Published online 17 September 2013

Nucleic Acids Research, 2013, Vol. 41, No. 21 ***e198***
doi:10.1093/nar/gkt834

DEXUS: identifying differential expression in RNA-Seq studies with unknown conditions

Günter Klambauer, Thomas Unterthiner and Sepp Hochreiter*

Institute of Bioinformatics, Johannes Kepler University, A-4040 Linz, Austria

Received June 17, 2013; Revised August 22, 2013; Accepted August 27, 2013

ABSTRACT

Detection of differential expression in RNA-Seq data is currently limited to studies in which two or more sample conditions are known a priori. However, these biological conditions are typically unknown in cohort, cross-sectional and nonrandomized controlled studies such as the HapMap, the ENCODE or the 1000 Genomes project. We present DEXUS for detecting differential expression in RNA-Seq data for which the sample conditions are unknown. DEXUS models read counts as a finite mixture of negative binomial distributions in which each mixture component corresponds to a condition. A transcript is considered differentially expressed if modeling of its read counts requires more than one condition. DEXUS decomposes read count variation into variation due to noise and variation due to differential expression. Evidence of differential expression is measured by the informative/noninformative (I/NI) value, which allows differentially expressed transcripts to be extracted at a desired specificity (significance level) or sensitivity (power). DEXUS performed excellently in identifying differentially expressed transcripts in data with unknown conditions. On 2400 simulated data sets, I/NI value thresholds of 0.025, 0.05 and 0.1 yielded average specificities of 92, 97 and 99% at sensitivities of 76, 61 and 38%, respectively. On real-world data sets, DEXUS was able to detect differentially expressed transcripts related to sex, species, tissue, structural variants or quantitative trait loci. The DEXUS R package is publicly available from Bioconductor and the scripts for all experiments are available at http://www.bioinf.jku.at/software/dexus/.

INTRODUCTION

The advent of next-generation sequencing has greatly expanded our knowledge about transcriptomes. New transcripts and splice variants have been found and break points of known transcripts determined more accurately (1–6). However, in RNA-Seq experiments, quantification of the expression of transcripts can be difficult (7). Without biological variability, transcripts that are differentially expressed between two conditions can be detected reliably (8). In studies with biological variability, however, detection of differential expression between two conditions remains challenging (9). A transcript that is differentially expressed between many conditions is hard to detect because read count variation due to differential expression and due to high overdispersion can only be distinguished with many samples and high coverage. See Supplementary Section S2 for more details. To detect differentially expressed transcripts, we therefore assume that the number of conditions is small compared with the number of samples.

Identifying differential expression is currently limited to particular study designs

Current methods for analyzing RNA-Seq data can identify differential expression between two conditions. For example, in a case-control study, only transcripts that are differentially expressed between cases and controls can be identified. Similarly, in a randomized controlled study, differential expression between treated and untreated subjects can be detected. These study designs can be generalized to more case groups or more treatments, which leads to multiple (more than two) known conditions. For example, multiple conditions may be due to different tissue types, as in the 'Allen Brain Atlas' (10), the 'Gene Expression Nervous System Atlas' (11), and the 'BioGPS' (12).

Identification of differential expression in RNA-Seq data requires a priori known conditions. In cohort, cross-sectional and nonrandomized controlled studies, the biological conditions are unknown or only partially known. Cohort and cross-sectional studies are observational studies in which the conditions of the subjects are unknown. Examples of observational studies include the HapMap (13), ENCODE (6) and the 1000 Genomes (14) project, for which RNA-Seq data are available (15,16). Nonrandomized controlled studies are treatment studies in which conditions such as genetic, environmental

*To whom correspondence should be addressed. Tel: +43 732 2468 4520; Fax: +43 732 2468 4539; Email: hochreit@bioinf.jku.at or treatment effects are not completely known. In nonrandomized controlled studies, unknown genetic variations such as single-nucleotide polymorphisms (SNPs), copy number variations and unknown environmental factors may result in differential expression between treated subjects. Furthermore, individual unknown treatment effects may cause variation in gene expression, for instance, responses of cell lines to the addition of compounds (17). Other examples are found in oncology, where unknown cancer subtypes or unknown cancer stages are characterized by a particular gene expression profile (18,19).

In nonclinical studies, the conditions are also often unknown. During development, the transcriptome regulates and controls cell growth, differentiation, movement and morphogenesis. Genes are differentially expressed between different time points and between different tissues; even within one tissue, gene expression may vary spatially. For two samples taken at different times or from different locations it is often unknown whether the conditions differ. Another example is *in vivo* or *in vitro* gene expression in mice treated with drug candidates (20,21). Unknown factors such as individual responses or side effects lead to differentially expressed transcripts between the samples.

The detection of differential expression in RNA-Seq studies with unknown conditions is important to obtain new biological knowledge. Current RNA-Seq methods, however, require the conditions to be known. For microarray data, a method for identifying unknown conditions in gene expression has been suggested (22). However, this method cannot be applied to RNA-Seq data with unknown conditions because a primary modeled factor is required and the noise is assumed to be Gaussian, which is not appropriate for RNA-Seq count data (23). We therefore present DEXUS, a method capable of detecting differential expression in RNA-Seq studies with unknown conditions.

A summary of study designs and methods that can detect differential expression in them is shown in Table 1.

Existing methods for detecting differential expression in RNA-Seq data

Methods that detect differential expression in RNA-Seq data are usually based on read counts, i.e. the number of reads mapping to a DNA region that is transcribed, such as a gene or an exon (32). These methods compare read counts for two conditions. If read counts show a large and consistent difference between the conditions, then the according transcript is differentially expressed. In this subsection, we review methods that detect differential expression in RNA-Seq data. Many methods model read counts by a negative binomial distribution because even after normalization the read counts have high variance. Therefore, we divide methods into two classes: those which do not use negative binomials (class A) and those which do (class B).

The following methods belong to class A.

DEGSeq (28) assumes that the log fold change of mean read counts between the two conditions follows a normal distribution given the log average expression. A differentially expressed gene is identified by a small *P*-value by means of this distribution.

NOISeq (30) also considers the log fold change of read counts between two given conditions together with their absolute difference. Empirical distributions are calculated using all pairs of replicates from different conditions. NOISeq identifies a gene as differentially expressed if the log fold change of read counts and the absolute difference of read counts between the two conditions have both a small *P*-value for the empirical distributions.

SAMSeq (27) performs a Wilcoxon test for each transcript testing the counts of one condition against the counts of the other. Because standard normalization techniques are not applicable, subsampling is used to normalize the read counts. SAMSeq requires a relatively high number of samples per condition to obtain significance for differential expression.

PoissonSeq (29) fits a Poisson log-linear model to the read counts after transforming them. A score statistic on the model parameters determines the significance for differential expression.

The following class B methods use negative binomial distributions to model the read counts.

edgeR (25) uses a quantile-adjusted conditional maximum likelihood estimator for the overdispersion parameter of the negative binomial distribution. This estimator is more accurate than the standard maximum likelihood estimator when only few replicates per condition are available (33). Borrowing information across transcripts allows the dispersion parameter to be adjusted toward a consensus value using an empirical Bayes procedure (34). Finally, edgeR uses an exact test

Table 1. An overview of study designs and methods that can detect differential expression in them

Study design	DEXUS	DESeq	edgeR	baySeq	SAMSeq	DEGSeq	PoissonSeq	NOISeq	DSS
Two known conditions									
Case-control study	✓	✓	✓	✓	✓	✓	✓	✓	✓
Randomized controlled study	✓	✓	✓	✓	✓	✓	✓	✓	✓
Multiple known conditions									
Multiple case-control study	✓	✓	✓	✓	✓				
Multiple treatment RCS	✓	✓	✓	✓	✓				
Unknown conditions									
Cross-sectional study	✓								
Cohort study	✓								
Nonrandomized controlled study	✓								

Alongside DEXUS, we included DESeq (24), edgeR (25), baySeq (26), SAMSeq (27), DEGSeq (28), PoissonSeq (29), NOISeq (30) and DSS (31).

to determine whether the counts of the two conditions come from the same negative binomial distribution.

DESeq (24) pools together transcripts with similar expressions values to improve the estimate of the overdispersion parameter. The overdispersion is assumed to be a function of the mean read count and is therefore estimated per condition. To determine whether a transcript is differentially expressed, the distribution parameters of the two conditions are tested by an exact test for equality of means.

baySeq (26) determines the distribution of the overdispersion parameter by applying a quasi-likelihood method to the read counts of one condition. The resulting distribution is used as prior for estimating the overdispersion parameter when fitting the model to the read count data.

DSS (31) is similar to baySeq. A negative binomial distribution is fitted to the read count data using a prior on the overdispersion parameter. This prior is a log-normal distribution, whose parameters are optimized using the dispersion parameters of each condition. Finally, a Wald test is used to determine differential expression.

In summary, the class B methods, which use negative binomial distributions, i.e. DESeq, baySeq, DSS and edgeR, mainly differ in the way they estimate the overdispersion parameter. Estimating the overdispersion parameter is crucial for the performance and not trivial because the maximum likelihood estimator is biased and has high variance if the sample size is small (33). The subsequent statistical test has a smaller effect on the results than the parameter estimates (23,31).

Extensions to multiple known conditions

McCarthy *et al.* (32) extended the R package edgeR to more than two conditions. A generalized linear model is fitted to the data, and then coefficients are tested for being different from zero, which leads to the final *P*-values. Again, the estimation of the overdispersion parameter for a transcript borrows information from other transcripts. DESeq, baySeq and SAMSeq have also been extended to more than two conditions.

MATERIALS AND METHODS

Method overview

Our goal is to identify differentially expressed transcripts in studies with unknown conditions. A transcript is differentially expressed if the mean expression levels for different conditions are different and read counts are observed under more than one condition. Therefore we assume a small number of conditions because, as mentioned above, the detection of differential expression for many conditions is difficult. RNA-Seq expression data are usually represented as read counts per transcript, or alternatively by exon or gene. It was observed that read counts from a single condition follow a negative binomial distribution (24–26,31). DEXUS therefore models read counts as a finite mixture of negative binomial distributions.

The model that best explains the observed read counts is selected from a set of models. In a Bayesian framework, model selection is based on finding the parameter which maximizes the posterior, the maximum a posteriori (MAP) parameter. The MAP model is found by an expectation maximization (EM) algorithm, where E-step and M-step are alternated repeatedly. The E-step estimates the unknown conditions based on actual model parameters, and the M-step optimizes the model parameters based on the E-step estimates. Models that use only one condition to explain the read counts are preferred by means of a prior distribution. One condition is the null hypothesis, which is rejected only if the data show strong evidence for more than one condition. Therefore, the parameters of the prior distribution determine how much DEXUS prefers to select models that explain the data without differential expression. Consequently, via the prior parameters, DEXUS can be adjusted to have a low false discovery rate at the detection of differential expression.

In the following subsections, we first describe the model in more detail and then explain the EM algorithm for model selection. Model selection includes prior assumptions that lower the false discovery rate and lead to more accurate estimates. Finally, we show how to call differentially expressed transcripts on the basis of an informative/noninformative (I/NI) value.

The model

Read count x per transcript is explained by a mixture of n negative binomial distributions:

$$p(x) \; = \; \sum_{i=1}^{n} \alpha_i \; \mathrm{NB}(x\,;\,\mu_i,r_i) \qquad \mathbf{(1)}$$

where α_i is the probability of being in condition i out of n possible conditions. In condition i, read counts are drawn from a negative binomial distribution with mean μ_i and size r_i, where the size parameter r_i is the inverse of the overdispersion ϕ_i. Note that we use the (μ,r) instead of the usual (μ,ϕ) parametrization to locally accumulate parameters that are associated with large overdispersions. This accumulation is essential to define a prior within a Bayesian framework.

A nondegenerate DEXUS model is identifiable (see Supplementary Section S3.1.3), as required for the maximum likelihood and the maximum a posterior estimator to be consistent. Consistency means that the estimator converges to the true parameter values with more data points, which is important for identifying differential expression. If the mean read count exceeds the variance, the maximum likelihood estimate of r tends to ∞ and the negative binomial converges to a Poisson distribution (see Supplementary Section S3.2.2).

Model selection

We perform model selection in a Bayesian framework by maximizing the posterior, i.e. by a MAP approach (35–37). Therefore, the parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{r} = (r_1, \ldots, r_n)$ are considered to be random variables, and the likelihood $p(x)$ in Equation (1) becomes the conditional probability

$p(x|\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{r})$. The objective of the model selection is to maximize the posterior of the parameters:

$$p(\boldsymbol{\mu},\boldsymbol{r},\boldsymbol{\alpha}|x) = \frac{p(x|\boldsymbol{\mu},\boldsymbol{r},\boldsymbol{\alpha})\ p(\boldsymbol{\mu})\ p(\boldsymbol{\alpha})\ p(\boldsymbol{r})}{\int p(x|\boldsymbol{\mu},\boldsymbol{r},\boldsymbol{\alpha})\ p(\boldsymbol{\mu})\ p(\boldsymbol{\alpha})\ p(\boldsymbol{r})\ d\boldsymbol{\alpha}\ d\boldsymbol{r}\ d\boldsymbol{\mu}} = \frac{1}{c(x)}\ p(x|\boldsymbol{\mu},\boldsymbol{r},\boldsymbol{\alpha})\ p(\boldsymbol{\mu})\ p(\boldsymbol{\alpha})\ p(\boldsymbol{r}) \tag{2}$$

where the priors on $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$ and $\boldsymbol{r}$ are assumed to be independent of each other, and are defined in the following.

Dirichlet prior for probabilities of conditions

First we choose the prior $p(\boldsymbol{\alpha})$ on the probabilities of the conditions. Since the majority of transcripts in a data set are usually not differentially expressed, the model should favor explaining the read counts for a transcript with a single condition. The null hypothesis of one condition should only be rejected if the data contain strong evidence for more than one condition. The prior reduces the number of falsely discovered differentially expressed transcripts and therefore keeps the false discovery rate low. DEXUS uses a Dirichlet prior $p(\boldsymbol{\alpha})$ on $\boldsymbol{\alpha}$ with parameters γ to incorporate the preference for only one condition:

$$p(\boldsymbol{\alpha}) = b(\gamma) \prod_{i=1}^{n} \alpha_i^{\gamma_i - 1} \tag{3}$$

where $\boldsymbol{\alpha}$ is an n-dimensional probability vector. Each component α_i is distributed according to a beta distribution with $\text{mode}(\alpha_i) = (\gamma_i - 1)/\left(\sum_{i=1}^{n} \gamma_i - n\right)$.

To express the prior knowledge that most transcripts are not differentially expressed and are generated under only one condition, we set $\gamma_1 \gg \gamma_i$ (for $i > 1$). This setting assumes that most read counts are generated under condition $i = 1$, which we call the *major condition*, while conditions $i > 1$ are called *minor conditions*. The vector of hyperparameters $(\gamma_1,\gamma_2,\ldots,\gamma_n)$ can be simplified to one hyperparameter G (Supplementary Section S3.2.1). In Supplementary Section S3.4 we show that DEXUS is not sensitive to the choice of the hyperparameter G. Therefore DEXUS is easy to use as good results are obtained with the default setting of $G = 1$ (see Supplementary Section S3.4). Without having seen the data, we assume that only the major condition is present, which means that the transcript is not differentially expressed. Only when the data show strong evidence also for minor conditions, does the posterior assign nonzero probabilities to minor conditions and the transcript is called differentially expressed.

Truncated exponential priors for overdispersions

In DEXUS model selection, the second prior is on the size parameter r of the negative binomial distribution, which determines the overdispersion. A prior on r improves the estimation of r if the number of samples is small. The maximum likelihood estimator of r is biased for few samples and overestimates the true size parameter (38,39), as confirmed in Supplementary Section S3.2.5. In a Bayesian approach, the influence of the prior decreases with an increasing number of samples, and therefore the MAP estimator is asymptotically (number of samples tending to infinity) unbiased.

To keep the estimate of r small, the prior pushes r toward zero. We choose an exponential distribution as prior:

$$p(r) = \eta\, e^{-\eta r}, \tag{4}$$

where η is a hyperparameter.

Like DESeq (24), we truncate the size parameter at the right-hand tail by using the constraint $r \leq r_{\max}$. The upper bound $r_{\max}$ on the size parameter is equivalent to a lower bound on the overdispersion and ensures minimal overdispersion for the read counts of each transcript. Further, this bound makes the parameter space compact, which is required for a consistent estimator. The same exponential prior is used for each component of $\boldsymbol{r}$. The hyperparameter η for the exponential prior on r is transformed to a hyperparameter θ (see Supplementary Section S3.2.5). Like the hyperparameter G, also θ is robust and supplies good results with $\theta = 2.5$.

Uniform priors for means

Finally, DEXUS model selection uses a prior on the mean μ of the negative binomial distribution. If in one condition all read counts were close to zero (transcripts are not present), the estimate of the mean of the negative binomial would not converge. Therefore, μ_i is lower bounded by $\mu_{\min}$. To ensure a compact parameter space as required for a consistent estimator, we use a uniform prior on μ_i on the compact interval $[\mu_{\min},\mu_{\max}]$, where $\mu_{\max}$ can be set to the largest observed read count.

In summary, DEXUS has only few parameters which in most applications need not be adjusted by the user, as their default values give good results.

EM algorithm

With the priors defined, the model with maximum parameter posterior can be selected. The EM algorithm (40) is used to minimize an upper bound on the negative log-posterior of the parameters. The E-step of the EM algorithm estimates the probability that a read count is generated under a particular condition. The M-step optimizes the model parameters based on the E-step estimates.

In the DEXUS model, α_i is the probability of condition i without observing any data. The model posterior $\tilde{\alpha}_{ik}$ estimates the probability that read count x_k is generated under condition i (the probability of condition i after observing data x_k):

$$\tilde{\alpha}_{ik} = \frac{\alpha_i\ \text{NB}(x_k\,;\,\mu_i,r_i)}{\sum_{i=1}^{n} \alpha_i\ \text{NB}(x_k\,;\,\mu_i,r_i)}\ . \tag{5}$$

This equation is the E-step (expectation step) of the EM algorithm. Using the posterior estimates $\tilde{\alpha}_{ik}$, we obtain following update rules for the M-step (maximization step):

- estimate for α_i:

$$\hat{\alpha}_i = \frac{1}{N} \sum_{k=1}^{N} \tilde{\alpha}_{ik} \tag{6}$$

- $\boldsymbol{\mu}$ update:

$$\mu_i = \frac{\frac{1}{N}\sum_{k=1}^{N} \tilde{\alpha}_{ik}\, x_k}{\hat{\alpha}_i}. \qquad (7)$$

- $\boldsymbol{r}$ update:
  The new r_i is obtained by solving the following equation for r_i:

$$\sum_{k=1}^{N} \tilde{\alpha}_{ik}\, \psi(x_k+r_i) \;-\; N\,\hat{\alpha}_i\, \psi(r_i) \;+ \qquad (8)$$

$$+\, N\,\hat{\alpha}_i\, \log\left(\frac{r_i\,\hat{\alpha}_i}{\frac{1}{N}\sum_{k=1}^{N} \tilde{\alpha}_{ik}\, x_k + r_i\,\hat{\alpha}_i}\right) \;-\; \eta \;=\; 0,$$

  where ψ is the digamma function. The equation is solved numerically for r_i by means of a 'bisection' procedure.

- $\boldsymbol{\alpha}$ update:

$$\alpha_i = \frac{\hat{\alpha}_i + \frac{1}{N}\,(\gamma_i - 1)}{1 + \frac{1}{N}\left(\sum_{i=1}^{n} \gamma_i \;-\; n\right)}. \qquad (9)$$

The complete derivation of the EM algorithm can be found in the Supplementary Section S3.2.1.

μ_i and r_i are initialized by using the results of k-means clustering (see Supplementary Section S3.2.4). The values α_i are simply initialized with the maximum entropy setting $\alpha_i = 1/n$.

I/NI value: evidence of differential expression

The Bayesian framework allows definition of an I/NI call (36,37,41,42). The I/NI call serves to extract differentially expressed transcripts with a desired specificity (1 − significance level or 1 − type I error rate) or sensitivity (power or 1 − type II error rate). DEXUS first computes the I/NI value, which quantifies the contribution of differential expression to the read count variation. Transcripts are then called informative if the I/NI value exceeds a threshold.

Unlike ϕ_i or r_i, which capture noise variation, $\boldsymbol{\alpha}$ captures variation arising from differentially expressed transcripts. The posterior $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ indicates differential expression in the data in the form of minor conditions with probabilities larger than zero. The larger the posterior value $\hat{\alpha}_i$ of a minor condition $i > 1$, the stronger the evidence for its presence. Further, evidence is also required that the minor condition is different from the major condition in terms of mean read counts. Although identifiability of the DEXUS model ensures that the negative binomials of different conditions are different, they may still be close to one another. The more the mean μ_i of the minor condition $i > 1$ differs from the mean of the major condition, the stronger is the evidence that the minor condition is different from the major condition. In conclusion, evaluating the evidence of differential expression (the I/NI value) should consider two factors: (i) $\hat{\alpha}_i$ as the evidence for the presence of the minor condition $i > 1$; (ii) the difference between the means of the major and minor conditions as evidence that they are indeed different.

The difference between the means is expressed by the log difference $\left|\log(\mu_i) - \log(\mu_1)\right|$. Factor (I) is incorporated into the I/NI value by weighting these differences by $\hat{\alpha}_i$, which yields

$$\begin{aligned}\mathrm{I/NI}(\widehat{\boldsymbol{\alpha}},\boldsymbol{\mu}) &= \sum_{i=1}^{n} \hat{\alpha}_i \left|\log\left(\frac{\mu_i}{\mu_1}\right)\right| \\ &= \sum_{i=1}^{n} \hat{\alpha}_i \left|\log(\mu_i) - \log(\mu_1)\right| .\end{aligned} \qquad (10)$$

The I/NI value is the expected log fold change of read counts with respect to the mean read count of the major condition given a noise-free model. 'Noise-free' refers to the assumption that under each condition, only the mean read count is observed. For a mathematical interpretation of the I/NI value see Supplementary Section S3.3.2.

Experiments

We evaluated DEXUS on simulated and real-world data sets. The simulated data sets had various library sizes, different numbers of replicates and different ratios between mean read counts under the different conditions. DEXUS was tested on the following real-world RNA-Seq data sets: (i) 'Nigerian HapMap', based on 69 Nigerian HapMap individuals, (ii) 'European HapMap', based on 60 European HapMap individuals, (iii) 'Primate Liver', based on liver tissue samples from humans, chimpanzees and rhesus macaques, (iv) 'Maize Leaves', using samples from different locations of maize plant leaves, and (v) 'Mice Strains', based on different strains of mice (Supplementary Section S4.2.4).

First we report the performance of DEXUS on 2400 simulated data sets for which the conditions were known but withheld from DEXUS. We then present tests on real-world data sets with either unknown conditions ('Nigerian HapMap', 'European HapMap') or partially known conditions ('Primate Liver', 'Maize Leaves'). In the latter case the conditions were withheld from DEXUS to ascertain whether it can identify them.

DEXUS for known conditions

Before we tested DEXUS on data with unknown conditions, we assessed how well it performs if the conditions of interest are known. For known conditions, DEXUS estimates only the parameters of a negative binomial for each condition. Therefore, we compared the parameter estimates of DEXUS to previously suggested methods in terms of detecting differentially expressed transcripts, namely the following eight state-of-the-art methods: DSS (31), DESeq (24), baySeq (26), edgeR (25), DEGseq (28), NOISeq (30), PoissonSeq (29) and SAMseq (27).

If only few samples per condition are available, the performance of DEXUS is below the best performing other methods. For medium and large sample numbers and small library size (1e6) DEXUS is second and third best method. For medium and large sample numbers and large library sizes (1e7 and 1e8) DEXUS outperforms all other methods. The experiments and the respective results are described in detail in Supplementary Section S4.2.

Simulated RNA-Seq data

Generating simulated RNA-Seq data

We simulated data sets from different experimental settings following the suggestions of Robinson *et al.* (34), Hardcaste and Kelly (26) and Wu *et al.* (31). For all samples of a data set, the library size was 1e6, 1e7 or 1e8 to cover a wide range of applications. Keeping the library size and the read quality constant for each sample in a data set avoids the need for normalization of the read counts, i.e. it avoids normalization biases. For each experiment, we choose a particular number of replicates per condition to evaluate DEXUS for different sample sizes and for unbalanced data. In case of two conditions, the numbers of replicates were 6/6, 9/3, 10/2, 11/1, 12/12, 18/6, 20/4, 22/2 (condition1/condition2). Each experiment consisted of 100 data sets with 10 000 transcripts each. The conditions were known and used for evaluation but withheld from DEXUS.

For the simulation we assumed that under condition i the reads x for a transcript are distributed according to a negative binomial $\mathrm{NB}(x\,;\,\mu_i,r_i)$. After Wu *et al.* (31), we took the mean μ_i and the size r_i from the 'Mice Strains' benchmark RNA-Seq data set (43) using only data from one particular biological condition. For a randomly selected transcript, the value μ_i was obtained as the median read count of the condition.

The overdispersion $\phi_i = 1/r_i$ tends to decrease with increasing mean read counts (see Supplementary Figure S15). Therefore, we fitted a regression line to overdispersion values by least squares. After sampling μ_i values, the corresponding ϕ_i values were obtained by means of the regression line to which zero-one normally distributed noise was added. Thirty percent of the genes were chosen to be differentially expressed. Differential expression was modeled by adjusting the means of the negative binomials to obtain log fold changes of 0.5, 1 and 1.5 between the mean of the major and the minor condition. The fold change values are randomly chosen with equal probability, such that all 3-fold change categories have about the same number of genes in each data set.

Evaluation criteria for simulated RNA-Seq data

We formulate the detection of differential expression as a classification task: DEXUS must decide whether a transcript is differentially expressed (positive prediction) or not (negative prediction). For the simulated data, we knew which transcripts were differentially expressed (the positives) and which were not (the negatives). DEXUS ranks the transcripts by the I/NI value from Equation (10). For a given I/NI threshold (the I/NI call), we can determine true positives, false positives, true negatives and false negatives. Using these numbers, we computed the specificity and the sensitivity of DEXUS. The specificity corresponds to '1 − significance level' or '1 − type I error rate'. The type I error rate is the ratio between false detections and all negatives. The sensitivity corresponds to the 'power' or '1 − type II error rate'. The type II error rate is the ratio between missed positives and all positives.

Results on simulated RNA-Seq data

We tested DEXUS on the simulated RNA-Seq data using its default parameters. Table 2 shows the results in terms of sensitivity and specificity for library size 1e8 at different thresholds for the I/NI value. Transcripts with an I/NI value above the threshold are called informative or (equivalently) differentially expressed. Results for other library sizes are presented in Supplementary Tables S12 and S13. The specificity of DEXUS is high across various numbers of replicates, whereas the sensitivity varies considerably. High specificity means that few transcripts are falsely identified as being differentially expressed. In highly unbalanced experiments, i.e. 11/1 and 22/2 replicates, differentially expressed transcripts are detected only at low I/NI thresholds of 0.025 and 0.05. Note that the minor condition $i = 2$ (smaller subgroup) leads to a small α_2 and therefore to a small I/NI value. For unbalanced data, the few minor condition samples must be distinguished from random outliers of the major condition.

Real-world RNA-Seq data

'Nigerian HapMap'

Pickrell *et al.* (16) sequenced RNA from 69 Nigerian HapMap individuals to study expression quantitative trait loci (eQTLs). The read count data were provided by the ReCount repository (44). As in previous experiments, DEXUS was applied to these data with its default parameters and ranked genes according to the I/NI value. The read counts of top-ranked genes and the conditions identified by DEXUS are visualized as a heatmap in Figure 1.

Five out of the 12 top-ranked genes are located on the Y chromosome (RPS4Y1, CYorf15A, EIF1AY, TMSB4Y, RPS4Y2). For these genes, the identified conditions are related to the sex. For four of the 12 top-ranked genes, at least one eQTL is known. For ZFP57, the associated eQTL is the SNP rs1736924 with a minor allele frequency (MAF) of 0.14 (16). CDH1 has 6 eQTLs, one of which is SNP rs7196495 with a MAF of 0.22 (45). CLLU1OS possesses the eQTL SNP rs12580153 with a MAF of 0.19 (46). L1TD1 has 2 eQTLs, one of which is SNP rs12137088 with a MAF 0.30 (47). Because the MAFs are high, it is plausible that the minor alleles are observed in the HapMap data set and that they lead to differential expressions of the associated genes. The conditions that were found by DEXUS are related to the alleles of corresponding SNPs.

Because the HapMap samples are lymphoblastoid cells, we confirmed that the genes detected by DEXUS are indeed expressed in lymphoblastoid cell lines. The gene NLRP2, ranked 11th by DEXUS, is expressed in lymphoblastoid cells but with large variability (48), as shown in Supplementary Figure S17. NLRP2 is expressed in the HapMap individuals, but in some at a low level.

Table 2. The performance of DEXUS in terms of sensitivity and specificity in detecting differential expression with unknown conditions

I/NI threshold	0.025		0.05		0.1	
C1/C2	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
6/6	0.893 ± 0.003	0.775 ± 0.009	0.951 ± 0.002	0.720 ± 0.009	0.985 ± 0.002	0.646 ± 0.009
9/3	0.893 ± 0.004	0.827 ± 0.006	0.951 ± 0.002	0.766 ± 0.007	0.985 ± 0.001	0.580 ± 0.008
10/2	0.893 ± 0.003	0.819 ± 0.008	0.950 ± 0.002	0.656 ± 0.009	0.985 ± 0.001	0.325 ± 0.009
11/1	0.893 ± 0.003	0.677 ± 0.009	0.951 ± 0.002	0.351 ± 0.008	0.985 ± 0.001	0.020 ± 0.003
12/12	0.945 ± 0.002	0.735 ± 0.008	0.982 ± 0.001	0.665 ± 0.008	0.996 ± 0.001	0.610 ± 0.009
18/6	0.945 ± 0.003	0.816 ± 0.008	0.982 ± 0.002	0.743 ± 0.009	0.996 ± 0.001	0.570 ± 0.011
20/4	0.945 ± 0.003	0.810 ± 0.008	0.982 ± 0.002	0.625 ± 0.009	0.996 ± 0.001	0.308 ± 0.009
22/2	0.946 ± 0.002	0.650 ± 0.009	0.982 ± 0.001	0.325 ± 0.008	0.996 ± 0.001	0.006 ± 0.002
Mean	**0.919 ± 0.028**	**0.764 ± 0.069**	**0.966 ± 0.017**	**0.606 ± 0.172**	**0.991 ± 0.006**	**0.383 ± 0.261**

The first column ‘C1/C2’ contains the numbers of replicates for the first and second condition. The other columns list sensitivity and specificity (with standard deviations) of DEXUS at different I/NI thresholds as the average across 100 data sets. The last row (‘Mean’) gives the average of the results in the columns. The library size was 1e8 for all experiments.

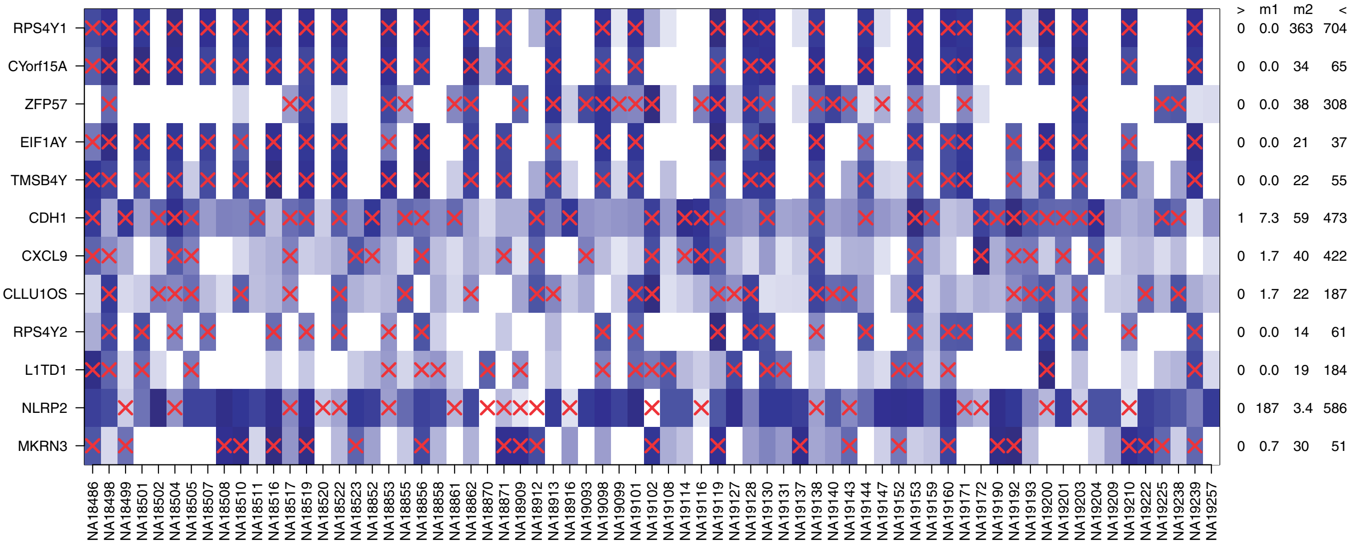

Figure 1. Heatmap of the normalized read counts of the 12 genes with the largest I/NI values for the ‘Nigerian HapMap’ data set. Colors range from white for low expression to blue for high expression. Different individuals are denoted along the x-axis, while the top-ranked genes are denoted by their gene symbols along the y-axis. Red crosses indicate samples that belong to the minor condition. At the right side of the heatmap, each gene is annotated by the minimum (‘>’), the median of two conditions (‘m1’ and ‘m2’) and the maximum (‘<’) read count.

Schlattl *et al.* (49) identified a copy number variable region that partially covers NLRP2 and causes its differential expression. Therefore, the conditions that DEXUS identified for NLRP2 may be related to copy number states of the samples. Copy number states may also be responsible for differential expression of MKRN3, which was ranked 12th by DEXUS. Pinto *et al.* (50) and Redon *et al.* (51) identified a copy number variable region covering MKRN3. However, interpreting the MKRN3 conditions is difficult because only the paternal copy of MKRN3 is expressed.

We analyzed the I/NI value ranking of transcripts: genes on the X chromosome were ranked significantly higher than other genes (P = 3.0e−12), which can be explained by sex-related transcripts. An analogous test for the Y chromosome was not significant because too few genes were expressed. However, as already mentioned, five out of the 12 top-ranked genes are located on the Y chromosome. At an I/NI threshold of 0.1, DEXUS called 366 differentially expressed genes. Gene set enrichment analysis showed that the called genes are associated with the extracellular region and the plasma membrane. In total, 20 significant GO terms were found, including ‘extracellular space’, ‘extracellular region part’ and ‘plasma membrane part’ with P = 2.5e−5, P = 8.8e−5 and P = 0.01, respectively. ‘Cell–cell signaling’, ‘chemokine receptor binding’ and ‘chemokine activity’ were also significant at P = 4.0e−3, P = 8.0e−4 and P = 9.8e−4 (P-values were corrected for multiple testing by means of the Benjamini–Hochberg procedure). These GO terms are in agreement with characteristics of lymphoblastoid cells. Supplementary Table S18 provides a complete list of all significant GO terms.

‘European HapMap’

We analyzed the RNA-Seq data of 60 individuals from the HapMap cohort from Montogmery *et al.* (15), which were

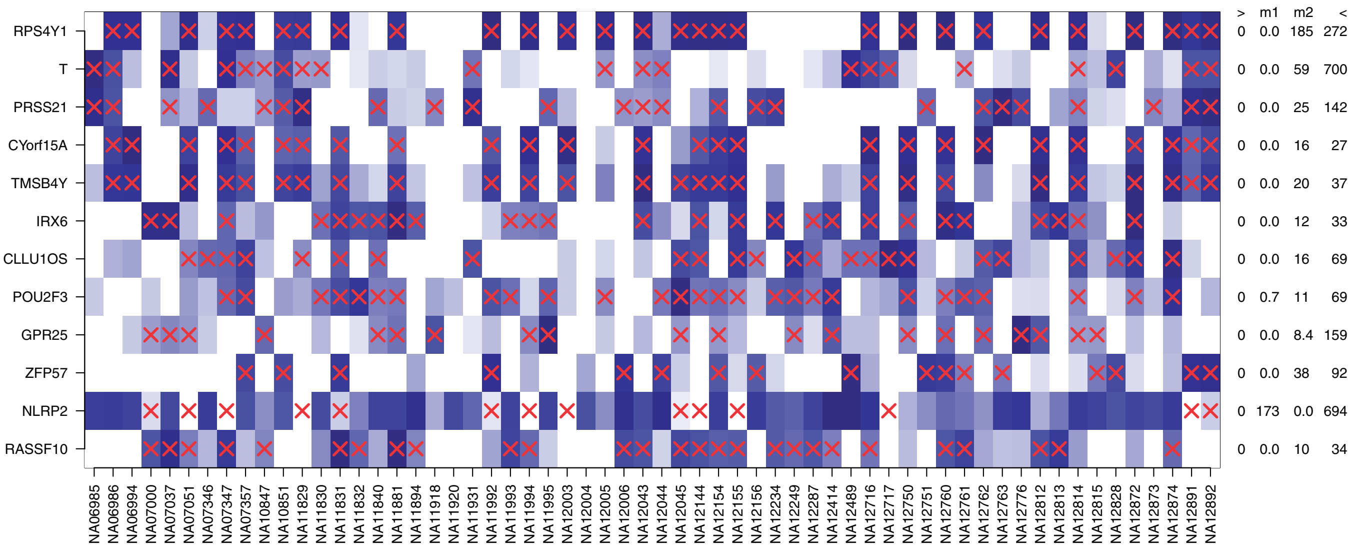

Figure 2. Heatmap of the normalized read counts of the 12 genes with the largest I/NI values for the 'European HapMap' data set. Colors range from white for low expression to blue for high expression. Different individuals are denoted along the x-axis, while the top-ranked genes are denoted by their gene symbols along the y-axis. Red crosses indicate samples that belong to the minor condition. At the right hand side of the heatmap, each gene is annotated by the minimum ('>'), the median of two conditions ('m1' and 'm2') and the maximum ('<') read count.

provided by the ReCount repository (44). Again, DEXUS was applied to these data with its default parameters and ranked genes according to the I/NI value. The read counts of top-ranked genes and the identified conditions are visualized as a heatmap in Figure 2.

RPS4Y1 is the gene with the largest I/NI value, differentially expressed between males and females, and located on the Y chromosome. The genes CYorf15A and TMSB4Y, ranked fourth and fifth according to the I/NI value, are also located on the Y chromosome. As in the 'Nigerian HapMap' data set, ZFP57 was detected as being differentially expressed. In addition to ZFP57, two other of the 12 top-ranked genes have eQTLs. CLLU1OS has as eQTL the SNP rs12580153 with a MAF of 0.19 (46). POU2F3 has as eQTL the SNP rs2847497 with a MAF of 0.14 (52). As in the 'Nigerian HapMap' data set, some top-ranked genes, such as NLRP2 (again rank 11), were differentially expressed owing to variable copy numbers (49). For the genes T, PRSS21 and RASSF10, DEXUS identified two conditions for which an explanation remains to be found and which may indicate a hitherto unknown source of variability in gene expression. The second-ranked gene T, the third-ranked gene PRSS21 and the 12th-ranked gene RASSF10 are known to be expressed in B-lymphoblastoid cells (6,12), the cell type of the HapMap samples. The high expression variability of T and PRSS21 in the B-lymphoblastoid cell line has already been reported by the ENCODE Project (6). The ENCODE Project expression values for the genes T, PRSS21 and RASSF10 are visualized in Supplementary Figures S19, S20 and S21.

Analyzing the I/NI value ranking, we found that genes on the X chromosome are ranked significantly higher ($P = 8.0\mathrm{e}{-6}$, Wilcoxon test). The analogous test for the Y chromosome yielded no significant results, as too few genes were expressed. However, three out of the 12 top-ranked genes with the largest I/NI value are located on the Y chromosome.

At an I/NI threshold of 0.1, DEXUS called 680 differentially expressed genes. Gene set enrichment analysis showed that the called genes are associated with ion transport. Significant Gene Ontology (GO) terms were 'ion transport', 'potassium ion transport' with $P = 0.04$ and $P = 4.3\mathrm{e}{-03}$, respectively. Again 'plasma membrane part' was significant at $P = 0.027$. Although 36 of the 680 genes were related to 'cell–cell signaling' and 6 to 'chemokine activity', these GO terms were not significant in this data set after correction for multiple testing by means of the Benjamini–Hochberg procedure. A table of all significant GO terms can be found in Supplementary Table S19.

'Primate Liver'

Blekhman *et al.* (53) investigated the differences in alternative splicing in liver tissue between humans, chimpanzees and rhesus macaques. For this purpose, they performed RNA-Seq on three male and three female liver samples from each species. They focused on the expression values of exons that had reliably determined orthologs in all species. Read counts for exons were provided by Blekhman *et al.* (53), who used gene models from Ensemble (Release 50). After pooling technical replicates, DEXUS ranked genes according to the I/NI value using its default parameters. The 10 top-ranked genes are visualized in Figure 3, which shows clear differential expression between the species. For these genes, and without having been provided with this information, DEXUS determined one of the three species as minor condition. Interestingly, out of the 10 top-ranked genes, six are human pseudogenes, AC010591.10, AC105383.3, AC093874.3-1, AC105383.3, AL132855.4 and UOX, which are inactive in humans because of recent structural

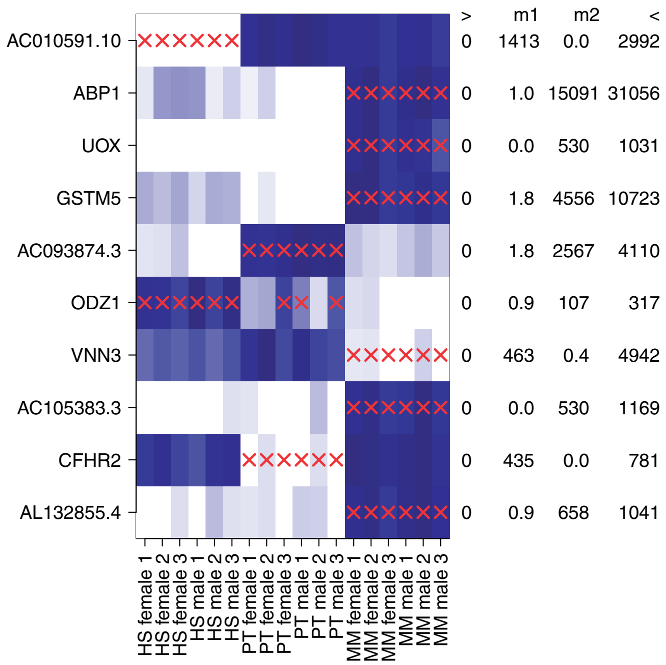

Figure 3. Heatmap of the normalized read counts of the 10 genes with the largest I/NI values for the 'Primate Liver' data set. Colors range from white for low expression to blue for high expression. The x-axis shows female and male individuals from the three species human *Homo sapiens* (HS), chimpanzee *P. troglodytes* (PT) and rhesus macaques *M. mulatta* (MM). The y-axis displays top-ranked genes indicated by their gene symbols. Red crosses mark samples that were assigned to the minor condition. At the right side of the heatmap, each gene is annotated by the minimum ('>'), the median of two conditions ('m1' and 'm2') and the maximum ('<') read count.

rearrangements (54). Because the rearrangements are recent, their orthologs can be identified reliably in other primates. Differential expression is detected because these orthologs are still transcribed in *Pan troglodytes* and in *Macaca mulatta*.

Several of the 10 top-ranked genes are associated with liver pathways and are therefore expressed in liver tissues. Differential expression of these genes between species may have arisen from different diets. Examples of such genes are the human pseudogene UOX, which catalyze the oxidation of uric acid to allantoin in *M. mulatta*, ABP1 and GSTM5, which participate in degradation and detoxification pathways, VNN3, which helps to recycle vitamin B5, and CHFR2, which is associated with lipoproteins.

Thresholding the I/NI call at 0.1, DEXUS called 3384 genes (16% of all genes) as differentially expressed. A gene set enrichment analysis found the GO terms 'intrinsic to plasma membrane' ($P = 7.9e{-7}$) and 'integral to plasma membrane' ($P = 4.0e{-6}$) to be significant. Thus, genes that encode membrane proteins seem to be differentially expressed between species more often than other genes. Interestingly, also 'response to extracellular stimulus', 'response to nutrient' and 'response to nutrient levels' were significant (all P-values $<7.6e{-5}$), which supports the hypothesis that some genes are differentially expressed owing to the different diets of the species. All P-values were corrected by means of the Benjamini–Hochberg procedure.

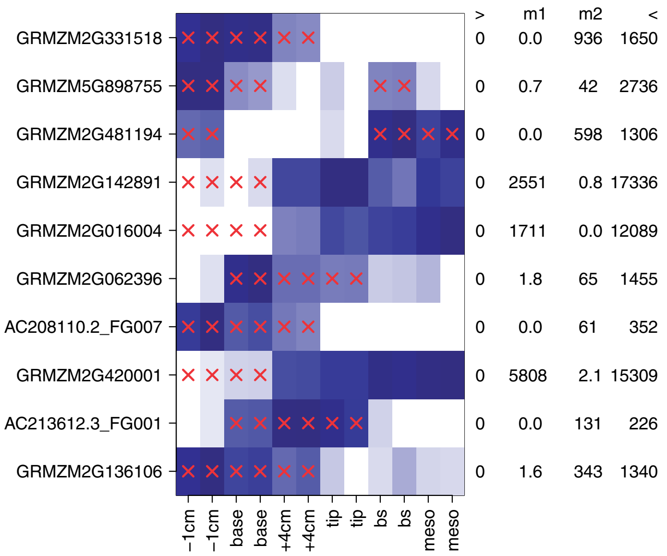

Figure 4. Heatmap of the normalized read counts of the 10 genes with the largest DEXUS I/NI values for the 'Maize Leaves' data set. Colors range from white for low expression to blue for high expression. The x-axis shows samples from different locations on the maize plant leaf. The y-axis displays different genes denoted by their gene symbols. Red crosses indicate that the according samples belong to the minor condition. At the right hand side of the heatmap, each gene is annotated by the minimum ('>'), the median of two conditions ('m1' and 'm2') and the maximum ('<') read count.

'Maize Leaves'

Using RNA-Seq data from various locations on maize plant leaves, Li *et al.* (55) studied the developmental dynamics of the maize transcriptome. For each location, two biological replicates were sequenced with Illumina's Genome Analyzer II. The reads were mapped to the TE-masked *Zea maize* ZmB73 reference genome version 2 (AGPv2), release 5a, using the GSNAP splicing short read mapper (56). We counted the overlaps between mapped reads and the *Z. maize* gene definitions from the Ensemble Plants database (Release 14). Reads that have multiple possible alignments or that overlap with more than one gene were discarded. DEXUS was applied to this data with its default parameters.

Figure 4 shows the genes with the largest I/NI value and the conditions that were identified by DEXUS. DEXUS found differences in gene expressions between different locations on the leaf despite this information being withheld. Further, it almost always assigned the two replicates to the same condition without knowledge of replicates or leaf locations. Thus, DEXUS assigns conditions reliably.

Eight of the 10 top-ranked genes were also measured by means of microarrays across different leaf locations of *Z. mays* (57). In this microarray experiment, all eight genes show an absolute log fold change of at least 1 between base and tip. Six of these eight genes show an absolute log fold change greater than four.

The two remaining genes, GRMZM2G331518 and AC213612.3_FG001, were not annotated on the microarray. The function of the top-ranked gene GRMZM2G331518 is not known. However, the associated peptide is similar to the defensin-like protein 91 of *Arabidopsis thaliana*, which plays a role in immune response. The gene ranked ninth, AC213612.3_FG001, is a glycine-rich cell wall structural protein, which is compatible with cell walls at different locations having different structure.

At a threshold of 0.1 for the I/NI call, DEXUS called 15 756 differentially expressed genes. Gene set enrichment analysis using the R package goseq (58) yielded to the significant GO terms 'chloroplast' ($P = 1.8e{-}92$) and 'plasma membrane' ($P = 1.3e{-}34$). Further, the GO terms 'cytosolic ribosome' ($P = 9.8e{-}32$), 'chloroplast thylakoid membrane' ($P = 5.4e{-}31$) and 'chloroplast stroma' ($P = 1.8e{-}30$) were significant. All *P*-values were corrected by means of the Benjamini–Hochberg procedure. It is plausible that that chloroplast also differs at different locations on the maize plant leaf. The GO term 'cell wall' was highly significant ($P = 3.9e{-}18$), which supports the above-mentioned hypothesis that the cell walls differ at different locations on the plant leaf.

CONCLUSION

We have introduced DEXUS, an algorithm that identifies differentially expressed transcripts in RNA-Seq data with unknown conditions. DEXUS is appropriate for use with data from cohort, cross-sectional and nonrandomized controlled studies, where conditions are often unknown. In experiments with simulated and real-world data with known conditions, DEXUS successfully found differential expressed transcripts and conditions, although the conditions were withheld from DEXUS. For HapMap individuals, DEXUS detected differentially expressed transcripts, the vast majority of which are related to sex, eQTLs or structural variants. We envisage that DEXUS will evolve into an important tool for analyzing RNA-Seq data in studies with unknown conditions and thus for obtaining new biological and medical knowledge.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Funds from the Institute of Bioinformatics, Johannes Kepler University Linz.

Conflict of interest statement. None declared.

REFERENCES

1. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.
2. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
3. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
4. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
5. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
6. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
7. Labaj,P.P., Leparc,G.G., Linggi,B.E., Markillie,L.M., Wiley,S.H. and Kreil,D.P. (2011) Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, **27**, i383–i391.
8. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
9. Hansen,K.D., Wu,Z., Irizarry,R.A. and Leek,J.T. (2011) Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.*, **29**, 572–573.
10. Jones,A.R., Overly,C.C. and Sunkin,S.M. (2009) The Allen Brain Atlas: 5 years and beyond. *Nat. Rev. Neurosci.*, **10**, 821–828.
11. Heintz,N. (2004) Gene expression nervous system atlas (GENSAT). *Nat. Neurosci.*, **7**, 483.
12. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
13. International HapMap 3 Consortium, Altshuler,D.M., Gibbs,R.A., Peltonen,L., Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
14. The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
15. Montgomery,S.B., Sammeth,M., Gutierrez-Arcelus,M., Lach,R.P., Ingle,C., Nisbett,J., Guigo,R. and Dermitzakis,E.T. (2010) Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, **464**, 773–777.
16. Pickrell,J.K., Marioni,J.C., Pai,A.A., Degner,J.F., Engelhardt,B.E., Nkadori,E., Veyrieras,J.-B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2010) Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, **464**, 768–772.
17. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.-P., Subramanian,A., Ross,K.N. *et al.* (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
18. The Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
19. Lal,A., Lash,A.E., Altschul,S.F., Velculescu,V., Zhang,L., McLendon,R.E., Marra,M.A., Prange,C., Morin,P.J., Polyak,K. *et al.* (1999) A public database for gene expression in human cancers. *Cancer Res*, **59**, 5403–5407.
20. Uehara,T., Ono,A., Maruyama,T., Kato,I., Yamada,H., Ohno,Y. and Urushidani,T. (2010) The Japanese toxicogenomics project: application of toxicogenomics. *Mol. Nutr. Food Res.*, **54**, 218–227.
21. Chen,M., Vijay,V., Shi,Q., Liu,Z., Fang,H. and Tong,W. (2011) FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov. Today*, **16**, 697–703.
22. Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161.
23. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics*, **11**, 94.

24. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
25. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
26. Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
27. Li,J. and Tibshirani,R. (2011) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data. *Stat. Methods Med. Res.*, **22**, 519–536.
28. Wang,L., Feng,Z., Wang,X., Wang,X. and Zhang,X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.
29. Li,J., Witten,D.M., Johnstone,I.M. and Tibshirani,R. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, **13**, 523–538.
30. Tarazona,S., Garca-Alcalde,F., Dopazo,J., Ferrer,A. and Conesa,A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, **21**, 2213–2223.
31. Wu,H., Wang,C. and Wu,Z. (2012) A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, **14**, 232–243.
32. McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
33. Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.
34. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
35. Hochreiter,S., Clevert,D.-A. and Obermayer,K.A. (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.
36. Clevert,D.-A., Mitterecker,A., Mayr,A., Klambauer,G., Tuefferd,M., Bondt,A.D., Talloen,W., Göhlmann,H. and Hochreiter,S. (2011) cn.FARMS: a latent variable model to detect copy number variations in microarray data with a low false discovery rate. *Nucleic Acids Res.*, **39**, e79.
37. Klambauer,G., Schwarzbauer,K., Mayr,A., Clevert,D.-A., Mitterecker,A., Bodenhofer,U. and Hochreiter,S. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.
38. Lloyd-Smith,J.O. (2007) Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS One*, **2**, e180.
39. Piegorsch,W.W. (1990) Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, **46**, 863–867.
40. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. B Met.*, **39**, 1–38.
41. Talloen,W., Clevert,D.-A., Hochreiter,S., Amaratunga,D., Bijnens,L., Kass,S. and Göhlmann,H. (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.
42. Talloen,W., Hochreiter,S., Bijnens,L., Kasim,A., Shkedy,Z. and Amaratunga,D. (2010) Filtering data from high-throughput experiments based on measurement reliability. *Proc. Natl Acad. Sci. USA*, **107**, 173–174.
43. Bottomly,D., Walter,N.A.R., Hunter,J.E., Darakjian,P., Kawane,S., Buck,K.J., Searles,R.P., Mooney,M., McWeeney,S.K. and Hitzemann,R. (2011) Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One*, **6**, e17820.
44. Frazee,A.C., Langmead,B. and Leek,J.T. (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449.
45. Zeller,T., Wild,P., Szymczak,S., Rotival,M., Schillert,A., Castagne,R., Maouche,S., Germain,M., Lackner,K., Rossmann,H. *et al.* (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One*, **5**, e10693.
46. Dimas,A.S., Deutsch,S., Stranger,B.E., Montgomery,S.B., Borel,C., Attar-Cohen,H., Ingle,C., Beazley,C., Arcelus,M.G., Sekowska,M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.
47. Veyrieras,J.-B., Kudaravalli,S., Kim,S.Y., Dermitzakis,E.T., Gilad,Y., Stephens,M. and Pritchard,J.K. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.*, **4**, e1000214.
48. Halbritter,F., Vaidya,H.J. and Tomlinson,S.R. (2011) GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods*, **9**, 7–8.
49. Schlattl,A., Anders,S., Waszak,S.M., Huber,W. and Korbel,J.O. (2011) Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res.*, **21**, 2004–2013.
50. Pinto,D., Marshall,C., Feuk,L. and Scherer,S.W. (2007) Copy-number variation in control population cohorts. *Hum. Mol. Genet.*, **16**, R168–R173.
51. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
52. Schadt,E.E., Molony,C., Chudin,E., Hao,K., Yang,X., Lum,P.Y., Kasarskis,A., Zhang,B., Wang,S., Suver,C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.
53. Blekhman,R., Marioni,J.C., Zumbo,P., Stephens,M. and Gilad,Y. (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, **20**, 180–189.
54. Balasubramanian,S., Zheng,D., Liu,Y.-J., Gang Fang,A.F., Carriero,N., Robilotto,R. and Philip Cayting,M.G. (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol.*, **10**, R2.
55. Li,P., Ponnala,L., Gandotra,N., Wang,L., Si,Y., Tausta,S.L., Kebrom,T.H., Provart,N., Patel,R., Myers,C.R. *et al.* (2010) The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.*, **42**, 1060–1067.
56. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
57. Sekhon,R.S., Lin,H., Childs,K.L., Hansey,C.N., Buell,C.R., deLeon,N. and Kaeppler,S.M. (2011) Genome-wide atlas of transcription during maize development. *Plant J.*, **66**, 553–563.
58. Young,M.D., Wakefield,M.J., Smyth,G.K. and Oshlack,A. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.*, **11**, R14.