ENGELHARDT, B. E., NKADORI, E., VEYRIERAS, J.-B., STEPHENS, M., GILAD, Y. and PRITCHARD, J. K. (2010a). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464** 768–772.

PICKRELL, J. K., PAI, A. A., GILAD, Y. and PRITCHARD, J. K. (2010b). Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet*. **6** e1001236.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. and SMYTH, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. **43** e47.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

SARTOR, M. A., TOMLINSON, C. R., WESSELKAMPER, S. C., SIVAGANESAN, S., LEIKAUF, G. D. and MEDVEDOVIC, M. (2006). Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics* **7** 538.

SHEIKH, B. N., DOWNER, N. L., PHIPSON, B., VANYAI, H. K., KUEH, A. J., MCCARTHY, D. J., SMYTH, G. K., THOMAS, T. and VOSS, A. K. (2015). MOZ and BMI1 play opposing roles during Hox gene activation in ES cells and in body segment identity specification in vivo. *Proc. Natl. Acad. Sci. USA* **112** 5437–5442.

SHI, W., OSHLACK, A. and SMYTH, G. K. (2010). Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res*. **38** e204.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*. **3** Article 3. MR2101454

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Stat*. **33** 1–67. MR0133937

WRIGHT, G. W. and SIMON, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19** 2448–2455.

ZHOU, X., LINDSAY, H. and ROBINSON, M. D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res*. **42** e91.

B. PHIPSON
MURDOCH CHILDRENS RESEARCH INSTITUTE
50 FLEMINGTON ROAD
PARKVILLE, 3052
VICTORIA
AUSTRALIA
E-MAIL: belinda.phipson@mcri.edu.au

S. LEE
I. J. MAJEWSKI
W. S. ALEXANDER
G. K. SMYTH
THE WALTER AND ELIZA HALL
INSTITUTE OF MEDICAL RESEARCH
1G ROYAL PARADE
PARKVILLE, 3052
VICTORIA
AUSTRALIA
E-MAIL: smyth@wehi.edu.au