Pardiff: Inference of Differential Expression at Base-Pair Level from RNA-Seq Experiments

Bogdan Mirauta[1], Pierre Nicolas[2], and Hugues Richard[1]

[1] Génomique des microorganismes, UPMC and CNRS UMR7238, Paris, France
{bogdan.mirauta,hugues.richard}@upmc.fr
http://www.lgm.upmc.fr/parseq
[2] Mathématique Informatique et Génome, INRA UR1077, Jouy-en-Josas, France
pierre.nicolas@jouy.inra.fr

Abstract. In the field of RNA-Seq transcriptomics, detecting differences in expression levels between two data-sets remains a challenging question. Most current methods consider only point estimates of the expression levels, and thus neglect the uncertainty of these estimates. Further, testing for differential expression is often done on predefined regions. Here, we propose Pardiff, a method that reconstructs the profile of differential expression at a base-pair resolution and incorporate uncertainty via the use of a Bayesian framework. This method is built on our approach, Parseq, to infer the transcriptional landscape from RNA-seq data.

A program, named Pardiff, implements this strategy and will be made available at: http://www.lgm.upmc.fr/parseq/.

1 Introduction

Various technologies allowing genome wide profiling of the transcriptional activity have emerged in the last two decades. The microarray technology first monitored those types of changes at the gene, and then down to the exon level [7], [6]. More recently, the surge in throughput from sequencing technologies, like Illumina or SOLiD, has raised this resolution to the basepair. Deep sequencing of the transcriptome (RNA-Seq) consists in random shearing of transcripts, followed by amplification and sequencing of the RNA population. An RNA-Seq experiment produces millions of reads which, after alignment to the reference genome, serve to estimate the expression landscape [15].

One traditional question is, starting from two or more controlled experiments, to identify the set of elements that exhibit differential expression (DE). Answering this question is an important step towards formulating a biological hypothesis or for instance deriving disease biomarkers. In statistical terms, the question translates into detecting significant changes of expression level, after accounting for the sources of experimental and biological variability. However, this traditional statistical standpoint does not consider the magnitude of the effect, and authors proposed to overcome this limitation by directly testing whether fold change is above a given level [13].

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 418–427, 2013.

The analysis of DE usually starts from predefined units of possible change, such as genes, exons or transcript isoforms. In this case, after cumulating counts at unit level, one can estimate the statistical significance of differences [1], [17], [18]. Replicating the data sets permits the control of expression variability and mitigates the incertitude on expression level estimation. A challenging problem arises when these units of potential DE are not known. In this case, one faces the problem of having to jointly delineate the boundaries of the DE regions and estimating the magnitude of the DE.

Approaches to tackle this problem group in two categories. The first category consists of estimating the transcript structures from the different datasets (used separately or jointly) and then applying DE detection methods on predefined units. The second category computes the DE profile from the read coverage in the two conditions and then segment this profile into DE regions. The first option is made possible by several methods [11], [14], [19] that deal with transcript reconstruction. In this case, the DE units are derived from the estimated transcript structure and DE changes within those units remain invisible. Inferring DE regions directly from the DE profile provide a more detailed view of DE landscape but may raise problems concerning the correspondence to previous annotation. A work in progress belonging to this second category of approaches is presented in [10] where, from multiple replicates, coverage difference at position resolution are used to reconstruct DE regions. When replicates are not available alternative ways to control the incertitude should be considered.

Here we propose another method that also belongs to the second category of approaches and provides at base-pair resolution, both the regions whose expression changed above a given fold and an estimate of the change magnitude. Our approach builds upon a statistically sounded model, Parseq [14], for the analysis of the transcriptional landscape whose inference recovers the posterior distribution of expression levels for each genomic position. We complement this information by inferring regions which are statistically differentially expressed at a minimal fold change. After a description of the Parseq model, we will present our method, Pardiff which detects DE regions at a given fold change from RNA-Seq data. Then we will illustrate the relevance of such a strategy on semi-synthetic data-sets derived from an RNA-Seq experiment conducted on *S. cerevisae*.

2 Methods

2.1 Reconstruction of Transcriptional Profiles with Parseq

Given an RNA-Seq experiment, after alignment of the reads to the genome, we observe at each position t the counts y_t of reads starting at this position. We denote the transcription level by u_t (u_t and v_t in the case of two conditions). This level is by construction proportional to the expectation of y_t.

Our aim is to reconstruct the trajectory $\mathbf{u} = (u_t)_{t\geq 1}$, i.e. estimate expected values and credibility intervals along the genome and identify breakpoints from the sequence of read counts $\mathbf{y} = (y_t)_{t\geq 1}$. For this purpose we consider a State Space Model where u_t is a hidden variable taking values on the real half line

$[0; +\infty)$ whose distribution depends on u_{t-1} via a Markov transition kernel and y_t is an observation whose emission distribution depends on u_t. This framework allows accounting for the longitudinal dependency between the u_t's and provides great flexibility in the modeling of y_t given u_t. However, parameter inference and trajectory reconstruction is more challenging than in a classical HMM where only discrete values are considered for the hidden variable.

We recall shortly the characteristics of the Markov transition kernel and emission model underlying Parseq, more details are provided in [14].

Longitudinal model of transcriptional level Following the work of [16] on tiling array data, the Markov transition writes as a mixture of different change types, aiming at differentiating expressed ($u_t > 0$) and non expressed ($u_t = 0$) regions:

$$k(u_t; u_{t-1}) = \mathbf{1}_{\{u_{t-1}=0\}}\left[(1-\eta)\delta_0(u_t) + \eta f(u_t)\right] \\ +\mathbf{1}_{\{u_{t-1}>0\}}\left[\alpha\delta_{u_{t-1}}(u_t) + \beta f(u_t) + \beta_0\delta_0(u_t) + \gamma\, g(u_t; u_{t-1}, \lambda)\right],$$

where $\mathbf{1}$ denotes the indicator function indicating the expression status at $t-1$, and δ_x denotes the Dirac delta function with mass at point x that serves to give a non-zero probability for unchanged expression and for changes to 0 at t. The parameters $\eta \in (0,1)$ and $(\alpha, \beta, \beta_0, \gamma) \in (0,1)^4$ with $\alpha+\beta+\beta_0+\gamma = 1$ define the probabilities of the different types of moves. The terms $f(u_t; \zeta)$ and $g(u_t; u_{t-1}, \lambda)$ are probability densities for the transcription level u_t, at the beginning of a transcribed region (occurring with probability η when $u_{t-1} = 0$) or after a shift (probability β when $u_{t-1} > 0$), and after a drift (probability γ when $u_{t-1} > 0$) respectively. The density $f(u_t; \zeta)$ corresponds to an exponential distribution of rate ζ and the parameter λ defines the average relative change caused by drifts.

Read count emission model Due to the sampling nature of the RNA-Seq experiment, we model the distribution of the counts y_t as a negative binomial distribution with expectation that depends on u_t. Previous analysis have revealed protocol-specific effects that influence the scale of the observed counts. Thus, we integrate two types of effect: (1) a position scaling term ν_t related to the effect of k-mer composition [12] and (2) a short range correlation term s_t modeled by a second sequence of Markov-correlated hidden variables with mean 1 [14]. Integrating the variability sources and adding the possibility of outliers, our read count emission model is:

$$y_t \mid u_t, s_t \sim (1-\varepsilon_b-\varepsilon_o)\mathcal{NB}(\phi, u_t s_t \nu_t) + \varepsilon_b\, \mathcal{P}_{-\{0\}}(a\nu_t) + \varepsilon_0\, \mathcal{U}(0 \ldots b),$$

where the parameters $(\varepsilon_b, \varepsilon_0) \in (0,1)^2$ and $\varepsilon_b + \varepsilon_0 \leq 1$ correspond to the probability of two different types of outliers, $\mathcal{NB}(\phi, u_t s_t \nu_t)$ is the negative binomial distribution with mean $u_t s_t \xi_t$ and overdispersion ϕ, $\mathcal{P}_{-\{0\}}(a)$ is the zero-truncated version of the Poisson distribution with mean a and $\mathcal{U}(0 \ldots b)$ is the discrete uniform distribution over $(0 \ldots b)$. In practice a and b are respectively set to the mean and the maximum values of the observed read counts.

Characterizing $u|y$ For the expression reconstruction, we use a recent Sequential Monte Carlo method known as Particle Gibbs (PG) that makes possible to obtain exact joint samples of the hidden trajectory and parameters given the data [2].

The result of the PG algorithms consists in trajectories drawn from $\mathbf{u} \mid \mathbf{y}$ and give access, for each position to $u_t|y$ through a sample of N particles $(u_t^{(i)})_{i=1}^N$

Note on Normalization. When we consider more than one condition, the depth of sequencing (e.g. the total number of reads produced) will directly affect u_t, the inferred transcription level. Under a perfectly controlled experiment, u_t is expected to scale linearly with the depth, and thus people proposed to scale u_t accordingly. However, [4] highlighted that in most transcriptomes a small fraction of the genes makes up most of the molecular mass, and thus simple scaling could lead to very unstable normalization. Following classical strategy for microarray data, we can perform scalar or quantile normalization on the set of expression levels [3]. The analysis of semi-synthetic data-sets that served here to compare the perfomance of the methods did not necessitate a normalization step.

2.2 Statistics to Detect Differential Expression

We want to provide a statistically sounded way of estimating the fold change and calling regions exhibiting DE above a given fold change level c between two data-sets. We base our method on the separate estimation of the expression level on each sample, denoted $\mathbf{u} = (u)_{t\geq 1}$ and $\mathbf{v} = (v)_{t\geq 1}$, with Parseq algorithm. Our goal is to estimate the fold change at a base-pair resolution, that is to estimate the change between u_t and v_t where t is the position on the genome. For this reason, all our variables refer to the position resolution and we often omit to write the position index t.

The ratio distribution has already attracted attention in sample survey and many other areas. Multiple approximation were proposed, either from large sample or hypothesizing a Gaussian distribution of the variables. A more general approach was also proposed [9] to derive confidence intervals.

A direct point estimate is provided by the ratio of posterior means $\hat{r}_{\mathrm{RM}} = \frac{\bar{u}}{\bar{v}}$ where $\bar{u}$ and $\bar{v}$ are expectations of the posterior distributions of the expression levels u and v in the two conditions as sampled with Parseq MCMC algorithm. To incorporate the information on uncertainty embeded in the posterior distribution we also considered fold-change estimate based on the posterior distribution of the ratio $r = \frac{u}{v}$. A natural way of doing it is to consider the empirical distribution of the sample $(r)_{1\leq i\leq N^2} = (\frac{u^{i_u}}{v^{i_v}})_{1\leq i_u\leq N, 1\leq i_v\leq N}$ where N is the sample size drawn from each posterior using Parseq MCMC algorithm and i_u and i_v the corresponding sample indexes.

Here, we also analyzed the results obtained with another approximation of the posterior distribution of the ratio $r = \frac{u}{v}$ build on the hypothesis that the posteriors on expression levels u and v can be well approximated by a gamma distribution. Namely, $l \sim \gamma(\kappa_l, \theta_l), \quad l = \{u, v\}$, where the parameters κ_l and θ_l represent the shape and the scale parameters of the gamma distribution. In practice, examination of the posterior distributions suggest that this assumption is roughly justified for all of our experiments. Rescaling v by $\frac{\theta_u}{\theta_v}$ brings the

two gamma distributions to the same scale while keeping the shapes unchanged allowing an explicit form of the ratio distribution. The ratio $\tilde{r} = \frac{u}{v} \cdot \frac{\theta_v}{\theta_u}$ has a Beta prime distribution $\mathcal{B}'(\kappa_u, \kappa_v)$ with density

$$\pi(\tilde{r}) = \frac{\tilde{r}^{\kappa_u - 1} \cdot (1+\tilde{r})^{-(\kappa_v + \kappa_u)}}{\beta(\kappa_u, \kappa_v)} \tag{2.1}$$

where β refers to the beta function.

The parameters κ and θ can be estimated for each individual posterior using the method of the moments or by maximum likelihood. By default we use the moment estimates of κ and θ which gives $\hat{\kappa}_l = \bar{u}_l^2/\hat{\sigma}_l^2$ and $\hat{\theta}_l = \hat{\sigma}_l^2/\bar{u}_l$, where $\hat{\sigma}_l^2$ is the sample estimate for the variance.

Estimating Fold Change and Differential Expression. We approximate the fold change and the differential expression above a given threshold c by using these three estimation methods:

1. **RM** - the point estimate based on the ratio of posterior means $\hat{r}_{\mathrm{RM}}$
2. **DR-e** - the posterior distribution of the ratio r as approximated by its empirical distribution;
3. **DR-β'** posterior distribution of the ratio r as derived from the Beta prime approximation $\tilde{r} \sim \mathcal{B}'(\kappa_u, \kappa_v)$.

We derive the DE at a given fold change c from the cumulative probability above c (tail function). In each method we obtain the complementary cumulative probability as follows:

$$\textbf{RM: } \mathbf{1}_{\{\hat{r}_{\mathrm{RM}} \geq c\}}; \textbf{ DR-e: } \frac{1}{N^2} \sum_{i_{u,v}}^{1:N^2} \mathbf{1}_{\{\frac{u^{i_u}}{v^{i_v}} \geq c\}} \text{ and } \textbf{DR-}\beta'\textbf{: } \int_{c\frac{\theta_v}{\theta_u}}^{\infty} \beta'_{\kappa_u,\kappa_v}(r)\,\mathrm{d}r.$$

Also, we determine positions having a given fold change c. We define a precision level and for this level build a precision interval $[c_1, c_2]$ around the target fold value. We then identify positions with point estimators in this interval or with a cumulative probability $P(c_1 \leq r \leq c_2)$ greater than a probability threshold.

Annotation of Differentially Expressed Regions. Read coverage variability induces uncertainty in the estimation of the expression level, which in turn can lead to discontinuities in the annotation of DE regions. In order to cluster positions, we used a local score approach, defined by the classical recurrence relation:

$$s_t = \max\{s_{t-1} + \log z_t - m, 0\},$$

where, in the context of DE region detection, $z_t = \pi(r_t \geq c)$ and the score s_t is the signal in which we search enriched regions. The penalty m is set higher than the average of z_t, in practice at 0.5. We selected regions with positive score from the first positive value to the maximum local score in the region. The score was set back to a null value after the end position of each of these segments to avoid overlooking downstream high scoring segments.

3 Results

The difficulty raised by evaluating our strategy on real data motivated the use of semi-synthetic datasets. The relevance on real cases is shown in the paragraph on detection of DE positions.

We started from a RNA-Seq experiment which was published in a study on regulatory non-coding RNAs in *S. cerevisae* [8] and sequenced on a SOLiD platform (Short Read Archive identifier SRR121907). Currently available RNA-Seq simulators (simNGS, Flux simulator) do not account for coverage variability as we observe on real datasets. Thus, we decide to generate synthetic data by taking real transcript expression values and generating counts according to a dispersion estimated by Parseq on real datasets.

We simulated data for the first 6 chromosomes of *S.cerevisiae* using transcripts from the SGD annotation [5]. For the "wild" data set ($\mathbf{v}$) we set the expression level for each transcript to the value computed from real data. This value is obtained by averaging the counts of reads corresponding to each transcript. For the "mutant" data set ($\mathbf{u}$) we used the same expression levels but we over expressed randomly 15% of the transcripts (corresponding to 200 transcripts) with folds change values of 1/4, 2, 4 or 8. To augment resemblance to real data we integrated in both cases local coverage alterations (s) as estimated on the real data. Conditioning on the expression profile and local alterations we sampled read counts: $y_l \mid l, s_l, l \in \{u, v\}$ according to a Negative Binomial distribution. We used mean ($\mu = l \cdot s_l$) and over-dispersion (ϕ) parametrization. The parameter ϕ was set to 2.9, a value estimated by Parseq on the real dataset. We then ran Parseq to estimate the expression profile for both data sets and obtained 2 samples of expression trajectories u^i and v^i, i=1:N. For each condition we run 2200 Parseq sweeps with a thinning step of 10 and we discard the 200 sweeps burn-in. Results using Parseq estimates were systematically compared with the estimation based on a sliding 100 bp window average of the read counts (SW). In order to avoid border effect, the SW estimate was constrained to the regions covered by at least one read.

For comparison at bp level we considered those positions where Parseq estimated average levels and SW values are above a background value (here 0.01 reads / bp). Reconstruction of DE regions included all values and we set to the background value all expression values below it. Estimation of parameters for the fold change distribution is done as described in the methods. However in some cases, degeneracy of the particles can lead to underestimate of the variance. We bound the coefficient of variation c_v to the maximum between 1‰low quantile $c_{v\,1‰}$ and 0.001 and then recalculate the variance: $\hat{\sigma}_{lt} = l_t \cdot c_{v\,1‰}$.

Given a level of fold change c, the results are assessed from three different standpoints: detection of positions with c fold change, of positions with at least c fold change, and the detection of DE regions of level c or above.

Results are reported in terms of sensitivity and positive predictive values (PPV) i.e. the fraction of true positives identified $\frac{TP}{TP+FN}$ and the percentage of positives from total predictions $\frac{TP}{TP+FP}$. Of main relevance, the comparison of results obtained using (1) RM, (2) DR-e and (3) DR-β' will motivate the choice of having sample estimates of expression level.

Table 1. Detection of change magnitude at position resolution. Synthetic data results. Positions expressed in any dataset lower than 0.01 reads/bp were disregarded. We show sensitivity and positive predictive values. Three fold values (2, 4, 8) were evaluated with a precision of ± 25% ; the threshold for the cumulative probability was set to 0.3.

	DR-e		**DR-β'**		**RM**		**SW**	
Fold	Sens.	PPV	Sens.	PPV	Sens.	PPV	Sens.	PPV
2	0.93	0.08	0.75	0.13	0.51	0.18	0.34	0.09
4	0.86	0.44	0.73	0.55	0.46	0.61	0.32	0.31
8	0.89	0.52	0.70	0.63	0.45	0.70	0.35	0.51

Fold Change Estimation. We consider a ± 25% precision around the correct fold change. Considering the theoretical cumulative betaprime function, for this precision interval, we threshold the cumulative probability (% of ratio values falling in the precision interval) at 0.3. Increasing this threshold will provide very high PPV but with significant sensitivity loss while, in reverse, at lower thresholds sensitivity reaches 1 but with very low PPV. All results based on Parseq expression level estimations are significantly better in both sensitivity and PPV that those obtained using SW (table 1). While DR-e and DR-β' show high sensitivity values for a moderate PPV decrease comparing to RM, the DR-β' seems to mediate better the trade-off between these two indicators.

Differential Expression. Estimation was done at bp precision for thresholds ranging from 2-fold to 8-fold. RM method performs better than SW mainly in terms of positive predictions (figure 1). DR-e and DR-β' results depend on the probability threshold. High sensitivity values are obtained by lowering the cumulative probability threshold to 0.25 with the cost of having PPV values similar to the SW method. We observe that, lowering the sample size, the PPV loss comparing to RM diminishes in the sensitivity - PPV trade. It is important to notice is the similar behavior of the DR-e and DR-β' sustaining the choice of the gamma distributions in modeling the expression level.

Differential Expression on Real Data. We analyzed the DE at position resolution on data from the study on regulatory non-coding RNAs, Xrn1-sensitive unstable transcripts (XUTs), in *S. cerevisae* [8]. XUTs accumulate in the mutant condition and their loci thus correspond to DE regions. As in [8], we scaled the reconstructed mutant expression profile so that levels of tRNA and snoRNA is equal between the two data sets and we excluded already annotated regions [5] from the DE analysis. To minimize the detection of UTRs we also excluded an additional 100 bp on both sides of each annotated gene. For DE thresholds ranging from 2 to 8 we compute the sensitivity and PPV in detecting XUT positions as annotated in the study. Results are shown in figure 2. As for the synthetic data, DR-e and DR-β' methods allow a good control of sensitivity with slight PPV changes.

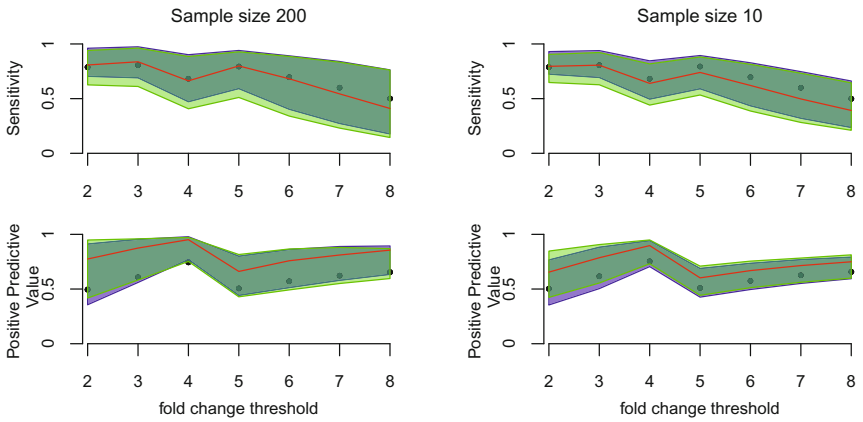

Fig. 1. Detection of DE at position level. X-axis: DE threshold. Y-axis: Sensitivity (top) and PPV (bottom). Sample size 200 (left) and 10 (right). Methods: DR-e (blue band), DR-β' (green band), RM (red line) and SW (black dots). Borders for DR-e and DR-β' bands: Sensitivity - top and low represent the 0.25 and 0.75 cumulative probability thresholds; PPV - top and low represent 0.75 and 0.25 thresholds.

Table 2. Accuracy in 5' End detection of DE regions. Results are shown for 3 values of DE thresholds. Estimated DE regions below 100 bp were discarded.

	DR-e		**DR-β'**		**RM**		**SW**	
Fold	Sens.	PPV	Sens.	PPV	Sens.	PPV	Sens.	PPV
≥ 2	0.72	0.20	0.72	0.25	0.69	0.25	0.66	0.12
≥ 4	0.57	0.43	0.58	0.39	0.54	0.39	0.63	0.26
≥ 8	0.43	0.39	0.48	0.38	0.38	0.35	0.44	0.23

Detection of DE Regions. To evaluate DE region detection we compared borders of regions estimated as having a fold change at least the DE threshold against those of transcripts with simulated fold change above the same threshold. The absence of noise and of longitudinal bias in the synthetic datasets allowed a high detection of transcript and DE regions borders. Sensitivity reaches values above 50% for all methods and most DE thresholds (table 2). Parseq based approaches have a net improvement in PPV with DR-e and DR-β' having slightly higher sensitivity results than RM.

4 Conclusion

This paper describes a method to reconstruct the regions having significant changes in expression between 2 conditions without recurring to predefined annotation and data sets replicates. This method is based on estimates of DE at position level and mitigates the lack of replicates by accounting for the incertitude in expression level estimation.

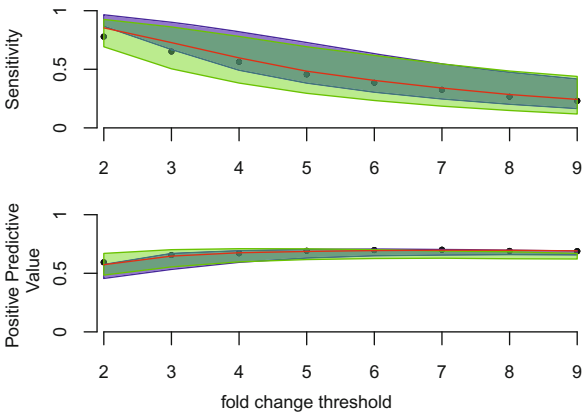

Fig. 2. Detection of XUTs at position level. Detection sensitivity and PPV computed on the whole repertoire of XUTs are shown as a function of the fold change threshold. Symbols are the same as in figure 1: DR-e (blue band), DR-β' (green band), RM (red line) and SW (black dots); the DR-e and DR-β' bands are delimited by the 0.25 and 0.75 cumulative probability thresholds.

Parseq, a probabilistic model for inferring expression profiles provides posterior estimates at position resolution which can be used to describe, directly or by approximation with a Betaprime distribution, the fold change distribution. Within this framework, we show improvements in the accuracy of predictions for methods based on empirical and betaprime approximation of ratio distribution, mainly for the estimation of the fold change.

This probabilistic model opens the way to more sophisticated approaches for the delineation of regions with constant a fold change, contributing to a better characterization of differences in expression.

References

1. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biology 11(10), R106 (2010), http://genomebiology.com/2010/11/10/R106
2. Andrieu, C., Doucet, A., Holenstein, R.: Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(3), 269–342 (2010)
3. Bolstad, B., Irizarry, R., Åstrand, M., Speed, T.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2), 185–193 (2003)
4. Bullard, J., Purdom, E., Hansen, K., Dudoit, S.: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11(1), 94 (2010), http://www.biomedcentral.com/1471-2105/11/94
5. Cherry, J.M., Hong, E.L., et al.: Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res 40(Database issue), D700–D705 (2012), http://dx.doi.org/10.1093/nar/gkr1029

6. Clark, T., Schweitzer, A., et al.: Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biology 8(4), R64 (2007)
7. DeRisi, J., Bittner, M.: Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. 14(4), 457–460 (1996)
8. van Dijk, E.L., Chen, C.L., et al.: Xuts are a class of xrn1-sensitive antisense regulatory non-coding rna in yeast. Nature 475(7354), 114–117 (2011), http://dx.doi.org/10.1038/nature10118
9. Fieller, E.C.: Some problems in interval estimation. Journal of the Royal Statistical Society. Series B (Methodological) 16(2), 175–185 (1954)
10. Frazee, A., Jaffe, A., Sabunciyan, S., Leek, J.: Differential expression analysis of rna-seq data at base-pair resolution in multiple biological replicates. Biostatistics (under revision)
11. Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., Regev, A.: Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotechnology 28(5), 503–510 (2010), http://www.nature.com/doifinder/10.1038/nbt.1633
12. Li, J., Jiang, H., Wong, W.H.: Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol. 11(5), R25 (2010)
13. McCarthy, D.J., Smyth, G.K.: Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 25(6), 765–771 (2009)
14. Mirauta, B., Nicolas, P., Richard, H.: Parseq: transcriptional landscape reconstruction from rna-seq data based on state-space models (submitted, 2013)
15. Mortazavi, A., Williams, B.A., et al.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5(7), 621–628 (2008), http://www.nature.com/doifinder/10.1038/nmeth.1226
16. Nicolas, P., Leduc, A., et al.: Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. Bioinformatics 25(18), 2341–2347 (2009)
17. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1), 139–140 (2010), http://bioinformatics.oxfordjournals.org/content/26/1/139.abstract
18. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L.: Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotech. 31(1), 46–53 (2013), http://dx.doi.org/10.1038/nbt.2450
19. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotech. 28(5), 511–515 (2010), http://dx.doi.org/10.1038/nbt.1621