Empirical Bayesian Analyses of High-Throughput Sequencing Data

Thomas J. Hardcastle[*]

Abstract

Methods for the analysis of high-throughput sequencing data must exploit the 'large *p*' nature of the data if they are to overcome the small sample sizes that are commonly available. This paper presents a flexible and powerful methodology for analysis of high-throughput sequencing data based on an empirical Bayesian approach.

The methods are demonstrated on two problems in high-throughput sequencing, that of differential expression discovery and of locus detection based on genome-aligned reads. For the application of differential expression, we show that the methods perform at least as well as any alternative approach. In the application to locus discovery, we show how, beginning with an initially poor approximation to the loci, we can use this empirical Bayesian approach to bootstrap to a much improved definition of the loci.

The methods developed here form a general strategy for the analysis of high-throughput sequencing data and may in principle be used with any set of models and distributions for the data. Novel modifications to the basic approach that reduce the computational effort required and increase the performance of these methods are introduced.

Key Words: empirical Bayesian methods, high-throughput sequencing, differential expression, locus detection

1. Introduction

The development of high-throughput sequencing technologies in recent years [3, 12, 13, 18] has led to a massive increase in genomic data represented by *counts*. In the raw form, these data consist of the number of times a particular sequence is observed in a sequenced *library*, whether the source is, for example, genomic DNA, DNA fragments produced by immunoprecipitation, messenger RNA (mRNA) or small RNAs (sRNA). In a number of applications made possible by high-throughput sequencing, we may wish to group multiple tags together and acquire a single count for that grouping. For example, in mRNA-seq analyses, the relevant count is that of the number of sequences that align to a particular gene, exon, or splice variant. Similarly, in analyses of sRNA-seq data, methods exist to group individual sequences into *loci* that represent sRNA precursors. In either case, for each distinct tag or grouping of tags, we have an ordered list, or *tuple*, of discrete counts with the sample order the same in each tuple.

Analyses of high-throughput sequencing data provide a classic example of the 'small *n*, large *p*' problem. In any given sequencing experiment the number of unique sequences present are likely to number in the millions, while, in the cases where a known grouping structure exists (for example, a known transcriptome) the number of tuples is likely to number in the tens of thousands. Conversely, because of the high costs of producing and sequencing biological samples, the number of samples available to any single analysis is likely to be small.

[*] Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge, United Kingdom

A key requirement of analysis methods in high-throughput sequencing data is therefore that they should exploit the 'large p' element of the data, that is, they should 'borrow' information across tuples in order to better analyse each individual tuple. A variety of methods have been suggested to achieve this borrowing of power, primarily in the area of differential expression discovery between samples. These include approaches for stabilisation of overdispersion through shrinkage towards a common estimator [17], and a local regression of shot noise to mean expression [1]. An empirical Bayesian approach is presented here that borrows information by inferring a distribution on the underlying parameters of the data by repeated sampling from within the data. This approach has the advantages that, in addition to increased performance, it allows great flexibility in the choice of models fitted to the data. This approach is demonstrated on the problems of differential expression detection [8] and locus discovery [9] from high-throughput sequencing data.

2. Methods

We begin by defining the concept of equivalence between two libraries on a given tuple. If for tuple i libraries j and k share the same parameters ζ on the distribution of their data then we say that these libraries are equivalent on tuple i. We assume that biological replicates are equivalent on all tuples by definition, and define differential expression between two libraries at a tuple as a non-equivalence between the libraries. Based upon this definition of equivalence, we can construct a set of models upon the data where each model represents a set of *equivalence classes* where, for tuple i, the libraries j and k belong to the same equivalence class if and only if their expression is equivalent at that tuple.

Analyses of the data then depend on an evaluation of the posterior likelihoods of each model at each tuple. Suppose that there exists some model M for the data D_i at the ith tuple. The posterior likelihood of the model M is

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

The likelihood of the data $\mathbb{P}(D_i)$ is easily estimated as the number of models is finite (although potentially large) and so this becomes a scaling factor such that the sum of $\mathbb{P}(M \mid D_i)$ over all models M is unity. The prior likelihood of a model $\mathbb{P}(M)$ is estimated based on the proportion of tuples whose data is best represented by that model (see Section 2.2). The principle challenge is thus to estimate the likelihood of the data given the model.

2.1 Likelihood of data given model

A model M defines equivalence classes $E_1, \cdots, E_m$. For each equivalence class E_q there exists a joint distribution θ_q on the underlying parameters of the data. Assuming independence between the θ_q then the likelihood of the data D_i given the model M can be expressed in terms of the data D_{iq}, the data from tuple i belonging to samples contained within equivalence class q. Then

$$\mathbb{P}(D_i \mid M) = \prod_q \int_{\zeta \in \theta_q} \mathbb{P}(D_{iq} \mid \zeta)\mathbb{P}(\zeta) d\zeta \tag{2}$$

We do not know the distribution θ_q. However, if we have a set Θ_q which samples from this distribution, then we can approximate the likelihood of the data given the model by

$$\mathbb{P}(D_i \mid M) \approx \prod_q \frac{1}{|\Theta_q|} \sum_{\zeta \in \Theta_q} \mathbb{P}(D_{iq} \mid \zeta)) \tag{3}$$

The set Θ_q can be acquired by sampling from the data. For a given model M we sample some tuple whose behaviour approximates this model. The parameters of M for this sampled tuple can be estimated from this tuple, typically through maximum or quasi-maximum likelihood methods. By repeatedly sampling without replacement from an appropriate set of tuples, the set Θ_q for each q is formed.

2.2 Prior likelihoods of models

The prior probabilities of each model $\mathbb{P}(M)$ are required to solve Eqn. 1. If these are estimable from other sources, this may provide the optimum solution. However, in most cases a reasonable estimate of prior probabilities will not be available. Hardcastle & Kelly [8] suggested an iterative approach to estimating the proportion of tuples represented by each model and employing these proportions as the prior. This approach did not account for the propensity of models defined by higher numbers of equivalence groups (and hence higher numbers of parameters) to over-fit, resulting in an over-estimation on the proportion of data represented by such models. An alternative approach that better estimates the proportion of tuples represented by each model is suggested here.

Given the Θ_q, Eqn. 3 defines the likelihood of a tuple i for a given model M. If we calculate this likelihood for all biologically plausible models on the ith tuple, we can use the Bayesian information criterion (BIC) to select amongst the models. Given this selection for a representative (and approximately independent) set of tuples, we can calculate the observed proportion of the set selected as best represented by a given model. As before, we use these proportions as prior likelihoods of the models. Figure 1 compares these methods of prior estimation to the true proportion of differentially expressed tuples in the simulated data suggests that the iterative method tends to over-estimate the proportion of differentially expressed genes, whilst the BIC method tends to under-estimate this proportion. However, for increasing numbers of libraries, the BIC method gives substantially better estimation of the true proportion than the iterative methods.

2.3 Bootstrapping a weighted sampling

In general, it is not possible to know in advance a set of tuples whose behaviour is unambiguously defined by one model M. However, if the likelihood p_k that tuple k is represented by model M is known, the approximation made in Eqn. 3 by weighting the numerical integration (following Evans & Swartz [6]) as

$$\mathbb{P}(D_i \mid M) \approx \prod_q \frac{1}{\sum p_k} \sum_{\zeta_k \in \Theta_q} p_k \mathbb{P}(D_{iq} \mid \zeta_k)) \qquad (4)$$

The values for p_k can be estimated from Eqn. 1. Given an intial (unweighted) approximation Θ_q it is thus possible to estimate the p_k for each sampled tuple k, allowing an improved approximation of the estimated values for $\mathbb{P}(D_k \mid M)$ and hence of $\mathbb{P}(M \mid D_k)$. This allows a further improvement in the calculation of the weights used in Eqn. 4. Thus, by iteratively bootstrapping from a relatively poor approximation Θ_q, a closer approximation to a sampling from the true θ_q can be acquired.

2.4 Stratified sampling

In general, both the estimation of parameters and the solution to the numerical integration given in Eqns 3 and 4 will require a numerical approach. When sampling from a large

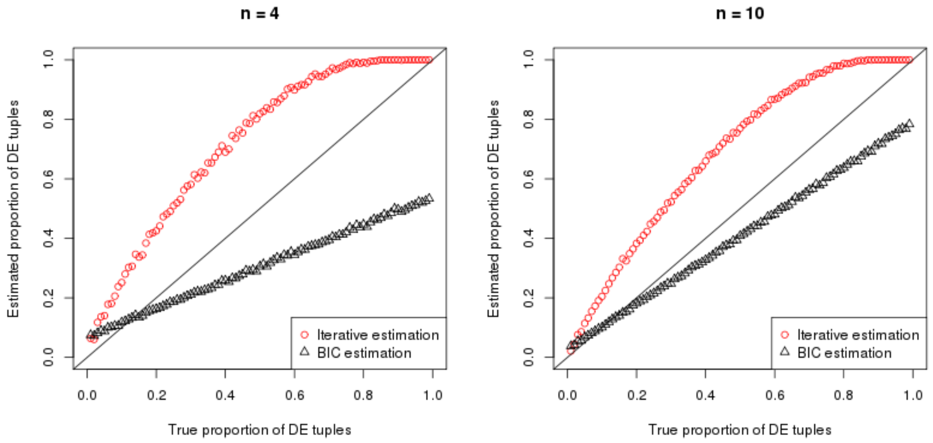

Figure 1: Comparison of methods for estimation of prior likelihood of models in an analysis of pairwise differential expression from simulated data with either 4 or 10 libraries. The proportion of differentially expressed genes, used as the prior likelihood for a model of differential expression is estimated either through the iterative methods described in Hardcastle & Kelly [8] or through the BIC estimation described here.

number of tuples, it is computationally intensive to sample a large proportion of the data. For the most part, good approximations to the underlying distributions θ_q can be acquired through a sampling of (on the order of) ten thousand tuples. However, where the underlying distribution contains long tails in the distribution of one or more parameters, it is possible that the sampling will fail to adequately describe the distribution in these tails.

Figure 2 shows an example of this in which one of the parameters is (up to scaling) an estimate of the mean expression of the tuples. While the distribution estimated from a subsampling of the data fits closely to that estimated from all data, it fails to fit in the tail as no tuples were sampled in this region. The distribution acquired by subsampling will provide poor estimates for the likelihood of the models for those tuples whose parameters lie far outside the range of the samplings.

This problem can be overcome by adopting a stratified sampling approach to constructing the Θ_q. In this instance, the tuples are split by average expression into one thousand quantiles from which an equal number of tuples are sampled (or the total number of tuples in each quantile, where this is lower. The sum over the sampled values used to estimate $\mathbb{P}(D_i \mid M)$ in Eqns 4 is then weighted as

$$\mathbb{P}(D_i \mid M) \approx \prod_q \frac{1}{\sum p_k w_k} \sum_{\zeta_k \in \Theta_q} w_k p_k \mathbb{P}(D_{iq} \mid \zeta_k)) \tag{5}$$

where, if ζ_k is sampled from a stratum of size S and a total of s values are sampled from this stratum, $w_k = \frac{s}{S}$.

3. Applications

The approach described above is highly flexible both in terms of the modelling structure, the distributional assumptions, and the parameterisation of those distributions. The appli-

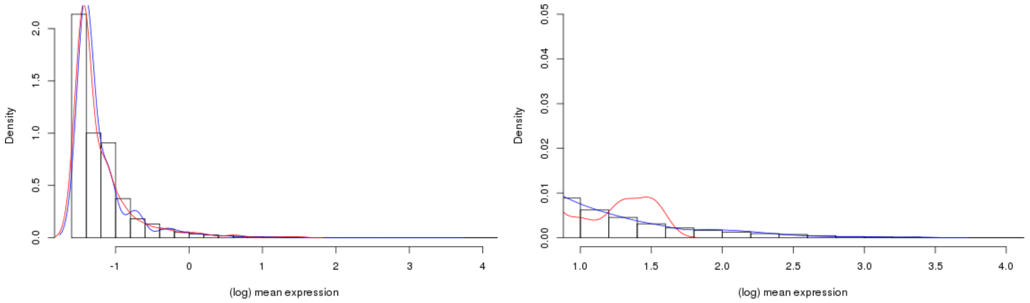

Figure 2: Sampling from a set of high-throughput sequencing data. The true log-mean expression (scaled by normalising factors) of the whole dataset is shown as a histogram. The density of the sampled distribution is shown in red, whilst the (weighted) density of the stratified sampling is shown in blue. Whilst both samplings approximate the distribution of the true data well in the body of the distribution, the stratified sampling approximates the true distribution in the tail much more closely.

cation of these methods to two significant aspects of high-throughput sequencing analysis is discussed below.

The empirical Bayesian methods described above do not imply and specific distributional assumptions about the data. However, in applying these methods, some appropriate choice of distributions based on the methods by which the data are produced must be made. High-throughput sequencing data are acquired through processes that give rise to Poisson distributed data if biological variation between samples is ignored. Biological variation leads to over-dispersion relative to the Poisson distribution. Since under-dispersion will not occur within these data, the negative binomial is commonly used [1, 8, 16] to model the data. A sample A_j has an associated 'library scaling factor' l_j, where this is chosen to account for the variation in sequencing depths between libraries [4, 17]. A scaling factor for the tuple, λ_i, may also be used. This is usually descriptive of the length of the genomic object described by the tuple i but might also be used to correct for GC enrichment biases [15] or other tuple specific factors [1].

The count u_{ij} observed at sample j for tuple i is then assumed to be distributed negative binomially, with mean $\mu_q l_j \lambda_i$ and dispersion ϕ_q. Then the parameterization of this distribution can be defined as

$$\mathbb{P}(u_{ij}; \lambda_i, l_j, \phi_q, \mu_q) = \frac{\Gamma(u_{ic} + \phi_q^{-1})}{\Gamma(\phi_q^{-1}) u_{ic}!} \left(\frac{1}{1 + \lambda_i l_j \mu_q \phi_q} \right)^{\phi_q^{-1}} \left(\frac{\lambda_i l_j \mu_q}{\phi_q^{-1} + \lambda_i l_j \mu_q} \right)^{u_{ic}} \tag{6}$$

If we assume independence between the u_{ij} conditional on μ_q, ϕ_q then

$$\mathbb{P}(D_{iq} | \mu_q, \phi_q) = \prod_{j \in E_q} \mathbb{P}(u_{ij}; \lambda_i, l_j, \phi_q, \mu_q) \tag{7}$$

The joint distributions θ_q on the parameters (μ_q, ϕ_q) thus defines the likelihood of the data (via Eqn. 2) under the models.

[1] It is also possible to replace the tuple-specific scaling factor with one that is tuple/library specific; i.e., λ_{ij}. This can be useful, for example, in comparing gene expression between multiple species whose orthologues have different genomic lengths.

3.1 Differential Expression Detection

The first application to high-throughput sequencing of this empirical Bayesian approach was in the detection of differential expression within defined tuples [8]. Patterns of expression are defined in terms of equivalence of expression between one or more samples, allowing multiple patterns of expression to be examined within a single analysis.

In the simplest case of a model for pairwise differential expression, suppose that there exist sets of samples from conditions A and B. In the case where two biological replicates exist for each condition, there are four libraries, A_1, A_2, B_1, B_2, where A_1, A_2 and B_1, B_2 are biological replicates. In most cases, it is reasonable to suppose that at least some of the tuples may be unaffected by our experimental conditions A and B. The count data for each sample in these tuples will then share the same underlying parameters. The model for non-differential expression is thus defined by the equivalence class $\{A_1, A_2, B_1, B_2\}$.

However, some of the tuples may be influenced by the different experimental conditions A and B. For such a tuple, the data from samples A_1 and A_2 will share the same set of underlying parameters, the data from samples B_1 and B_2 will share the same set of underlying parameters, but, crucially, these sets of parameters will not be identical. The model for differential expression between condition A and condition B is defined by the equivalence classes $\{A_1, A_2\}$ and $\{B_1, B_2\}$.

3.1.1 Sampling Θ_q

If E_q is an equivalence class, then the sampled set Θ_q might be derived by sampling some tuple k and, from the data D_{kq} associated with that equivalence class and tuple, estimating values for (μ_k, ϕ_k). However, suppose that the tuple sampled shows genuine differential expression within the samples defined by E_q. The estimate of dispersion is then likely to be substantially over-estimated. Since it is not known in advance which tuples are genuinely differentially expressed, it becomes difficult to estimate the dispersions. However, it can be assumed by definition that there is no differential expression within sets of biological replicates. This allows estimates of dispersion based on the replicate structure of the data. Consider the sets $\{F_1, \cdots F_s\}$ where $i, j \in F_r$ if and only if sample A_j is a replicate of A_i.

Given this structure for the data, the dispersion of the data for the kth tuple is estimated by quasi-likelihood methods [14] by initially defining $\hat{\mu}_{kr} = \langle\{\frac{u_{kj}}{l_j \lambda_k} : j \in F_r\}\rangle$, and choosing ϕ_k such that

$$2\sum_r \sum_{j \in F_r} \left\{ u_{kj} \log\left[\frac{u_{kj}}{l_j \lambda_k \hat{\mu}_{kr}}\right] - (u_{kj} + \phi_k^{-1}) \log\left[\frac{u_{kj} + \phi_k^{-1}}{l_j \lambda_k \hat{\mu}_{kr} + \phi_k^{-1}}\right] \right\} = n - 1 \tag{8}$$

This value for ϕ_k is then used to re-estimate the values $\hat{\mu}_{kr}$ by maximum likelihood methods, choosing the values for $\hat{\mu}_{kr}$ that maximise the likelihoods

$$\mathbb{P}(D_{kr} \mid \phi_k, \hat{\mu}_{kr}) = \prod_{j \in F_r} \frac{\Gamma(u_{kj} + \phi_c^{-1})}{\Gamma(\phi_c^{-1}) u_{kj}!} \left(\frac{1}{1 + l_j \lambda_k \hat{\mu}_{kr} \phi_c}\right)^{\phi_c^{-1}} \left(\frac{l_j \lambda_k \hat{\mu}_{kr}}{\phi_c^{-1} + l_j \lambda_k \hat{\mu}_{kr}}\right)^{u_{kj}} \tag{9}$$

for each r. Iterating on these estimations of ϕ_k and $\hat{\mu}_{kr}$ until convergence defines the value for ϕ_k for the kth tuple. The value μ_{kq} can then be estimated for any equivalence class E_q by fixing the dispersion parameter as ϕ_k and finding the value of μ_{kq} that maximises the likelihood of the associated data D_q. The set $\Theta_q = \{(\mu_{kq}, \phi_k)\}$ is thus acquired by repeating this process for multiple k.

This method of estimating the dispersion assumes that the dispersion of a tuple is constant across different sets of samples. In most cases, where the number of samples is low, this is likely to be the best approach. Where there is some expectation that the dispersion will be substantially different between sets of replicates, there may be advantages to estimating the dispersions individually for each of the different sets of samples in each model, while still considering the replicate structure within these sets. This is easily done by restricting the data (and corresponding replicate structure) to D_{qc} when estimating the dispersion in Eqn. 8. In general, no substantial differences between these approaches is found in simulation studies.

3.1.2 Comparison of Methods

Comparisons of these empirical Bayesian methods for detection of differential expression on both real and simulated datasets to alternative methods for the detection of differential expression may be found in [8]. In general, these methods perform at least as well, and often better than, alternative methods for the detection of differential expression. Figure 3 shows a comparison of methods for detection of pairwise differential expression on simulated data. Where few libraries are available, these methods (implemented in the `baySeq` R package) show equal performance with the other best-performing method, `edgeR` [17]. Where a greater number of libraries are available, the empirical Bayesian methods outperform all alternative methods tested. Independent comparisons are carried out in [5] and [10] and support the claim that the empirical Bayesian methods are amongst the best available for detection of differential expression.

3.2 Locus Detection

The same general approach to analysis of high-throughput sequencing has also been applied to the problem of locus detection [9]. This study was motivated by the sequencing of small RNAs (sRNAs), which arise from some longer precursor element. This precursor cannot itself be sequenced due to its transitory nature, however, small RNAs, if stabilised by association with an RNA induced silencing complex (RISC) [7] can be sequenced with relative ease. Where sRNAs derive from the same precursor, the sequenced reads align to the genome in close proximity to each other and with non-independent abundances. However, because the affinity of the RISC to individual sRNAs is highly variable we see strong accumulation biases of sRNAs upon the genome. In combination with the presence of background noise as a result of sequencing errors and the presence of breakdown products from longer RNA molecules, amongst other factors, this makes the precise definition of locus boundaries a non-trivial task.

In a full analysis of small RNA loci, methods are required for the analysis of data from multiple experimental conditions, in which small RNAs from a particular locus may be expressed under some conditions but not others. This is achieved in Hardcastle *et al* [9] by first calculating likelihoods for each set of experimental conditions independently and then taking an algorithmic approach to combine the loci called for each set of experimental conditions into a single set of defined loci. Here we consider only the application of the empirical Bayesian methods and so restrict our analysis to a consideration of data from replicate group r.

The first requirement of such an analysis is the ability to evaluate the likelihood that some defined region may be considered as showing locus-type expression within replicate

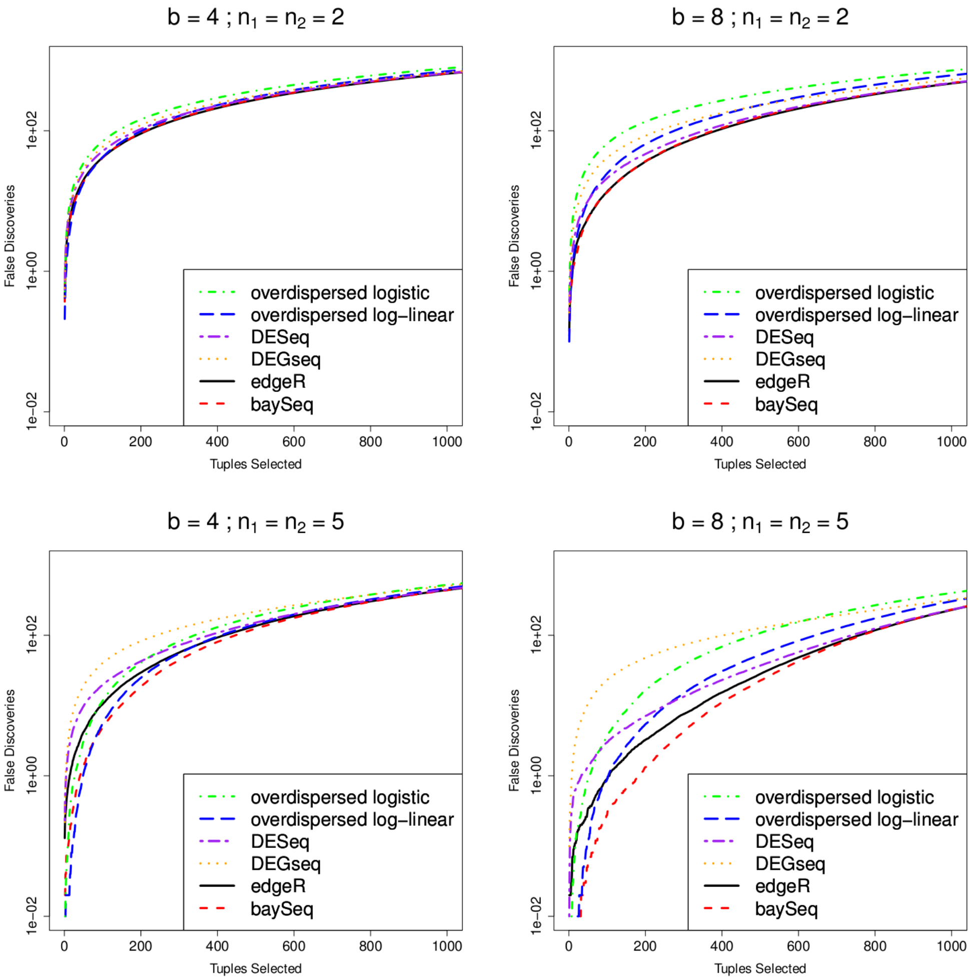

Figure 3: Comparison of performance of various methods for detection of pairwise differential expression in simulations of high-throughput sequencing data, showing the number of false discoveries identified in the first *n* tuples selected by each method. The logistic [2] and log-linear [11] methods are classical methods that account for over-dispersion but do not borrow power between tuples. The `DEGseq` [19], `DESeq` [1] and `edgeR` methods [17] are methods developed for high-throughput sequencing and all make use of the 'large *p*' nature of the data to borrow power between tuples in some manner. The `baySeq` method is the implementation of the empirical Bayesian methods [8] described here.

The simulations used to generate these data are described in [8]. Briefly, in each simulation there exist ten thousand tuples of which one thousand are differentially expressed. The fold change between differentially expressed tuples is given by the parameter *b*. The number of libraries in the first and second group of biological replicates is n_1 and n_2 respectively.

group r. There are thus two models for each tuple, the first, M_N, modelling the data as being expressed under background, or null-type conditions, the second, M_L, as being expressed under locus-type conditions.

3.2.1 Sampling a weighted Θ_r

The two models of locus-type and null-type expression both define equivalence of expression between all samples within a replicate group. Simply sampling randomly selected regions of the genome is therefore not sufficient to define the sets Θ_r for each model. Instead, we use the weighted approach defined by Eqn. 4.

An heuristic approach is used to give a first approximation to those regions of the genome that may be defined as loci. From this, a non-overlapping set of 'segments' that covers the genome can be defined, with each segment identified as either a locus or a null. If the kth segment is sampled, then the parameters ϕ_k and μ_{kr} can be calculated as in Eqns. 8 and 9. For both models, Θ_r is formed from $\{\mu_{kr}, \phi_k\}$. However, for the model of locus-type expression, the initial weighting p_k is one if the kth segment is defined as a locus by the heuristic method, and zero otherwise. Similarly, for the model of null-type expression, the initial weighting p_k is one if the kth segment is defined as a null by the heuristic method, and zero otherwise.

The weightings on the sets Θ_r allow an initial estimate of the likelihood that each segment is a locus through an application of Eqn. 4 (or, more usually, Eqn. 5 as the total number of segments is likely to be large). However, the initial definitions given by the heuristic method for each segment as a locus or null may be incorrect. The bootstrapping procedure described in Section 2.3 is thus applied to refine the weightings given these initial estimates until the estimated likelihoods converge.

3.2.2 Likelihoods of Classification

The process defined above gives the likelihoods, for an existing set of segments, that each segment is a locus; that is, that it is a region associated with sequenced reads. These segments are initially defined heuristically, however, given these likelihoods, the empirical Bayesian methodology can be applied to refine the classification.

Suppose that there exists some set of candidate loci for a given replicate group. Two models exist for these data; a model $M_{L'}$ in which the candidate locus exists within some true locus, and a model $M_{N'}$ in which the candidate locus exists within a null region. For each candidate, the likelihood of that candidate locus existing within some true locus can be calculated, given an appropriate sampling of Θ'_r.

The set Θ'_r is derived by sampling a non-overlapping set of candidate loci that lie wholly within a segment of an existing locus map. Suppose that a sampled candidate locus l_k lies within a segment s_m. The segment s_m has an estimated likelihood of being a locus (as defined above) of p_m. For the model $M_{L'}$ the set Θ'_r is formed from the parameters $\{\mu_{kr}, \phi_k\}$ estimated from the data D_{kr} (via Eqns. 8 and 9) associated with the candidate locus l_k, and weighted by p_m, the likelihood that the containing s_m is a true locus. For the model $M_{N'}$ the set Θ'_r is formed from the parameters $\{mu_{kr}, \phi_k\}$ as before, but weighted by $1 - p_m$, the likelihood that the containing s_m is a true null. It is thus possible to calculate for any given candidate locus the likelihoods of these competing models. These likelihoods are combined (see Hardcastle *et al*, 2011 [9]) to define a new segmentation of the genome. Figure 4 compares the loci defined by these methods to the initial heuristically defined loci

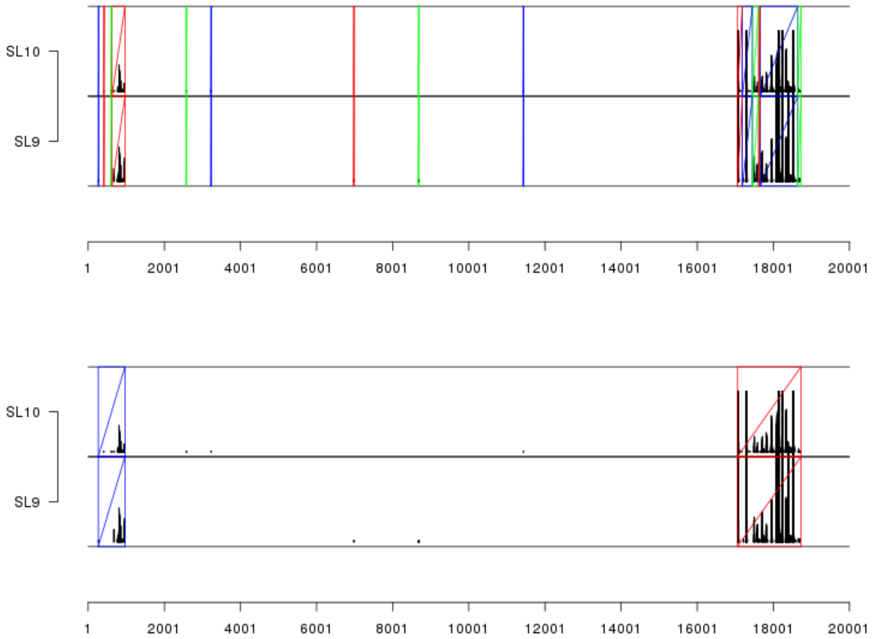

Figure 4: Plots of the first twenty thousand bases on chromosome one of *Arabidopsis thaliana* showing the small RNA loci discovered by heuristic methods (top) and empirical Bayesian methods (bottom) on the basis of a pair of biological replicates. Small RNA reads are mapped back to the genome; the number of reads at any base upon the genome is plotted in black while the coordinates of the loci are shown as red, green and blue rectangles. Note the over-segmentation that occurs in the heuristic methods, and the (unreplicated) background noise visible between bases 2000 and 12000.

on a small region of the genome. There is a clear gain in performance acquired through the empirical Bayesian methods over the initial approximation acquired through heuristic methods.

4. Discussion

We present here a general strategy for analysis of count data from high-throughput sequencing data. In summary, this method establishes a set of models upon the genome. Each of these models describes some distribution for the data within a given tuple, the parameters of which are distributed according to some unknown, model specific distribution. These distributions can be estimated empirically from the data, exploiting the 'large p' nature of high-throughput sequencing, by sampling from a set of tuples which approximate the model under consideration and estimating the parameters on the distribution of the data through maximum or quasi-maximum likelihood methods. Given this approximate distribution on the parameters, the likelihood of each model can be calculated.

This strategy can be applied to any parameterisable distribution for high-throughput sequencing data, and any set of models can in theory be established. Two factors act to complicate the extent which these methods are generalisable to any situation. The first is that the parameterisation of the distributions for the data must be such as to allow a good estimation of the parameters from the data. This can be seen in the parameterisation used for the negative binomial distribution in Eqn. 6. If rather than the dispersion ϕ we were to use the 'size' parameter $\frac{1}{\phi}$, for many tuples (with dispersion close to zero) this would result in extremely large and sparsely distributed estimates for this size parameter. Since it is generally required that the parameters be calculated numerically, such a situation is likely to lead to reduced precision in calculating the parameters, leading to reduced accuracy in likelihood estimation. A second difficulty emerges in sampling from sets of tuples that approximate any given model being considered. This problem is reduced by weighting the sampling (Eqns. 4 and 5) and bootstrapping from an initial weighting to acquire an improved approximation. Nevertheless, this requires an initial approximation (usually derived heuristically) in order to begin the bootstrapping process.

These factors can usually be addressed with an appropriate parameterisation and sampling strategy. Where this is done, the application of the empirical Bayesian approach gives both high performance and flexibility. In the application of these methods to discovery of pairwise differential expression in high-throughput sequencing data, they perform as well or better than any alternative approach. Moreover, unlike the majority of competing methods developed for high-throughput sequencing analysis, this approach has the advantage that multiple models for diverse patterns of differential expression can be evaluated simultaneously (see Hardcastle & Kelly [8] for examples).

The empirical Bayesian approach has also successfully been applied to the problem of sRNA locus detection in replicated high-throughput sequencing data. Several points of interest are exemplified by this application of the methods. In sRNA locus detection, the two competing models for count data in a tuple, that of expression from a null, or background region of the genome, and that of expression from a locus. These two models have the same pattern of equivalence and consequently the sampling used to generate approximations to the distributions thus becomes of prime importance. In the first instance, a heuristic method is used to give an initial approximation to the loci and null segments, from which a bootstrapped estimation of likelihoods can be calculated. These likelihoods are then used to weight the sampling of candidate loci used to define a second pair of models that address the likelihood that a candidate locus lies within a true locus; significantly, the weightings of the sampled loci are derived from the containing segment rather than the candidate locus itself.

The basic methods here are applicable to any model-based analysis of high-throughput sequencing data. The prime restriction to their use is the heavy computational requirements of numerical estimation of parameters through maximum likelihood methods, and the numerical integration used to approximate the likelihood of data given the model. However, the methods are embarassingly parallel and hence, given sufficient computational power there is no practical limit on their use. These methods thus form a highly flexible and generalisable strategy for analysis of high-throughput data of all kinds. Furthermore, by applying these methods to diverse types of high-throughput sequencing data (e.g., mRNA-Seq, methyl-Seq, sRNA-Seq, *et cetera*) it becomes simple to compare results from these diverse data types. This allows for integrated downstream analyses of these data, which is of great value in a systems approach to biology.

References

[1] Simon Anders and Wolfgang Huber, *Differential expression analysis for sequence count data.*, Genome Biology **11** (2010), no. 10, R106.

[2] K A Baggerly, L Deng, J S Morris, and C M Aldaz, *Overdispersed logistic regression for SAGE: modelling multiple groups and covariates*, BMC Bioinformatics **5** (2004), 144.

[3] D R Bentley, *Whole-genome re-sequencing*, Curr. Opin. Genet. Dev. **16** (2006), 545–552.

[4] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit, *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.*, BMC Bioinformatics **11** (2010), no. 1, 94.

[5] Francesca Cordero, Marco Beccuti, Maddalena Arigoni, Susanna Donatelli, and Raffaele A Calogero, *Optimizing a massive parallel sequencing workflow for quantitative miRNA expression analysis.*, Plos One **7** (2012), no. 2, e31630.

[6] Michael Evans and Tim Swartz, *Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems*, Statistical Science **10** (1995), no. 3, 254–272 (EN).

[7] S M Hammond, E Bernstein, D Beach, and G J Hannon, *An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells.*, Nature **404** (2000), no. 6775, 293–6.

[8] Thomas J Hardcastle and Krystyna A Kelly, *baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.*, BMC Bioinformatics **11** (2010), no. 1, 422.

[9] Thomas J Hardcastle, Krystyna A Kelly, and David C Baulcombe, *Identifying small interfering RNA loci from high-throughput sequencing data.*, Bioinformatics **28** (2012), no. 4, 457–63.

[10] Vanessa M Kvam, Peng Liu, and Yaqing Si, *A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data.*, American Journal of Botany **99** (2012), no. 2, 248–56.

[11] J Lu, J K Tomfohr, and T B Kepler, *Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach*, BMC Bioinformatics **6** (2005), 165.

[12] E R Mardis, *The impact of next-generation sequencing technology on genetics*, Trends Genet. **24** (2008), 133–141.

[13] M Margulies, M Egholm, W E Altman, S Attiya, J S Bader, L A Bemben, J Berka, M S Braverman, Y J Chen, Z Chen, S B Dewell, L Du, J M Fierro, X V Gomes, B C Godwin, W He, S Helgesen, C H Ho, G P Irzyk, S C Jando, M L Alenquer, T P Jarvie, K B Jirage, J B Kim, J R Knight, J R Lanza, J H Leamon, S M Lefkowitz, M Lei, J Li, K L Lohman, H Lu, V B Makhijani, K E McDade, M P McKenna, E W

Myers, E Nickerson, J R Nobile, R Plant, B P Puc, M T Ronan, G T Roth, G J Sarkis, J F Simons, J W Simpson, M Srinivasan, K R Tartaro, A Tomasz, K A Vogt, G A Volkmer, S H Wang, Y Wang, M P Weiner, P Yu, R F Begley, and J M Rothberg, *Genome sequencing in microfabricated high-density picolitre reactors*, Nature **437** (2005), 376–380.

[14] J A Nelder, *Quasi-likelihood and psuedo-likelihood are not the same thing*, Journal of Applied Statistics **27** (2000), 1007–1011.

[15] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit, *GC-content normalization for RNA-Seq data.*, BMC Bioinformatics **12** (2011), no. 1, 480.

[16] M D Robinson and G K Smyth, *Small-sample estimation of negative binomial dispersion, with applications to SAGE data*, Biostatistics **9** (2008), 321–332.

[17] Mark D Robinson and Alicia Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.*, Genome Biology **11** (2010), no. 3, R25.

[18] S C Schuster, *Next-generation sequencing transforms today's biology*, Nature Methods **5** (2008), 16–18.

[19] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang, *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.*, Bioinformatics **26** (2010), no. 1, 136–8.