Empirical likelihood tests for nonparametric detection of differential expression from RNA-seq data

Article

Accepted Version

Empirical likelihood tests for nonparametric detection of differential expression from RNA-seq data

Thomas Thorne*, thomas.thorne@ed.ac.uk,
School of Informatics, University of Edinburgh, EH8 9AB.

January 18, 2016

Abstract

The availability of large quantities of transcriptomic data in the form of RNA-seq count data has necessitated the development of methods to identify genes differentially expressed between experimental conditions. Many existing approaches apply a parametric model of gene expression and so place strong assumptions on the distribution of the data. Here we explore an alternate nonparametric approach that applies an empirical likelihood framework, allowing us to define likelihoods without specifying a parametric model of the data. We demonstrate the performance of our method when applied to gold standard datasets, and to existing experimental data. Our approach outperforms or closely matches performance of existing methods in the literature, and requires modest computational resources. An R package, `EmpDiff` implementing the methods described in the paper is available from `http://homepages.inf.ed.ac.uk/tthorne/software/packages/EmpDiff_0.99.tar.gz`.

Keywords: Differential expression, RNA-Seq, Transcriptomics.

1 Introduction

Tests for differential expression allow us to generate hypotheses about the regulatory mechanisms behind the differing phenotypes observed between experimental conditions, and can be applied as a first step towards determining the genes involved and prioritising targets for further investigation. Tools that allow for the fast and robust determination of genes that are differentially expressed between sets of samples are of great value to experimentalists in identifying the knock-on effects of perturbations, or in identifying candidate genes responsible for changes observed between cases and controls. The comparatively simple nature of the task allows us to deliver tools that are widely applicable and require little to no expert tuning of the underlying algorithm.

When working with data collected from RNA-seq experiments, after alignment of reads to the genome, the data consist of count values of the number of mapped reads to each gene. Since samples may not all be sequenced to the same

*to whom correspondence should be addressed

depth, some form of normalisation is generally applied to compensate for this – here we apply the approach of Anders and Huber (2010). Some approaches also attempt to correct for the effects of gene length on numbers of mapped reads, but since in our methods, outlined below, we only seek to compare expression between samples on a gene by gene basis, this is not necessary.

Existing methods for the detection of differential expression from RNA-seq data typically apply either parametric Bayesian models, or frequentist nonparametric approaches. It is important to note that unlike microarray expression data, where the data are typically assumed to be normally distributed, this is not the case with RNA-seq count data, so methods for detection of differential expression in microarray data cannot be applied unmodified. The most straightforward parametric model of a Poisson distribution, having only the mean as a parameter, typically underestimates the variance of the data. When choosing parametric models for RNA-seq data, existing methods have typically considered negative binomial distributions (Leng et al., 2013; Love et al., 2014; Robinson et al., 2010; Hardcastle and Kelly, 2010), with the aim of addressing the overdispersion seen when applying Poisson distributed models. However this complicates the inference procedure as additional parameters need to be inferred. Many approaches incorporate empirical Bayes estimates (Leng et al., 2013; Love et al., 2014; Robinson et al., 2010; Hardcastle and Kelly, 2010) to attempt to learn hyperparameters from the observed data, making use of the large numbers of genes typically considered to share information on the distribution of counts between genes.

Nonparametric schemes also exist in the literature, for example the approach of Li and Tibshirani (2013) (SAMSeq) that applies a Wilcoxon test with a resampling scheme to compensate for sequencing depth, and Tarazona et al. (2011) (NOISeq) that compares fold changes to empirically estimated noise distributions. Here we explore an application of the empirical likelihood methods of Owen (1988), that provide a nonparametric framework in which we can estimate the likelihood of observed data. This frees us from making assumptions about the distribution of the data required to derive a traditional parametric model from which the likelihood can be calculated, instead deriving likelihoods empirically from the data. These approaches are also computationally efficient, making them well suited to the large data sets generated by RNA-seq experiments. The generality of the empirical likelihood approach makes it an ideal candidate for consideration in the detection of differential expression from RNA-seq data, where work on parametric models has shown the benefit of utilising distributions that better fit the observed data.

2 Methods

We work with RNA-seq count data presented as a set of values for a gene i of $x^1_{i1}, \ldots, x^1_{im}$ and $x^2_{i1}, \ldots, x^2_{in}$ for two conditions with m and n samples to be compared. For each gene we aim to test the hypothesis that the two conditions have differing mean (normalised) counts, which we will denote as differential expression (DE) versus the null hypothesis of a shared mean between the two conditions, corresponding to non-differentially expressed genes (NDE). Thus we will consider methods to estimate the likelihood of a particular value of the mean given observed counts for a condition. In the following we assume a single

gene is under consideration and drop the subscript i for convenience.

2.1 Empirical likelihood

The empirical likelihood framework assigns a likelihood to a set of independent and identically distributed observations by performing constrained optimisation on a set of probability weights assigned to each observation. Given $x_1, \ldots, x_n$ we maximise

$$f(x) = \prod_{i=1}^{n} p_i, \tag{1}$$

under the constraints $\sum_{i=1}^{n} p_i = 1$, $\forall i, 0 \leq p_i \leq 1$. Finally the model is imposed by adding a constraint

$$\sum_{i=1}^{n} t(x_i, \theta) p_i = 0. \tag{2}$$

When considering the mean μ of X we set equation 2 to $\sum_{i=1}^{n} x_i p_i - \mu = 0$. Then it is possible to use Lagrange multipliers to derive the optimal p_i (see supplementary material for further details) and it is straightforward to apply numerical root finding algorithms to determine the optimal values of the p_i for a given μ and so determine the value of the empirical likelihood. This likelihood can then be applied as an approximation in a scenario where a traditional likelihood derived from a parametric model would be used.

2.2 Euclidean likelihood

Using the methods described above, it is not possible for μ to escape the convex hull of the data, so that the support of the likelihood is constrained by the minimum and maximum of the observations. In the setting of testing for differential expression of genes between two conditions, we may often encounter the case where the convex hulls of observations (and so the support of the likelihoods) under the two conditions do not intersect one another. We could simply take this as evidence of differential expression, but there is no straightforward way to assess the statistical significance of such cases. The Euclidean likelihood, introduced by Baggerly (1998), circumvents this by instead of utilising the standard likelihood of equation 1, defining (Owen, 2001)

$$\log f(x|\mu) = -\frac{1}{2} \sum_{i=1}^{n} (np_i - 1)^2, \tag{3}$$

a measure of the divergence of the p_i from the optimum (for the mean) of $\frac{1}{n}$. We also no longer constrain the p_i to be positive, leaving us with the constraints $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} x_i p_i - \mu = 0$. Seeking to maximise $\log f(x|\mu)$ we arrive at an expression for the log likelihood as

$$\log f(x|\mu) = -\frac{n}{2s} (\bar{x} - \mu)^2, \tag{4}$$

where s is the sample variance, see supplementary material for a derivation. It should be noted that this is equivalent to utilising the asymptotic normal

distribution of the mean (under no assumptions about the distribution of the data) to define an approximate likelihood (Pawitan, 2001), and so we might expect to be able to improve on this approach.

2.3 Penalised empirical likelihood

The method outlined above does not enjoy the same properties as the empirical likelihood, and will not generally yield the same results. A solution to this problem is to apply the penalised empirical likelihood of Bartolucci (2007), whereby a penalty term that approximates the likelihood outside of the convex hull of the data is introduced, but the likelihood converges to the empirical likelihood as a tolerance term h tends to zero. Using the form of the penalty term suggested in Bartolucci (2007) we define the penalised empirical likelihood as

$$f(x|\mu) = \max_{\nu} g(x|\nu) \exp -\frac{n}{2h^2 S}(\mu - \nu)^2 \tag{5}$$

where $g(x|\nu)$ is the traditional form of the empirical likelihood for the mean, and ν is constrained so that $\min(x) \leq \nu \leq \max(x)$. As h tends to zero the penalty on $\nu \neq \mu$ increases and we recover the empirical likelihood. We set h to the optimal value of $h = \sqrt{(1 - \frac{\epsilon}{n})^{-1} - 1}$ as described in Bartolucci (2007).

2.4 Likelihood ratio test

To test for differential expression of each gene we apply a simple likelihood ratio test comparing the hypothesis of a shared mean $\mu_{12} = \bar{x}^{1,2}$ and independent means $\mu_1 = \bar{x}^1$ and $\mu_2 = \bar{x}^2$. This gives us a likelihood ratio r of

$$r = 2 \log \left(\frac{f(x_1^1, \ldots, x_m^1|\mu_1) f(x_1^2, \ldots, x_n^2|\mu_2)}{f(x_1^1, \ldots, x_m^1|\mu_{12}) f(x_1^2, \ldots, x_n^2|\mu_{12})} \right). \tag{6}$$

Once likelihood ratios are calculated for each gene, we calculate p-values of differential expression for each gene, using the $F_{1,n}$ distribution as suggested in Owen (2001), and then adjust for multiple testing using the Bonferroni correction. This is the most conservative approach – other methods, for examples those that control the false discovery rate (FDR) are of course equally applicable.

2.5 Data preprocessing and software

Data were preprocessed by filtering out genes with no expression in either condition, and by removing genes with less than 10 counts in both conditions. Subsequently counts were normalised between samples using the method described in Anders and Huber (2010).

Our approach is implemented as a software package `EmpDiff` for the R statistical environment (R Core Team, 2015) and is available to download from the author's website (`http://homepages.inf.ed.ac.uk/tthorne/software/packages/EmpDiff_0.99.tar.gz`). Below we refer to the Euclidean likelihood method as EmpDiff Euclidean and the penalised empirical likelihood as EmpDiff penalised.

3 Results

3.1 MAQC Data

The MicroArray Quality Control (MAQC) project 2 (Canales et al., 2006; MAQC Consortium et al., 2006) provides us with an opportunity to benchmark our differential expression test against a gold standard list of differentially expressed (DE) and non-differentially expressed (NDE) genes. Count data was taken from the ReCount database (Frazee et al., 2011) and qRT-PCR data obtained from the Gene Expression Omnibus datbase (Barrett et al., 2013; Edgar et al., 2002) (accession number *GSE5350*, `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5350`).

Following the procedures described in Bullard et al. (2010), differentially expressed genes were defined as genes in the qRT-PCR samples with mean absolute $\log_2$ fold change greater than 2, and non-differentially expressed genes as having mean $\log_2$ fold change less than 0.2. This gave a set of 297 DE and 97 NDE genes. We restrict our analyses to genes identified as either DE or NDE from the qRT-PCR data (although benchmarked methods are run on the full set of genes in the RNA-seq data so that empirical Bayes approaches are able to take advantage of all of the data). Count data and lists of DE and NDE genes are available using the dataset `maqc` in the `EmpDiff` R package.

We applied several existing methods from the literature to the data to compare performance with our approach, considering baySeq (Hardcastle and Kelly, 2010), DESeq2 (Love et al., 2014), EBSeq (Leng et al., 2013), edgeR (Robinson et al., 2010), NOISeq (Tarazona et al., 2011), and SAMSeq (Li and Tibshirani, 2013).

The R package `PRROC` (Grau et al., 2015) was used to plot Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves for the methods, see figure 1, and to calculate the Area Under the Curve (AUC) for each curve, tabulated in table 1. To generate the ROC and PR curves, the terms are defined as follows:

$$\text{Sensitivity} = \frac{\text{No. of correctly predicted DE genes}}{\text{No. of DE genes}} \tag{7}$$

$$\text{FPR} = \frac{\text{No. of incorrectly predicted DE genes}}{\text{No. of NDE genes}} \tag{8}$$

$$\text{Precision} = \frac{\text{No. of correctly predicted DE genes}}{\text{No. of genes predicted DE}} \tag{9}$$

$$\text{Recall} = \text{Sensitivity} \tag{10}$$

As can be seen from figures 1 both of the empirical likelihood approaches perform the best out of those considered in terms of the AUC, with the Euclidean likelihood scheme being indistinguishable from the penalised empirical likelihood approach.

3.2 SimSeq simulated data

We also utilised the SimSeq algorithm (Benidt and Nettleton, 2015) to simulate RNA-seq data with known DE and NDE genes. The `SimSeq` R package was used

to simulate RNA-seq counts for 10 sets of 5, 10 and 15 samples each from two conditions, with 1000 DE genes and 4000 NDE genes, using the `kidney` dataset provided, as taken from the KIRC RNA-seq dataset (The Cancer Genome Atlas Research Network, 2013) in The Cancer Genome Atlas (TCGA) database.

We again compare the methods considered in section 3.1, with the exception of SAMSeq due to the difficulty in batch processing with the currently available software, using the known lists of DE and NDE genes to plot ROC and PR AUC with the `PPROC` R package. The plots of AUC in figure 2 show that our approach performs well in terms of the AUROC for small numbers of samples, and comparably to baySeq for larger numbers of samples. DESeq2 appears to have superior performance for larger numbers (10 and 15) of samples, and performs well in the PR AUC. Considering PR, our approach is comparable to baySeq and EBSeq for smaller numbers of samples. There is little to separate the Euclidean likelihood from the penalised empirical likelihood.

One drawback of the SimSeq algorithm is that it is somewhat circular in the derivation of simulated differential expression – it relies on a test for differential expression to pick genes to sample and utilise as differentially expressed in the simulated data. However such a nonparametric approach is perhaps still preferable to synthetic data generated from parametric models that make strong prior assumptions on the distribution of the data, and so automatically favour schemes based on the same parametric model.

3.3 Aryl hydrocarbon receptor regulation in MCF-7 cancer cells

To explore the utility of our approach on an experimental data set, we consider RNA-seq experiments probing the effects of the aryl hydrocarbon receptor (AHR) transcription factor on gene expression in MCF-7 cancer cells (Salisbury et al., 2014). The data were taken from the GEO database (Barrett et al., 2013; Edgar et al., 2002), accession number *GSE52036* (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52036`). The data consist of RNA-seq counts for 57605 genes, with 6 control samples and 5 AHR knockdown samples, the aim being to identify genes in the regulatory network neighbourhood of AHR whose expression is altered by the knockdown of AHR.

Applying the Euclidean likelihood test for differential expression, out of the 57605 genes in the dataset, we find that 445 are differentially expressed at the 1% significance level. The 10 genes with highest likelihood ratio (after filtering of pseudogenes) are shown in table 2, along with their likelihood ratio and corresponding fold change. Of the genes listed, CYP1B1 (Yang et al., 2008; Salisbury et al., 2014) and HMOX1 (Lo and Matthews, 2013) are known to be regulated by AHR.

To test for enrichment of pathways within our gene set we applied the *g:Profiler* tools (Reimand et al., 2011, 2007) to the set of DE genes identified by our method. Results showing KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2002; Ogata et al., 1999) pathways that are significantly enriched in our set of DE genes are tabulated in table 3. In agreement with Salisbury et al. (2014), we find that a known 2,3,7,8 tetrachlorodibenzo-*p*-dioxin (TCDD) pathway is enriched in the set of DE genes, namely metabolism of xenobiotics by cytochrome P450 (Lo and Matthews, 2012; Dere et al., 2011).

For comparison we also apply the next best performing methods (table 1), baySeq and DESeq2, and the nonparametric SAMSeq method, to compare the coverage of differentially expressed genes identified. Applying firstly baySeq, 177 genes were found to be differentially expressed when taking a posterior probability of DE greater than 0.99 as the cutoff, with 137 of these found to be in common with those identified by EmpDiff. Then applying DESeq2 157 genes were found to be differentially expressed at the 1% significance level, with 54 of these found to be in common with those identified by EmpDiff. Finally utilising SAMSeq we find 308 DE genes with a local false discovery rate of less than 0.01, with 127 of these in common with EmpDiff. In figure 3 we show plots of fold change against mean absolute difference between the control and knockdown samples, highlighting genes identified as DE by the four methods considered. It is apparent that baySeq is more conservative in identifying genes with a large absolute difference in expression level but a low fold change. SAMSeq is less conservative but has a similar distribution of significant genes to DESeq2, and all three methods competing methods fail to identify genes with a high fold change as being differentially expressed, whereas EmpDiff labels large fold changes as significant. Depending on whether it is desirable to include genes with low expression but high fold change in the results, our approach can easily be tuned by simply filtering genes with low mean expression from the study.

4 Conclusions

We have introduced a novel approach to testing for differential expression in RNA-seq data that, in contrast to the many existing parametric methods, makes no assumptions about the distribution of the data. Applying our methodology to experimental and simulated data we see performance superior or comparable to existing approaches, and we further demonstrate the applicability of the approach through analysis of DE genes identified by our methods when applied to RNA-seq data from MCF-7 cancer cells. Our method performs particularly well on data with small numbers of samples (SimSeq), or with several technical replicates (MAQC data), and we show that it identifies genes with large fold changes as significant that are not found by other approaches. We find that the Euclidean likelihood performs as well as the penalised empirical likelihood, despite only being an approximation. The method is available as a software package, `EmpDiff` implemented in R from the author's website.

Acknowledgement

This work was supported by the University of Edinburgh Chancellor's Fellowship to T.T.

References

Anders, S. and W. Huber (2010): "Differential expression analysis for sequence count data," *Genome biology*, 11, R106.

Baggerly, K. A. (1998): "Empirical likelihood as a goodness-of-fit measure," *Biometrika*, 85, 535–547.

Barrett, T., S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva (2013): "NCBI GEO: archive for functional genomics data sets–update." *Nucleic Acids Research*, 41, D991–5.

Bartolucci, F. (2007): "A penalized version of the empirical likelihood ratio for the population mean," *Statistics & probability letters*, 77, 104–110.

Benidt, S. and D. Nettleton (2015): "SimSeq: a nonparametric approach to simulation of RNA-sequence datasets." *Bioinformatics*, 31, 2131–2140.

Bullard, J. H., E. Purdom, K. D. Hansen, and S. Dudoit (2010): "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, 11, 94.

Canales, R. D., Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid (2006): "Evaluation of DNA microarray results with quantitative gene expression platforms." *Nature biotechnology*, 24, 1115–1122.

Dere, E., R. Lo, T. Celius, J. Matthews, and T. R. Zacharewski (2011): "Integration of Genome-Wide Computation DRE Search, AhR ChIP-chip and Gene Expression Analyses of TCDD-Elicited Responses in the Mouse Liver," *BMC genomics*, 12, 365.

Edgar, R., M. Domrachev, and A. E. Lash (2002): "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." *Nucleic Acids Research*, 30, 207–210.

Frazee, A. C., B. Langmead, and J. T. Leek (2011): "ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinformatics*, 12, 449.

Grau, J., I. Grosse, and J. Keilwagen (2015): "PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R." *Bioinformatics*, 31, 2595–2597.

Hardcastle, T. J. and K. A. Kelly (2010): "baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, 11, 422.

Kanehisa, M. and S. Goto (2000): "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic Acids Research*, 28, 27–30.

Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya (2002): "The KEGG databases at GenomeNet." *Nucleic Acids Research*, 30, 42–46.

Leng, N., J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski (2013): "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments." *Bioinformatics*, 29, 1035–1043.

Li, J. and R. Tibshirani (2013): "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data." *Statistical Methods in Medical Research*, 22, 519–536.

Lo, R. and J. Matthews (2012): "High-resolution genome-wide mapping of AHR and ARNT binding sites by ChIP-Seq." *Toxicological Sciences*, 130, 349–361.

Lo, R. and J. Matthews (2013): "The aryl hydrocarbon receptor and estrogen receptor alpha differentially modulate nuclear factor erythroid-2-related factor 2 transactivation in MCF-7 breast cancer cells," *Toxicology and Applied Pharmacology*, 270, 139–148.

Love, M. I., W. Huber, and S. Anders (2014): "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology*, 15, 550.

MAQC Consortium, L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-h. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker (2006): "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." *Nature biotechnology*, 24, 1151–1161.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999): "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research*, 27, 29–34.

Owen, A. B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237–249.

Owen, A. B. (2001): *Empirical Likelihood*, CRC Press, Boca Raton, FL.

Pawitan, Y. (2001): *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.

R Core Team (2015): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Reimand, J., T. Arak, and J. Vilo (2011): "g:Profiler—a web server for functional interpretation of gene lists (2011 update)," *Nucleic Acids Research*, 39, W307–W315.

Reimand, J., M. Kull, H. Peterson, J. Hansen, and J. Vilo (2007): "g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments." *Nucleic Acids Research*, 35, W193–200.

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.

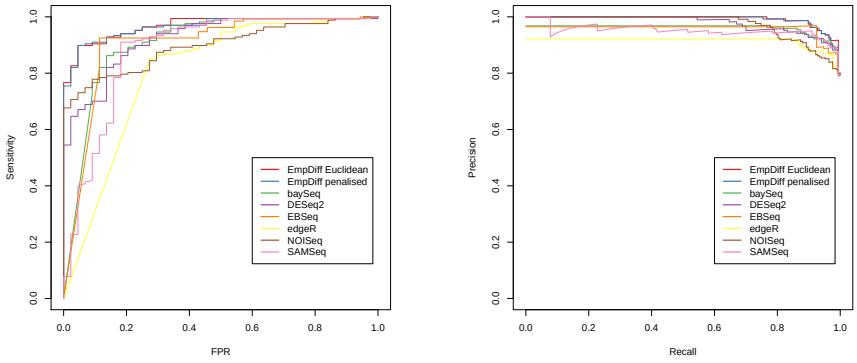
Salisbury, T. B., J. K. Tomblin, D. A. Primerano, G. Boskovic, J. Fan, I. Mehmi, J. Fletcher, N. Santanam, E. Hurn, G. Z. Morris, and J. Denvir (2014): "Endogenous aryl hydrocarbon receptor promotes basal and inducible expression of tumor necrosis factor target genes in MCF-7 cancer cells," *Biochemical Pharmacology*, 91, 390–399.

Tarazona, S., F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa (2011): "Differential expression in RNA-seq: a matter of depth." *Genome Research*, 21, 2213–2223.

The Cancer Genome Atlas Research Network (2013): "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, 499, 43–49.

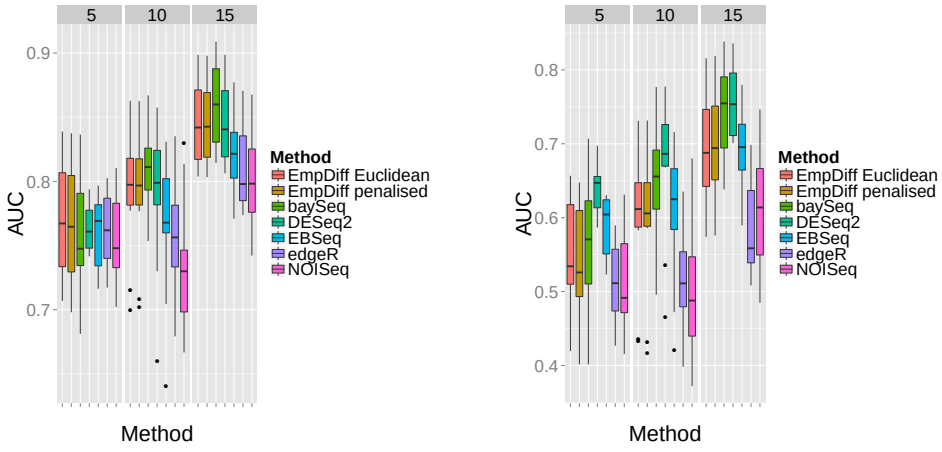
Yang, X., S. Solomon, L. R. Fraser, A. F. Trombino, D. Liu, G. E. Sonenshein, E. V. Hestermann, and D. H. Sherr (2008): "Constitutive regulation ofCYP1B1 by the aryl hydrocarbon receptor (AhR) in pre-malignant and malignant mammary tissue," *Journal of Cellular Biochemistry*, 104, 402–417.

List of Figures



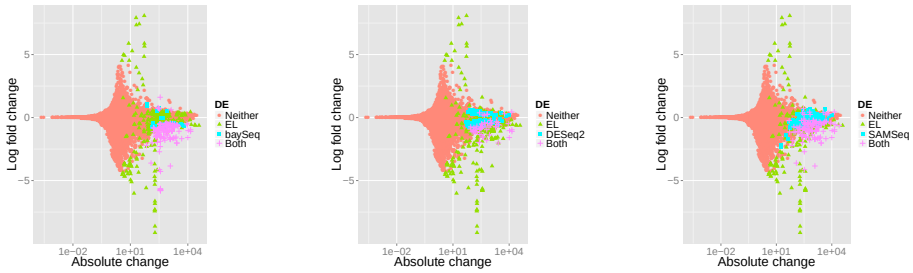
(a) (b)

Figure 1: (a) ROC and (b) PR curves for methods tested on the MAQC dataset. The two EmpDiff methods correspond to applying Euclidean likelihood or the penalised empirical likelihood.



(a) (b)

Figure 2: Area under curve for (a) ROC and (b) PR curves for methods on three sets of 10 different SimSeq simulated datasets, with 5, 10 and 15 replicates per condition. Again the two EmpDiff methods correspond to Euclidean and penalised empirical likelihood.



(a) (b) (c)

Figure 3: Plots of fold change against absolute difference between the control and AHR knockdown samples. Genes identified as DE by EmpDiff and baySeq or both (a), EmpDiff and DESeq2 or both (b), and EmpDiff and SAMSeq or both (c) are labelled.

List of Tables

	AUC	
	ROC	PR
EmpDiff Euclidean	0.97	0.99
EmpDiff penalised	0.97	0.99
baySeq	0.91	0.96
DESeq2	0.92	0.98
EBSeq	0.90	0.96
edgeR	0.80	0.91
NOISeq	0.90	0.97
SAMSeq	0.88	0.95

Table 1: AUC for ROC and PR curves when applying the different methods considered to the MAQC experimental data.

Ensembl ID	Name	Likelihood ratio	Log 2 fold change
ENSG00000027869	SH2D2A	2.624E+04	-7.426E+00
ENSG00000151632	AKR1C2	2.497E+04	-1.584E+00
ENSG00000027644	INSRR	2.388E+04	-7.320E+00
ENSG00000108602	ALDH3A1	2.302E+04	-3.845E+00
ENSG00000100292	HMOX1	1.834E+04	-2.025E+00
ENSG00000181577	C6orf223	1.809E+04	8.083E+00
ENSG00000231274	SBK3	1.443E+04	7.921E+00
ENSG00000138061	CYP1B1	1.148E+04	-1.037E+00
ENSG00000198400	NTRK1	8.444E+03	-5.310E+00
ENSG00000180061	TMEM150B	6.513E+03	7.349E+00

Table 2: Top 10 differentially expressed genes ranked by likelihood ratio, after filtering of pseudogenes. It can be seen that the ranking of genes does not simply correspond to the ordering of the fold changes.

ID	Pathway name	Adjusted p-value	No. of DE genes
KEGG:01100	Metabolic pathways	2.080E-03	44
KEGG:00140	Steroid hormone biosynthesis	3.450E-10	14
KEGG:00480	Glutathione metabolism	3.590E-05	9
KEGG:00500	Starch and sucrose metabolism	1.030E-05	10
KEGG:00980	Metabolism of xenobiotics by cytochrome P450	2.650E-12	17
KEGG:00053	Ascorbate and aldarate metabolism	5.060E-09	10
KEGG:00830	Retinol metabolism	4.060E-06	11
KEGG:05204	Chemical carcinogenesis	1.950E-10	16
KEGG:00983	Drug metabolism - other enzymes	2.050E-05	9
KEGG:00040	Pentose and glucuronate interconversions	5.260E-09	11
KEGG:00982	Drug metabolism - cytochrome P450	3.860E-08	13
KEGG:00860	Porphyrin and chlorophyll metabolism	6.840E-11	13

Table 3: Enrichment of KEGG and Reactome pathways within the set of DE genes identified by the Euclidean likelihood approach, as identified by the g:Profiler tool (Reimand et al., 2011, 2007). Adjusted p-values and numbers of DE genes within the pathway are also listed.