Generalised empirical Bayesian methods for discovery of differential data in high-throughput biology

Thomas J. Hardcastle *†

November 27, 2014

Abstract

High-throughput data are becoming ubiquitous in biological research, and rapidly changing technologies and application mean that novel methods for detecting differential behaviour that account for a 'small n, large P' setting are required at an increasing rate. The development of such methods is, in general, being done on an *ad hoc* basis, requiring further development cycles and a lack of standardization between analyses.

We present here a generalized method for identifying differential behaviour within high-throughput biological data through empirical Bayesian methods. This approach is based on our `baySeq` algorithm for identification of differential expression in RNA-seq data based on a negative binomial distribution, and in paired data based on a beta-binomial distribution. Here we show that the same empirical Bayesian approach can be applied to any parametric distribution, removing the need for lengthy development of novel methods for differently distributed data. We compare the application of these generic methods to methods developed specifically for particular distributions, and show equivalent or better performance. We additionally present a number of enhancements to the baySeq algorithm and a set of strategies to reduce the computational time required for complex data sets.

The methods are implemented in the **R** `baySeq` (v2) package, available at `http://www.bioconductor.org/packages/release/bioc/html/baySeq.html`.

*to whom correspondence should be addressed <tjh48@cam.ac.uk>

†Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, United Kingdom

1 Introduction

High-throughput data are becoming ubiquitous in biological research, and numerous statistical techniques have been developed to analyse these data, generally to identify patterns of difference between sets of biological replicates. Microarray technology led to a proliferation of methods [20, 32, 7, 24] designed to analyse data within a 'small n, large P' setting under assumptions of (log-) normality. The subsequent emergence of high-throughput sequencing (HTS) motivated the development of analysis methods which generally assumed some form of over-dispersed Poisson distribution [30, 2, 13].

The majority of these analytic methods seek not merely to adjust for the high-dimensionality of the data [3], but to exploit it through various forms of the 'borrowing' of information across the P dimension. However, many of the methods developed achieve this borrowing of information by exploiting specific features of the data. Consequently, while the methods developed for analysis of the negative-binomially distributed HTS data are conceptually similar to those previously developed for analysis of (log-) normally distributed data, the implementation of these methods is strongly divergent.

Novel technologies for high-throughput generation of biological data may require different distributional assumptions to current methods. More complex experimental designs, involving multi-dimensional data are describing diverse types of biological information within a single organism [36, 35, 37]. Complete analysis of single-cell sequencing also seems likely to require novel distributional assumptions [6, 17, 21], as do analyses of high-throughput quantitative proteomic data [8, 26]. While some of these challenges are beginning to be addressed, this is being done on an *ad hoc* basis, requiring further development cycles and a lack of standardisation between analyses.

We present here a generalised method for identifying differential behaviour within high-throughput data of all types. This approach is based on our `baySeq` algorithm for identification of differential expression in RNA-seq data based on a negative binomial distribution [13], and in paired data based on a beta-binomial distribution [14]. Here we show that the same empirical Bayesian approach can be applied to any parametric distribution, removing the need for lengthy development of novel methods for differently distributed data. We compare the application of these generic methods to methods developed specifically for particular distributions, and show equivalent or better performance. We additionally present a number of enhancements to the baySeq algorithm and a set of strategies to reduce the computational time required for complex data sets.

2 Methods

The generalisation of methods allows differential behaviour to be identified in any class (or combination of classes) of genomic event which can be detected through some application of high-throughput technologies. We consider the first dimension of the data to define a specific genomic event, and define the data

attached to a particular genomic event c as D_c. Thus, for the simple case of mRNA-seq, the D_c describes the number of sequenced reads for a gene c in each biological sample. The second dimension of the data gives an indexing of the samples; thus, D_{cj} refers to data from the jth sample for the cth genomic event. Further dimensions of the data may be used to refer to individual components of a genomic event; e.g., timepoints, marker classification, *et cetera*.

In addition to the sequenced (or other stochastic) high-throughput data, we may also consider sets of *observables*. These are known, fixed observations that influence the generation of the data. Typical examples of these observables include library scaling factors (a measure of the depth of sequencing for each sample) and coding sequence length (in mRNA-seq experiments).

2.1 Generalised empirical Bayesian methods

As previously [13, 14], we suppose that there exists some model M whose posterior likelihood, given the observed data, is to be estimated. The model is defined by the equivalence classes $\{E_1, ..., E_m\}$ such that samples i and j are equivalent on genomic event c if and only if D_{ci} and D_{cj} are drawn from identically parametrised distributions. For notational simplicity, we define the set D_{cE_q} as the data associated with equivalence class E_q. We similarly define the *replicate sets* $\{F_1, ..., F_k\}$ such that samples i and j are in the same replicate set F_r if and only if they are biological replicates, and define the set D_{cF_r} as the data associated with replicate set F_r.

The posterior likelihood of a model M for a genomic event c is then acquired by computation of

$$\mathbb{P}(M \mid D_c) = \frac{\mathbb{P}(D_c|M)\mathbb{P}(M)}{\mathbb{P}(D_c)} \tag{1}$$

The major challenge in estimating the posterior likelihood of M is in estimating $\mathbb{P}(D_c|M)$, the likelihood of the data for a particular genomic event c given the model. If $\theta_q = \{\phi_{q1}, ...\phi_{qn}\}$ is a random variable defining the parameters of the distribution of the data in D_{cE_q} and we assume that the θ_q are independent, then

$$\mathbb{P}(D_c \mid M) = \prod_q \int \mathbb{P}(D_{cE_q} \mid \theta_q)\mathbb{P}(\theta_q \mid M)\mathrm{d}\theta_q \tag{2}$$

If Θ_q is a set of values sampled from the distribution of θ_q, then $\mathbb{P}(D_c \mid M)$ can be approximated [11] by

$$\mathbb{P}(D_c \mid M) = \prod_q \frac{1}{|\Theta_q|} \sum_{\eta_q \in \Theta_q} \mathbb{P}(D_{cE_q} \mid \eta_q) \tag{3}$$

Similarly, posterior distributions on θ_q given a model M and the observed

data D_{cE_q}, can be estimated by weighting each η_q^h in Θ_q by ω_q^h, where

$$\omega_q^h = \frac{\mathbb{P}(D_{cE_q} \mid \eta_q^h)}{\sum_{\eta_q \in \Theta_q} \mathbb{P}(D_{cE_q} \mid \eta_q)} \tag{4}$$

We have previously shown for a negative-binomial distribution [13] and a beta-binomial distribution [14] how the data can be sampled to acquire the sets Θ_q, and thus approximate $\mathbb{P}(D_c|M)$ through such a numerical integration. These methods generalise to any parametric distribution for which there is some method for estimating the parameters given the observed data D_c. The implementation of this generalisation in baySeq v2 allows these empirical Bayesian methods to be applied to any distribution for which a maximum likelihood solution exists, including for multi-dimensional data, simply by defining a density function f such that $\mathbb{P}(D_{cj}|\eta) = f(D_{cj};\eta)$.

2.2 Sampling Θ_q

Given a density function $f(D;\eta)$, a model M, and a replicate structure $\{F_1, F_2, \cdots, F_k\}$, the sets $\{\Theta_1, ..., \Theta_m\}$ are acquired by sampling from the data. It is often convenient to assume that certain parameters of the distribution of the data are (marginally) identically distributed under all circumstances. In negative binomial modelling of high-throughput sequencing data, for example, the dispersion is commonly assumed to be fixed for any given transcript [30, 13]. This strategy reduces the number of parameters required to be estimated from the data and, especially for low numbers of replicates, will tend to increase the stability of the estimated values. We thus categorise the η_{qk} as either marginally identically distributed over all q and models M, or not.

Suppose that we sample the data for some genetic event h. We first consider the likelihood of the data as the product of the likelihood of the data within each replicate group

$$\mathbb{P}(D_h) = \prod_r f(D_{hF_r}; \eta_{r1}^h, ..., \eta_{rn}^h) \tag{5}$$

and choose η_{rk}^h to maximise this likelihood subject to the constraint that $\eta_{rk}^h = \eta_{sk}^h$ for all r, s if η_{qk}^h is assumed to be marginally identically distributed over all q and models M. For each equivalence class E_q, we then calculate

$$\mathbb{P}(D_{hE_q}) = f(D_{hE_q}; \eta_{q1}^h, ..., \eta_{qn}^h) \tag{6}$$

and maximise this likelihood subject to the constraint that $\eta_{qk}^h = \eta_{rk}^h$ for all k if η_{qk}^h is assumed to be marginally identically distributed over all q and M. This gives a single sampling of values for each $\eta_q^h \in \Theta_q$. We continue sampling (without replacement) to acquire sufficiently large Θ_q.

In both maximisations, we use the Nelder-Mead [25] algorithm as implemented in **R**'s `optim` function. This requires initial values to be provided. For

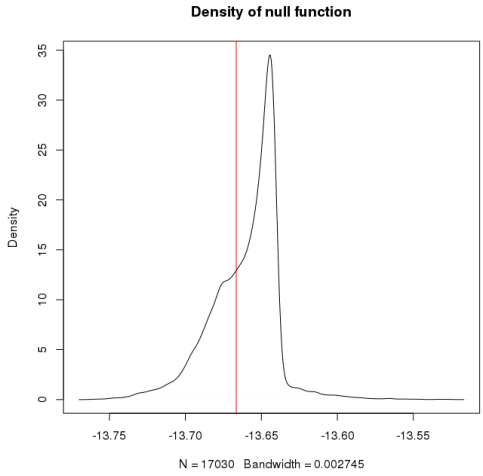

Figure 1: Distribution of the log of the parameter associated with the mean expression (scaled for library scaling factor and gene length) in a set of RNA-seq data derived from rat thymus in juvenile female individuals [37]. The tail of data to the left of the modal peak may be considered to represent non-expressed genes. The red line indicates the threshold level which minimises the intra-class variance.

optimal performance, these initial values should be as close as possible to the solution to the optimisation, and so `baySeq` v2 allows these to be specified as a function of the sampled data D_h.

Maximum likelihood solutions will not always be optimal. In certain circumstances we find increased performance by constraining the domain in which the function to be optimised operates. We give an example of this below when considering a zero-inflated negative binomial model.

2.2.1 Variations on hyperparameter distribution

In sequencing expression of various genomic events, it is not uncommon to find a subset of genomic events that are qualitatively different from the remainder of the data. In mRNA-Seq data, we expect a set of unexpressed genes to which only a small number of reads are assigned, either through sequencing error, misalignment, very low background levels of expression, *et cetera*. Figure 1 shows the distribution of the log of the parameter associated with the mean expression in a set of RNA-seq data assumed to be distributed negative binomially and equivalently across all samples. The tail of data to the left of the modal peak may be considered to represent non-expressed genes.

To distinguish between such qualitatively different events, we can construct additional models in baySeq v2. In the example above, we construct one model (M_{NDE}) for expressed but non-differentially expressed genes, and one model

(M_{NE}) for non-expressed genes. These two models are identical in terms of their equivalence classes, but will differ in the assumed distribution of hyper-parameters.

Two principal options exist for varying the assumed distributions of hyper-parameters between models that share the same equivalence classes. Since the purpose of the two models is to separate two qualitatively different sets of genomic events, we may find some function on the sampled values of hyper-parameters that splits the data. The data shown in Figure 1 can be split by minimising the intra-class variance [27], as shown. Sampled values mapping to the left of the threshold represent the distribution of data for M_{NE} while those to the right represent the distribution of data for M_{NDE}. Supplementary S1 describes an analysis based on such a modelling.

In some cases, the distinction between two quantitatively different models for gene expression introduces a natural choice of hyperparameter. For example, in paired data, a substantial proportion of the data may be equivalently expressed within all pairs, and this may be regarded as a qualitatively different scenario to equivalent expression across replicates but divergent expression within each pair. We have previously shown [14] that these cases can be analysed by constructing a model for equivalent expression across replicates. Assuming a beta-binomial distribution with parameters p, the proportion of counts observed in the first member of each pair, and ϕ, the dispersion, a set Θ_q can be constructed by maximum likelihood methods, as Section 2.2. We can then construct a second model describing equivalent expression within pairs in which the calculated values for ϕ are used for the dispersions but in which the values for p are set to 0.5, the value which corresponds to a hypothesis of balanced expression between pairs.

2.2.2 Bootstrapping weights on hyper-parameters

We can further refine the distributions on the different models by bootstrapping weights attached to the sampled hyperparameter estimate. We begin by adapting Equation 3 to allow for weightings on the sets θ_q associated with a given model, such that for model M the estimated hyper-parameters η_q^h are weighted by w_M^h.

$$\mathbb{P}(D_c \mid M) = \prod_q \frac{1}{\sum_h w_M^h} \sum_h w_M^h \mathbb{P}(D_{cE_q} \mid \eta_q^h) \tag{7}$$

Initially, these weights may be determined by a partition of the estimated hyper-parameters as described above, or may be identical over all hyper-parameters. These weights can then be used to give initial estimates of posterior likelihoods for each model, which can then be used to refine the weightings. Thus, if the hyper-parameters η_q^h are derived from a gene with an estimated posterior likelihood for some model M of p_M^h, the weighting for those hyper-parameters can be updated to $w_M^h = p_M^h$.

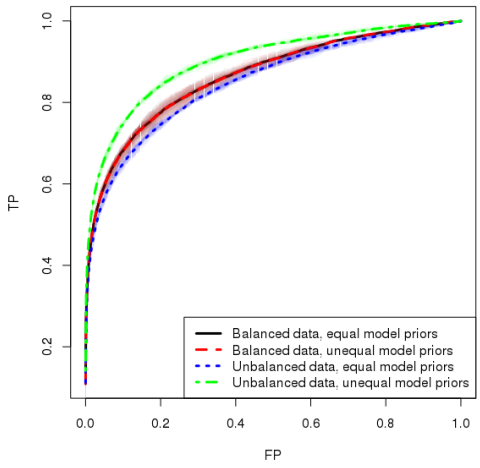

Figure 2: Average ROC curves showing performance of baySeq v2 on balanced and unbalanced differentially expressed data. Allowing unequal model priors for different sets of the data increases performance for unbalanced data.

2.3 Model priors

Model priors may be provided based on prior knowledge, or estimated empirically from the calculated $\mathbb{P}(D_c \mid M)$ values. If estimated empirically, the default behaviour is to calculate the $\mathbb{P}(D_c \mid M)$ for all models M and all c and use the Bayesian Information Criterion (BIC) to choose between each model for each c. The proportion of data for which a model M is selected using the BIC is taken as the prior value $\mathbb{P}(M)$. If no data are selected for a given model, the prior value is set to $1/n$, where n is the total number of genomic events, and the other priors adjusted accordingly. The use of the BIC gives better estimates (Figure S3) of the expected numbers of differentially expressed genomic events than the iterative method described in our previous work [13].

Rather than assume a single value for $\mathbb{P}(M)$, baySeq v2 now allows different subsets of the data to take different values for the model prior. This can substantially improve performance if there are strong reasons to suppose that different subsets of the data will display different proportions of differential expression. We can apply this in a variety of circumstances, but perhaps the most useful is in analyses of unbalanced differential expression, in which differential expression is primarily in one direction.

Figure 2 shows a reanalysis of the simulated data used in Soneson *et al* [33]. The data from which this figure derive consist of simulated data equivalent to 12450 genes from 10 samples, of which approximately 1250 are differentially expressed between the first five and second five samples. In one set of simulations, the differentially expressed genes are equally likely to be up-regulated

as down-regulated between the two groups in the data, while in the other, all differential expression is an up-regulation of the second group relative to the first. Allowing baySeq v2 to choose different model priors depending on which group has higher average expression gives a substantial increase in performance in the unbalanced case, while not affecting performance in the balanced data.

2.4 Computational Strategies

Calculating priors through numerical methods, and posterior likelihoods via Eqn. 3 or Eqn. 7 are computationally expensive steps that scale linearly with both the number of models to be evaluated and the number of genomic events being considered. Several strategies are proposed to mitigate the computational costs involved.

2.4.1 Stratified sampling

A minimum size of the sets Θ_q is required for accurate estimations of the posterior likelihood. The highest accuracy of estimations of posterior likelihood will generally be obtained by making Θ_q as large as possible, but this carries computational costs, making sampling from the data necessary. However, for the numerical approximation described in Eqn. 3 to provide a reasonable approximation to the true value of $\mathbb{P}(D_c|M)$, the sets Θ_q must contain values in the high probability mass regions of $\mathbb{P}(D_{cE_q}|\Theta_q)$. If the sets Θ_q are acquired by sampling uniformly from the data, this can present difficulties for estimating posterior probabilities for D_{cE_q} that lie in the tails of the hyperparameter space of θ_q. Increasing the sample size will resolve this issue, but at a computational cost. Instead, we propose a stratified sampling technique in which the data are stratified by some summary statistic and equal volumes of data are sampled from within each stratum. Each sampling is weighted proportionally to the total number of elements in the stratum such that Eqn. 7

$$\mathbb{P}(D_c \mid M) = \prod_q \frac{1}{\sum_h w_M^h s^h} \sum_h w_M^h s^h \mathbb{P}(D_{cE_q} \mid \eta_q^h) \tag{8}$$

such that s^h is the reciprocal of the stratum size from which the value η_q^h is sampled.

2.4.2 Consensus Priors

For large numbers of models, computational costs can be reduced substantially if we assume that the parameters are identically distributed for all models; that is, that $\Theta_q = \Theta$ for all q. In this case, Eqn. 3 becomes

$$\mathbb{P}(D_c \mid M) = \prod_q \frac{1}{|\Theta|} \sum_{\eta \in \Theta} \prod_{i \in E_q} \mathbb{P}(D_{ci} \mid \eta) \tag{9}$$

and Eqn. 8 becomes

$$\mathbb{P}(D_c \mid M) = \prod_q \frac{1}{\sum_h w_M^h s^h} \sum_h w_M^h s^h \prod_{i \in E_q} \mathbb{P}(D_{ci} \mid \eta^h) \tag{10}$$

The advantage of this formulation is that the values $\mathbb{P}(D_{ci} \mid \Theta)$ are identical for all models; consequently, these need only be calculated only once and the likelihood of the data under any model can then be evaluated by taking the appropriate product-sum-product, considerably reducing the computational cost.

In estimating a set Θ, those parameters assumed to be marginally identically distributed over all q and models M are estimated as previously described in Eqn 5. We then randomly select amongst the replicate sets a single set F_r and maximise the likelihood

$$\mathbb{P}(D_{hF_r}) = f(D_{hF_r}; \eta_1^h, ..., \eta_n^h) \tag{11}$$

as in Eqn. 6, subject to the constraint that $\eta_k^h = \eta_{rk}^h$ for all k, if η_k^h is assumed to be marginally identically distributed over all q and M. This gives a single sampling of values for each $\eta \in \Theta$.

2.4.3 The 'catchall' model

The number of potential models scales exponentially with the number of replicate groups. The number of biologically plausible models will in general be much smaller, and the number of biologically interesting models may be smaller still. In cases where it is not clear which models are biologically plausible, or where the number of plausible models exceeds the number of interesting models, the 'catchall' model provides a useful solution. This model assumes the data within each replicate group is distributed independently. Any data not well characterised by any other specified model will thus be best described by the 'catchall' model. Data for which the catchall model has a high posterior likelihood can then be examined for previously unspecified patterns of expression of interest.

3 Results

The baySeq v2 methods described above can be applied to any high-dimensional data provided a suitable distribution can be defined. To assess the general utility of this approach, we consider several differently distributed data sets for which different analytic methods have been specifically designed. We compare the performance of baySeq v2 to these specific methods in identifying differential expression.

3.1 Affymetrix Microarray Latin Square Data

Microarray data has conventionally been analysed under an assumption of (log) normality. We compare the performance of limma [32], an established method

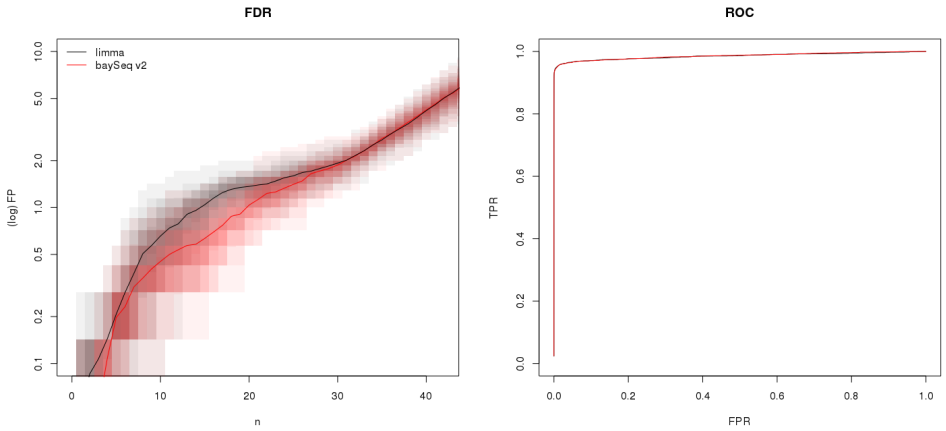

Figure 3: Average false discovery rates and ROC curves for baySeq v2 and limma from sampling of non-overlapping pairs of hybridisations in the Affymetrix HGU133A Latin Square data. Percentiles of false discovery rates across samplings are shown as transparent areas around curves.

for discovery of differential expression, to a baySeq v2 analysis using a normal distribution in which

$$\mathbb{P}(D_{cE_q}|\eta_q) = \prod_{D_{cj} \in D_{cE_q}} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(D_{cj}-\mu_q)^2}{2\sigma^2}} \tag{12}$$

where $\eta_q = (\mu_q, \sigma)$ with the standard deviation σ being assumed constant across different equivalence classes E_q.

Comparisons are made on the Affymetrix HGU133A Latin Square data [1]. These data consist of three technical replicates of 14 hybridisations in a human background, with 42 spiked transcripts at divergent concentrations in each hybridisation. We process the data using the RMA [16] algorithm using the alternate chip description file supplied with the data. Non-differentially expressed control spikes showed highly variable expression across arrays (Figure S5) and were removed from further analysis.

To assess performance of the methods, we select seven non-overlapping pairs of hybridisations in which to identify differential expression, and compute the average numbers of true and false positives for each comparison. We repeat the selection of pairs of hybridisations one hundred times, and show the distribution of false positives against oligonucleotides selected and of ROC curves (Fig. 3).

ROC curves are very similar between the two approaches, while false discovery rates are slightly lower in the baySeq v2 analysis, suggesting that a generalised empirical Bayesian approach can match or exceed the performance of a well-established method for microarray analysis.

3.2 Zero-inflated RNA-Seq data

Zero inflation occurs when two processes operate upon the data, the first, a binary distributed process that defines whether signal is present or absent, the second, a distribution on the size of the signal (which may itself be zero) if a signal is present. Zero-inflated negative binomial data may arise in a number of scenarios using the current generation of high-throughput sequencing technology. In cross-species analyses [5] in which the expression of gene homologues is being compared, some genes may have moved out of a given regulatory pathway and be unexpressed in some organisms. In meta-transcriptomic studies [12] the observed expression of a gene may be driven by a single organism which may or may not be present in the meta-sample. Similarly, in single-cell sequencing, the expression of genes may be much more of a stochastic on/off process than observed in a multi-cell profile [23]. Zero-inflation may also occur in genome-wide enrichment data as a result of low coverage and sequencing bias [28].

We compare the ShrinkBayes package [34], which has been developed to apply a generalised linear model based on a zero-inflated negative binomial model to a baySeq v2 analysis using a zero-inflated negative binomial model in which

$$\mathbb{P}(D_{cE_q}|\eta_q) = \prod_{D_{cj}\in D_{cE_q}} (1-\zeta)g(D_{cj},\mu_q l_j,\phi) + \zeta I_{D_{cj}=0} \tag{13}$$

where $g(D_{cj},\mu_q l_j,\phi)$ is the probability mass function of a negative binomial distribution with mean $\mu_q l_j$ and dispersion ϕ, where l_j is the library scaling factor [13] of library j. $I_{D_{cj}=0}$ is an indicator function which is 1 if $D_{cj}=0$ and 0 otherwise. $\eta_q = (\mu_q,\phi,\zeta)$, with the dispersion ϕ and proportion of zero inflation ζ being assumed constant across different equivalence classes E_q. In the event that no zeros appear in the reported expression for a gene, a maximum likelihood estimation of the ζ parameter (Eqn. 5) will be zero (up to computational precision). Similarly, since a highly dispersed negative binomial variable will be rich in zeros, a maximum likelihood estimation on a zero-inflated gene may report high ϕ and low ζ values. We find improved performance by limiting the domain of the function defined by Eqn. 13 such that $\zeta \geq \max_r\{1-2^{-1|D_{cF_r}|}\}$, that is, ζ must be greater or equal to that proportion of zero-inflation which gives a 50% chance of seeing no zeros within the smallest replicate group.

In the absence of a zero-inflated data set for which true positives are known, we assess performance on a simulated data set. We base this on previous simulations developed to describe high-throughput sequencing data [13, 30] in which data from ten thousand genes in ten samples are simulated from a negative binomial distribution, with means sampled from a SAGE dataset. To explore the effects of increased sequencing depths in zero-inflated background, we scale the mean expression by 1, 3, and 5. Dispersions for each gene are sampled from a gamma distribution with shape = 0.85 and scale = 0.5. Library sizes for each sample are sampled from a uniform distribution between 30000 and 90000. One thousand of the genes are simulated to have an eight-fold differential expression in either direction between the first and second sets of five samples each. For each gene, we then sample a proportion p_c of zero-inflation from a uniform

distribution between 0 and 0.5, and for each sample in that gene, replace the observed value with a zero with probability p_c.

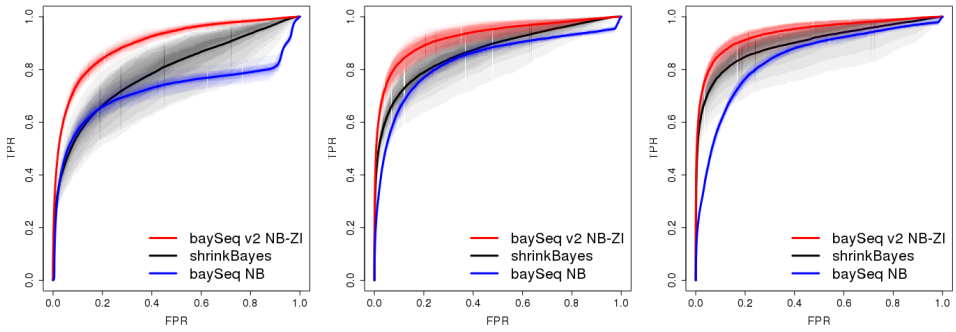

Figure 4: Average ROC curves for baySeq v2 (zero-inflated negative binomial), shrinkBayes and baySeq v2 (negative binomial) analyses of differential expression in zero-inflated negative binomially distributed data. Library scaling factors are increased by factors of 1, 3, and 5. Percentiles of true positive rates across samplings are shown as transparent areas around curves.

Figure 4 shows average ROC curves for the simulations, with mean expression scaled by 1, 3, and 5. In all cases, baySeq v2 with a zero-inflated negative binomial model strongly outperforms ShrinkBayes, which generally outperforms baySeq v2 using a negative binomial model. There is a general improvement in performance with increased sequencing depth for the two methods accounting for zero inflation which is reversed for the method which does not. This suggests that zero inflation becomes increasingly significant with higher sequencing depth, as might be expected.

3.3 Matched sample sequencing

The Rat BodyMap [37] project sequenced RNA-seq data from multiple organs from juvenile, adolescent, adult and aged Fischer 344 male and female rats. For each individual in this study, mRNA is sequenced from every available organ. We use this data to demonstrate a novel analysis based on a multinomial-Dirichlet distribution which allows us to identify changes in relative expression within the tissue types while accounting for individual-specific effects.

For simplicity, we consider the data from ten tissue types (adrenal gland, brain, heart, kidney, liver, lung, muscle, spleen, thymus, and uterus) in female rats, comparing four juvenile (2-week old) to four aged (104-week old) individuals. The data are thus multi-dimensional; for each gene and each individual, we have ten values giving the expression in each organ.

We construct a baySeq v2 analysis using a Dirichlet-multinomial analysis in which $\eta_q = (p_{q1}, p_{q2}, \phi)$. ϕ, the dispersion parameter, is assumed to be constant across equivalence classes. The values p_{q1} and p_{q2} represent the proportion of

expression in the tissues with highest and second highest mean expression in the gene being modelled, with the proportion of expression in the eight remaining tissues being modelled as $p_{qr} = \frac{1-p_{q1}-p_{q2}}{8}$. We adopt this strategy to reduce the dimensionality of the distribution being empirically estimated, and thus prevent the empirical distribution from being too sparse an estimate of the true distribution. We thus calculate the likelihood of the observed data as

$$\mathbb{P}(D_c|\eta_q) = \prod_{D_{cj} \in D_{cE_q}} \frac{\Gamma(\sum_k \alpha_{qjk})}{\Gamma(\sum_k \alpha_{qjk} + D_{cjk})} \prod_{k=1}^{10} \frac{\Gamma(\alpha_{qjk} + D_{cjk})}{\Gamma(\alpha_{qjk})}$$

where

$$\begin{aligned}
\alpha_{qjk_1} &= \tfrac{1}{\phi-1} \tfrac{p_{q1} L_{jk_1}}{\sum_j p_{q1} L_{jk_1}} \quad \text{if} \left\langle \frac{D_{cjk_1}}{L_{jk_1}} \right\rangle_{k_1} \geq \left\langle \frac{D_{cjk}}{L_{jk}} \right\rangle_{k} \forall k \\
\alpha_{qjk_2} &= \tfrac{1}{\phi-1} \tfrac{p_{q2} L_{jk_2}}{\sum_j p_{q2} L_{jk_2}} \quad \text{if} \left\langle \frac{D_{cjk_2}}{L_{jk_2}} \right\rangle_{k_2} \geq \left\langle \frac{D_{cjk}}{L_{jk}} \right\rangle_{k} \forall k \neq k_1 \\
\alpha_{qjk_r} &= \tfrac{1}{\phi-1} \tfrac{p_{qr} L_{jk_r}}{\sum_j p_{qr} L_{jk_r}} \quad \text{otherwise}
\end{aligned}$$

with L_{jk} the library scaling factor for the kth tissue of the jth individual.

We fit three models to these data. The first model describes genes with consistent levels of expression across all tissue types and ages. The second model describes genes with expression consistent between ages, but variable within tissue types. The third model describes genes for which the ratio of expression between tissues varies between juvenile and aged individuals. In the first two of these models, all individuals lie in the same equivalence class.

To distinguish between those genes which have consistent levels of expression across all tissue types and ages and those which have consistent levels of expression between ages but vary within tissue types, we compute priors for a single model of consistent expression between ages. For the model of consistent levels of expression across all tissue types and ages, we take the computed dispersion parameters and set $p_{qk} = \frac{1}{9}$ for all k, while for the model of consistent expression between ages with variable expression between tissues, we use the maximum likelihood estimated values of of p_{qk}. We initially weight the models by partitioning the values p_{q1} estimated for a model of consistent expression between ages (Figure S4) to minimise the intra-class variance [27] and use Eqn 7 to calculate posterior likelihoods based on these weighted values. For the model of consistent expression across all tissue types and ages, w_M^h is 1 if $f(\eta_q^h)$ is less than the partitioning threshold, 0 otherwise; for the remaining two models, w_M^h is 1 if $f\eta_q^h$ is greater than the partitioning threshold, 0 otherwise. We bootstrap these weightings as in Section 2.2.2 over five iterations.

Figure 5 shows the top ranked genes from each of the three models. The expected number of genes conforming to each model may be inferred by summing the posterior likelihoods estimated for each gene for that model. An estimated 133 genes are expected to be consistently expressed across all tissues and timepoints. 4603 genes are estimated as showing variability between tissues, but no

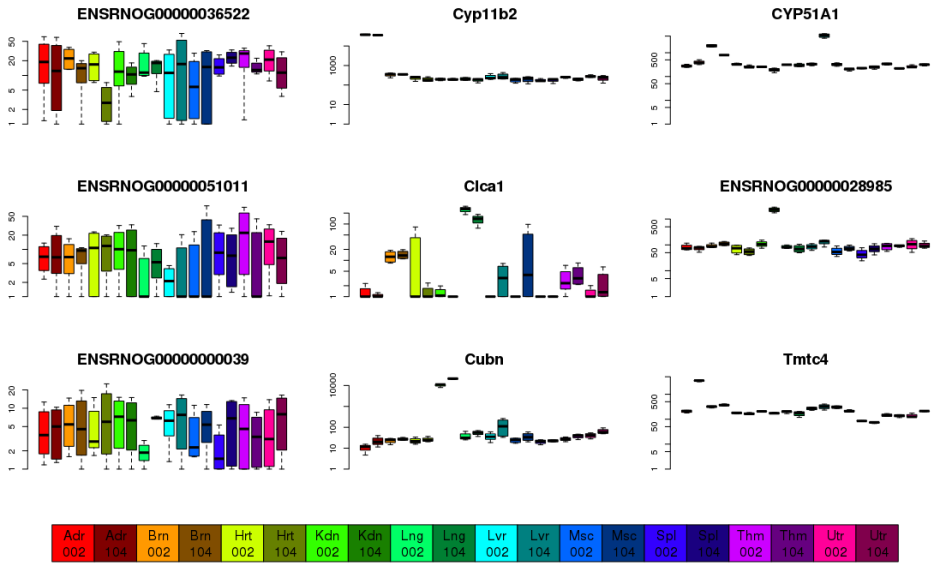

Figure 5: Expression levels of the top three identified genes from female rats for each of three models of expression; consistent expression across tissue types and ages (left), consistent expression across ages but variable between tissues (centre) and variable expression between ages (right).

differential behaviour over time, while 15430 genes are expected to show variable behaviour in one or more tissues over time. This unusually high proportion of differential behaviour may be accounted for by noting that we are considering ten distinct tissues, and differential behaviour in any one of those tissues will be sufficient to identify differential expression.

Analysis of the posterior distributions of the hyper-parameters (Eqn 4, Fig S6) allows a breakdown of the differential behaviour . Of the top 231 genes (selected as controlling family-wise error rate at 10%) that show variability between tissues and no differential behaviour over time, the gene is most abundantly expressed most frequently in brain tissue (36%), and most rarely in uterus tissue (1.2%). Of these 231 genes, 95 show a likelihood greater than 95% of the parameter p_{q2} exceeding the nominal average proportion of $\frac{1}{10}$. These genes thus show an increase in expression in two tissue types relative to the remaining tissues; of these, the most frequent pairing are between heart and muscle tissues (27) and kidney and liver (15).

For those genes that show a change in proportion of gene expression across tissues over time, we are similarly able to breakdown the discovered differential expression. Controlling family-wise error rate at 10%, we discover 10071 genes that show changes over time. The largest category of change (27%) is a reduction of relative expression in thymus tissue over time, presumably as a result of thymic involution [31]. However, in 1073 genes, this reduction in relative expres-

sion in thymus tissue correlates with an increase in relative expression in spleen tissue, suggesting a partial compensation mechanism may be in place. The genes showing a reduction in thymus show a strong enrichment for RNA-binding function [9] (Figure S8, Table S1), potentially linked to age-related processes [22]. Other large categories of change involve large changes in relative expression over time that nevertheless leave the gene maximally expressed in the same tissue (Figure S7).

3.4 Complex modelling and computational time

We next use a subset of the Rat BodyMap [37] data to demonstrate the use of various computational strategies to accelerate a complex modelling analysis. We begin by considering the RNA-seq data for each of the four time points (2, 6, 21 and 104 weeks) in the thymus of female rats. With four different experimental conditions, there are a total of fifteen possible models. This is sufficiently few models that we are able to evaluate, at some computational cost, posterior likelihoods for each model using Eqn 7, with the priors being sampled separately for each model as in Section 2.2. However, we can achieve a significant reduction in computational cost through the use of consensus priors (Eqn 10).

An alternative way to reduce the computational cost of this analysis is to restrict the groups to those which are biologically plausible or interesting, and to use a 'catchall' model to identify all genes which do not fit this model. We will suppose that we are primarily interested in genes which undergo a single change in expression between two consecutive timepoints where this change is maintained for all subsequent timepoints. Together with the 'catchall' model, and a model for non-differentially expressed genes, this requires the evaluation of five models in total. We refer to these models as NDE (no differential expression), LDE (late differential expression), in which change occurs between the third and fourth timepoints, MDE (median differential expression), in which change occurs between the second and third timepoints, EDE (early differential expression), in which change occurs between the first and second timepoints, and 'catchall'. We can achieve further reductions in computational cost by using consensus priors in this analysis.

To compare the performance of these approaches, we assume that the estimated posterior likelihoods for the complete fit without consensus priors are accurate. We can then estimate the number of true positives (and hence, the number of false positives) in each of the models in the restricted analysis for the various approaches as the sum of the posterior likelihoods of the complete fit for the first n selected genes. Figure 6 shows the results of these analyses. There is a marginal increase in false discoveries between the complete fit and the complete fit with consensus priors apparent for the NDE set and the MDE set, but in general, the use of consensus priors appears to cause only minor changes in performance for both the complete and reduced model fit. The false discovery rate does show a clear increase between the complete and reduced model fit, though this is not generally of large magnitude.

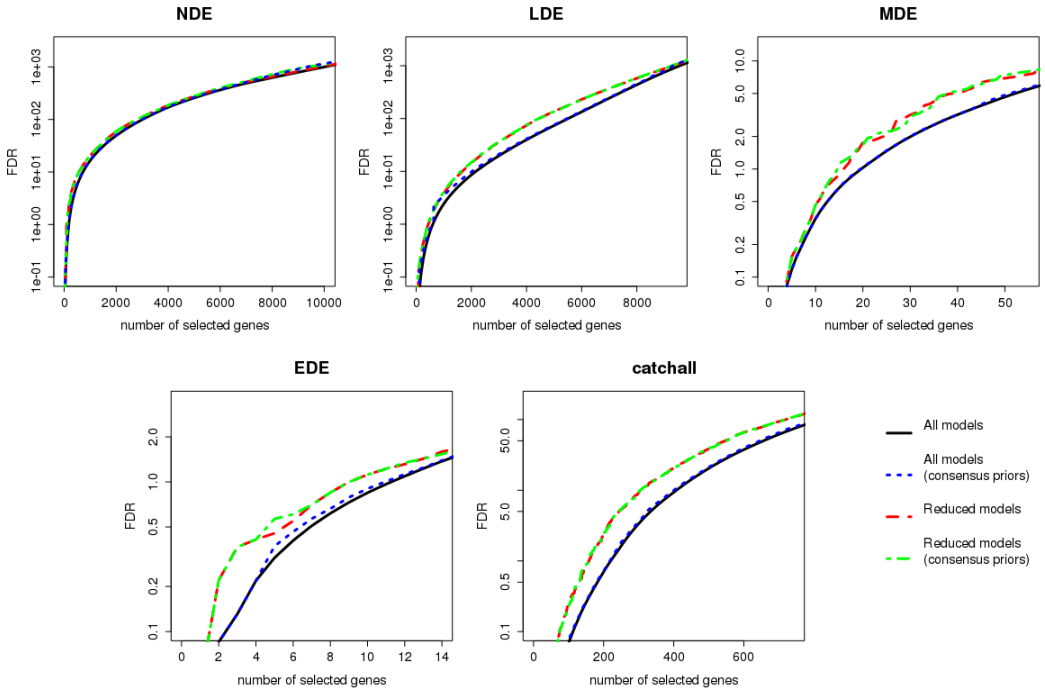

Figure 6: False discovery rates for four different strategies for fitting the five models showing conserved change in expression over time.

The time required for analysis was evaluated on an octo-core (2.50GHz) machine, running in parallel on all cores. Analysis of the complete model fit took 7.0h, while use of consensus priors reduced this to 1.7h. Analysis of the reduced model took 3.3h, while use of consensus priors in the reduced model took 1.4h. Given the similarity of performance between the complete model fit with and without consensus priors, it seems that the use of consensus priors will generally be preferable for the majority of analyses. It is also apparent from these data that the use of consensus priors scales well, with an increase from 5 models in the reduced model set to 15 in the complete set causing only an 18% increase in computational run time.

Using the complete model set allows the identification of further patterns of differential expression over time. Figure 7 shows the expected number of genes for each model, together with normalised and summarised expression values for the top ranked gene in the eight models with highest number of expected genes. These data suggest that, while the majority of genes are not differentially expressed across timepoints, almost as many genes show a change between the fourth time points and the three earlier points. The models with highest numbers of expected genes do generally preserve the ordering of timepoints, however, there are exceptional genes for which the second timepoint is distinct from all

other times ({1,3,4},{2}), and also for which the first and fourth time points differ from the second and third points ({1,4},{2,3}).

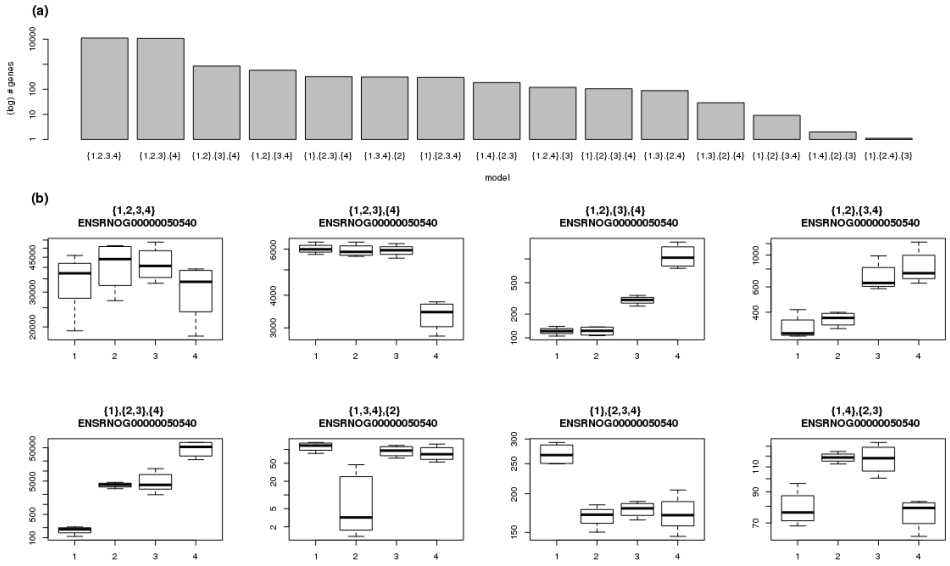

Figure 7: The expected number of genes belonging to each model (a). Normalised expression values of the top ranked genes for the eight models with highest expected number of genes, summarised by timepoint (b).

4 Discussion

We present here a highly flexible solution (baySeq v2) to the general problem of identifying differential behaviour in the 'large p, small n' sets of data that are becoming ubiquitous in biological experimentation (and elsewhere). Given any parametric distributional assumptions for which a hyper-distribution on the parameters can be inferred from the data, posterior likelihoods on diverse patterns of differential expression can be inferred through an empirical Bayesian analysis. We describe here methods to infer the hyper-distribution through maximum likelihood methods, but this is not essential; any methods to infer parameters from the data might be applicable.

We also introduce a number of further refinements to the basic concept. The use of a consensus empirical distribution on the hyper-parameters removes much of the computational cost of these analyses. We demonstrate this on an analysis of complex gene behaviour in a subset of the rat BodyMap data in which the time required for a 15 model analysis of 24750 genes in 16 samples is reduced by 75% through the use of consensus priors with little change in performance. Using these methods, we were able to identify replicated changes in patterns of differential expression over time and show that diverse sets of differential behaviour are present in these data.

Qualitatively distinct data may be distinguished through a weighting or modification of the empirical values representing a hyper-distribution on parameters. We show that this technique allows the identification of unexpressed genes in RNA-Seq data and consistent expression over multiple tissues and time points in matched samples (Figure 5). Bootstrapping can improve the weightings assigned to the sampled values and further improve performance. A natural extension of this approach would be to use distinct distributions for the different models, and this approach is currently under development.

Allowing model priors to vary over different subsets of the data can also improve performance (Figure 2). This may be valuable in a variety of cases where sufficient information is available to distinguish between large categories of genes; for example, if a transcription factor is known to bind to a specific set of gene promoters, this subset of genes is much more likely than its inverse to be differentially expressed if this transcription factor is misregulated). Some care may be needed with this approach in avoiding confirmation bias in downstream analyses of the sets of differentially expressed data.

We demonstrate the effectiveness of this approach by comparison with methods designed specifically for particular distributional assumptions. The limma [32] method is a well-established and widely used method for analysis of microarray data under an assumption of a log-normal distribution. We show on the Affymetrix HGU133A Latin Square data a minor improvement in performance of baySeq v2 over limma, under the same distributional assumptions (Fig 3). We next examined sets of simulated zero-inflated negative binomial data, and show substantial gains in accuracy using baySeq v2 with a zero-inflated negative binomial model over the ShrinkSeq method [34], an approach specifically designed for zero-inflated negative binomial data.

These comparisons do not necessarily imply that the accuracy of baySeq v2 will always match or exceed that of a method specifically designed for a particular set of distributional assumptions, but they do suggest that performance will generally be acceptable. Two major advantages derive from this. Firstly, this approach will substantially reduce development time for the analysis of data with novel distributional assumptions. This reduction in development time appears essential if methods for powerful statistical analyses of these data are to keep pace with the rapid development of new technologies and new applications of those technologies for generating large volumes of biological data. For example, single-cell sequencing appears to include a mixture of Bernoulli and Poisson noise [6], and is likely to require specific distributional assumptions to account for heterogeneity of expression within an individual [17]. The diverse classes of histone modification signatures [10, 15, 4] suggests that differential behaviour in histone modification between samples might be identified by a simultaneous analysis of quantitative values for all histone modifications, perhaps through an assumption of a multinomial-Dirichlet distribution. We describe the development of such a model here as a method for analysing multiple matched samples in RNA-seq data from diverse tissues of rat [37] over time. The results acquired through this analysis correspond on a broad scale to known interactions between tissues and their changes over time, and allow detailed comparison of gene behaviour between tissues. The relative ease with which distributional assumptions can be changed and modified using these methods also allows the rapid incorporation of significant observables into the models; for example GC-content [29], secondary structures [19], mapping uncertainties [18], *et cetera*.

The second major advantage of this approach is that it allows a standardisation of output over the diverse data-types. Furthermore, posterior likelihoods are easily manipulated and compared between analyses. For example, if a set of RNA-seq data is analysed under an assumption of negative binomially distributed data and a set of ChIP-Seq data under an assumption of a multinomial-Dirichlet distribution, it is straightforward to calculate (under an assumption of independence) joint likelihoods of specific patterns of differential expression of RNA-Seq and ChIP-Seq, and thus, for example, to order the set of overlapping gene/histone modifications by the likelihood that both are differentially expressed. Coupled with the capability of baySeq v2 to easily evaluate novel datatypes, this suggests that novel data sets can be readily incorporated with existing analyses.

References

[1] Affymetrix Latin Square Data for Expression Algorithm Assessment.

[2] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, Jan. 2010.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[4] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012.

[5] D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grützner, S. Bergmann, R. Nielsen, S. Pääbo, and H. Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–8, Oct. 2011.

[6] P. Brennecke, S. Anders, J. K. Kim, A. A. KoÅĆodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–5, Nov. 2013.

[7] X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics (Oxford, England)*, 6(1):59–75, Jan. 2005.

[8] B. Domon and R. Aebersold. Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology*, 28(7):710–21, July 2010.

[9] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, Jan. 2009.

[10] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–9, May 2011.

[11] M. Evans and T. Swartz. Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems. *Statistical Science*, 10(3):254–272, Aug. 1995.

[12] R. Fang, B. Wagner, J. K. Harris, and S. A. Fillon. Application of zero-inflated negative binomial mixed model to human microbiota sequence data. Jan. 2014.

[13] T. J. Hardcastle and K. A. Kelly. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, Jan. 2010.

[14] T. J. Hardcastle and K. A. Kelly. Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC Bioinformatics*, in press, 2013.

[15] G. Hon, W. Wang, and B. Ren. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Computational Biology*, 5(11):e1000566, Nov. 2009.

[16] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[17] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lönnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–7, July 2011.

[18] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, Feb. 2010.

[19] J. Li, H. Jiang, and W. H. Wong. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):R50, Jan. 2010.

[20] I. Lönnstedt and T. Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

[21] J. Lovén, D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee, and R. A. Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–82, Oct. 2012.

[22] K. Masuda, Y. Kuwano, K. Nishida, and K. Rokutan. General RBP expression in human tissues as a function of age. *Ageing Research Reviews*, 11(4):423–31, Sept. 2012.

[23] A. McDavid, G. Finak, P. K. Chattopadyay, M. Dominguez, L. Lamoreaux, S. S. Ma, M. Roederer, and R. Gottardo. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, 29(4):461–7, Feb. 2013.

[24] C. Murie, O. Woody, A. Y. Lee, and R. Nadon. Comparison of small n statistical tests of differential expression applied to microarrays. *BMC bioinformatics*, 10:45, Jan. 2009.

[25] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, Jan. 1965.

[26] T. Nilsson, M. Mann, R. Aebersold, J. R. Yates, A. Bairoch, and J. J. M. Bergeron. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature Methods*, 7(9):681–5, Sept. 2010.

[27] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[28] N. U. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12(7):R67, Jan. 2011.

[29] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480, Jan. 2011.

[30] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–7, Nov. 2007.

[31] D. P. Shanley, D. Aw, N. R. Manley, and D. B. Palmer. An evolutionary perspective on the mechanisms of immunosenescence. *Trends in Immunology*, 30(7):374–81, July 2009.

[32] G. K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1544–6115, Jan. 2004.

[33] C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):91, Jan. 2013.

[34] M. A. Van De Wiel, G. G. R. Leday, L. Pardo, H. v. Rue, A. W. Van Der Vaart, and W. N. Van Wieringen. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–28, Jan. 2013.

[35] L. Wang, Y. Xiao, Y. Ping, J. Li, H. Zhao, F. Li, J. Hu, H. Zhang, Y. Deng, J. Tian, and X. Li. Integrating multi-omics for uncovering the architecture of cross-talking pathways in breast cancer. *PloS one*, 9(8):e104282, Jan. 2014.

[36] S. H. Yoon, M.-J. Han, H. Jeong, C. H. Lee, X.-X. Xia, D.-H. Lee, J. H. Shim, S. Y. Lee, T. K. Oh, and J. F. Kim. Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome biology*, 13(5):R37, Jan. 2012.

[37] Y. Yu, J. C. Fuscoe, C. Zhao, C. Guo, M. Jia, T. Qing, D. I. Bannon, L. Lancashire, W. Bao, T. Du, H. Luo, Z. Su, W. D. Jones, C. L. Moland, W. S. Branham, F. Qian, B. Ning, Y. Li, H. Hong, L. Guo, N. Mei, T. Shi, K. Y. Wang, R. D. Wolfinger, Y. Nikolsky, S. J. Walker, P. Duerksen-Hughes, C. E. Mason, W. Tong, J. Thierry-Mieg, D. Thierry-Mieg, L. Shi, and C. Wang. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nature Communications*, 5:3230, Jan. 2014.