Bioinformatics Advance Access published August 13, 2016

Gene expression

SDEAP: A Splice Graph Based Differential Transcript Expression Analysis Tool for Population Data

Ei-Wen Yang [1,2]* and Tao Jiang [1,3,4]*

[1]Department of Computer Science and Engineering, University of California, Riverside, CA
[2]Department of Integrative Biology and Physiology, University of California, Los Angeles, CA
[3]Institute of Integrative Genome Biology, University of California, Riverside, CA
[4]MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST / Department of Computer Science and Technology, Tsinghua University, Beijing, China

*To whom correspondence should be addressed.

Associate Editor: Dr. Inanc Birol

Abstract

Motivation: Differential transcript expression (DTE) analysis without predefined conditions is critical to biological studies. For example, it can be used to discover biomarkers to classify cancer samples into previously unknown subtypes such that better diagnosis and therapy methods can be developed for the subtypes. Although several DTE tools for population data, *i.e.,* data without known biological conditions, have been published, these tools either assume binary conditions in the input population or require the number of conditions as a part of the input. Fixing the number of conditions to binary is unrealistic and may distort the results of a DTE analysis. Estimating the correct number of conditions in a population could also be challenging for a routine user. Moreover, the existing tools only provide differential usages of exons, which may be insufficient to interpret the patterns of alternative splicing across samples and restrains the applications of the tools from many biology studies.
Results: We propose a novel DTE analysis algorithm, called SDEAP, that estimates the number of conditions directly from the input samples using a Dirichlet mixture model and discovers alternative splicing events using a new graph modular decomposition algorithm. By taking advantage of the above technical improvement, SDEAP was able to outperform the other DTE analysis methods in our extensive experiments on simulated data and real data with qPCR validation. The prediction of SDEAP also allowed us to classify the samples of cancer subtypes and cell-cycle phases more accurately.
Availability: SDEAP is publicly available for free at https://github.com/ewyang089/SDEAP/wiki
Contact: {yyang027,jiang}@cs.ucr.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, RNA-Seq has taken a major role in the quantitative analysis of transcript expression and variant discovery and become a vital component of genomic and transcriptomic research (Trapnell *et al.*, 2010). For case-control studies, several differential transcript expression (DTE) analysis methods such as Cuffdiff 2 (Trapnell *et al.*, 2013), DESeq2 (Love *et al.*, 2014), edgeR (Robinson *et al.*, 2010a), and ALEXA-Seq (Griffith *et al.*, 2010) have been developed to discover genes that have differentially expressed transcripts whose abundance values alter between known biological conditions. In addition to the DTE methods, differential splicing (DS) analysis methods such as MISO (Katz *et al.*, 2010), FDM (Singh *et al.*, 2011), MATS (Shen *et al.*, 2012), DEXSeq (Anders *et al.*, 2012), and DiffSplice (Hu *et al.*, 2013) are focused on identifying difference in relative abundance of transcripts. Note that a change in the absolute abundance of a transcript may result from a change in the basal expression level of the corresponding gene, its splicing ratio or both (Trapnell *et al.*, 2013). In other words, DTE methods should be able to discover DS events but not vice versa. Moreover, the DTE and DS analyses may also find many important applications in population based studies, where predefined conditions are unavailable *a priori*. For example, a recent population study to improve the diagnosis and prognosis for breast cancer shows that triple-negative breast cancer can be further classified into six subtypes based on the differential analysis of the expression profiles of patients (Lehmann *et al.*, 2011).

Each of the six subtypes has different sensitivities to targeted therapies. To understand the functions and mechanisms during cell development or differentiation, differentially expressed transcripts are used to characterize cell types or specificity in a mixed population (Brennecke *et al.*, 2013; Buettner *et al.*, 2015; Trapnell, 2015). Due to the emergent demand for computational tools for DTE and DS analyses without known biological conditions, several methods have been proposed recently. SIBER and DEXUS test differential expression by looking at the numbers of reads mapped to individual genes or exons (Tong *et al.*, 2013; Klambauer *et al.*, 2013). An extended protocol of DESeq2 has been published recently to identify differentially expressed genes for single-cell (SC) RNA-Seq data based on ERCC spike-in data (Brennecke *et al.*, 2013). SigFuge compares the areas under normalized read-depth curves to call DS genes (Kimes *et al.*, 2014).

In a DTE analysis, the expression of genes (or transcripts) is summarized by numerical features, called expression features (*e.g.*, read counts of genes or exons in DEXUS and areas under normalized read-depth curves in SigFuge). The DTE of genes is assessed by calculating the variation of the expression features across the samples of different biological conditions. When the biological conditions of samples are not predefined, to test whether an expression feature has experienced differential expression, the input samples are usually clustered based on the expression levels of the feature across the samples. The clusters of the samples are then used as the biological conditions and the variation of the expression feature between the conditions is then measured statistically (Kimes *et al.*, 2014). Clearly, the correctness of the clustering is critical to the DTE analysis when the biological conditions is not available. Both SIBER and SigFuge assume that the input population consists of only two conditions and always cluster the samples into two clusters. This assumption is unrealistic in many applications and an incorrect partition (or clustering) of samples may lead to unreliable conclusions of differential expression tests. Although DEXUS also assumes binary conditions by default, it allows the user to input the number of the biological conditions, which will be used as the number of clusters on all expression features. However, specifying the correct number of the biological conditions directly from samples is difficult for a routine user. Moreover, since some expression features may exhibit little variability on a subset of the biological conditions, the optimal numbers of clusters for individual expression features may differ and may not always be equal to the number of biological conditions defined by the user, even when the user-defined number is correctly specified. Hence, a computational method that can determine the numbers of clusters automatically from data is urgently needed to conduct reliable DTE analyses when the biological conditions are not available. In addition to the challenge in specifying the number of clusters, these tools only report changes in coverage of exons of genes by RNA-Seq reads which do not directly suggest how and where transcript expression diverges as a result of alternative splicing events (ASEs). Such ASE information could be valuable in down-stream applications in their own right, *e.g.* as biomarkers in several cancer studies (Bonnal *et al.*, 2012).

To address these issues in the existing DTE analysis tools, several technical improvements have been made in a new DTE analysis algorithm, called SDEAP. In SDEAP, the number of clusters for expression feature is no longer fixed as two or required as a part of the input. Instead, a Dirichlet infinite mixture model is applied to determine the number of clusters for every expression feature by fitting the data optimally. To further discover ASEs, a graphical data structure, called the *splice graph*, is used in SDEAP to model the structures and abundance of all transcripts of a gene such that ASEs can be represented by decomposing the graph into *alternative splicing modules* (ASMs), as originally proposed in DiffSplice (Hu *et al.*, 2013). However, the graph modular decomposition algorithm employed in DiffSplice is not used here because we have found a counterexample to its correctness (see the Supplementary Materials for a detailed discussion). A corrected algorithm is described in this paper and implemented in SDEAP. Moreover, a method accounting for variability due to technical noise across biological or technical replicates has been adopted to reduce the number of false positives in SDEAP (Anders and Huber, 2010; Robinson *et al.*, 2010b; Brennecke *et al.*, 2013).

To assess the prediction accuracy of SDEAP, extensive experiments on both simulated and real data were conducted to compare SDEAP with DEXUS. SIBER was excluded from our comparisons because the performance of SIBER and DEXUS in detecting the variation of read coverage of genes and exons due to DTE has been compared in (Kimes *et al.*, 2014). DEXUS was shown to constantly outperform SIBER and, moreover, SIBER can only be run on large datasets of more than 50 RNA-Seq samples (Tong *et al.*, 2013). In our simulation experiments, SDEAP outperformed DEXUS by at least 0.17 in the area under precision-recall curve (or AUC_{pr}, which is used as the assessment of overall performance), on average. The numbers of conditions for at least 88% genes in the simulated data were correctly predicted by SDEAP. Although DS analysis is not the main purpose of SDEAP, we compared it with SigFuge in the detection of changes in relative abundance of transcripts by repeating the simulated experiments in (Kimes *et al.*, 2014). SDEAP discovered more DS genes than SigFuge without producing any false positives. Furthermore, the performance of SDEAP is also compared to that of three state-of-the-art DTE methods, namely DESeq2, Cuffdiff 2 and edgeR, that require known (binary) conditions as a part of the input using simulated datasets. When the numbers of individuals from each condition are not highly imbalanced, there is no noticeable difference between the prediction accuracy of SDEAP and that of the three methods. The time efficiency of SDEAP is discussed in the Supplementary Materials. To further demonstrate the utility of SDEAP in real biological applications, the DTE genes predicted by SDEAP were used as biomarkers to classify different cancer subtypes, cell types and cell-cycle phases on several recently published RNA-Seq datasets (Eswaran *et al.*, 2012; Sasagawa *et al.*, 2013). These applications are interesting because some critical diseases, *e.g.*, breast cancer (BC), are known as heterogeneous diseases with a variety of transcriptomic alterations that severely affect the diagnosis and prognosis of the diseases (Lehmann *et al.*, 2011). Finding DTE genes that could be used as transcriptomic biomarkers to identify subtypes of such diseases could be important for the design of clinical trials to investigate targeted treatments. Moreover, the expression patterns of transcripts in individual cells of different cell types or cell-cycle phases, which can be revealed by the SC RNA-Seq technology nowadays, are fundamental to studies on alternative cellular functions during the development of a tissue or an organ (Sasagawa *et al.*, 2013; Trapnell, 2015). Our real data experiments demonstrate that the classification of RNA-Seq samples using the ASEs from SDEAP is much more consistent with the real biological conditions (*i.e.*, cancer subtypes, cell types and cell-cycle phases) than using the differentially expressed exons predicted by DEXUS. The prediction results of both methods were also compared with the qPCR validations of gene expression. More validated DTE genes were covered by the prediction of SDEAP. These results suggests that SDEAP also performs DTE analysis well on real population data.

2 Methods

An expressed segment is an exonic region delimited by two exon boundaries. A splice graph $G(V \cup \{s,t\}, E)$ of a gene g is a weighted and directed acyclic graph where every vertex $v \in V$ denotes an expressed segment R_v. For every pair of vertices u and v, there is a directed edge (u,v) from u to v if the expressed segment R_v immediately follows R_u in some transcript of the gene g. In addition to the vertices V representing expressed segments, two artificial vertices s and t are included in G to indicate the beginning and end of all transcripts of the gene g, respectively. The vertex s is connected to every vertex corresponding to the very first

expressed segment of a transcript of the gene g and every vertex denoting the last expressed segment of a transcript is connected to t. In SDEAP, we assume that splice graphs are provided as the input. Given all RNA-Seq reads mapped to the gene g in an RNA-Seq sample, the weight of a vertex v, $w(v)$, is defined as the number of reads mapped to the region R_v and the weight of the edge (u, v), $w(u, v)$, is the number of reads that span the two expressed segments R_u and R_v.

A vertex $u \in V$ *pre-dominates* a vertex $v \in V$ if every path from the artificial vertex s to v contains u. A vertex $w \in V$ *post-dominates* a vertex $v \in V$ if every path from v to the artificial t contains w. An ASM (or alternative splicing module) is an induced subgraph $H(s_1, t_1) = \{V_H, E_H, s_1, t_1\}$ of G with the entry s_1 and the exit t_1 outside H that satisfies the following conditions (Hu *et al.*, 2013): (1) (Single entry) All edges from $(G-H)$ to H come from s_1; (2) (Single exit) All edges from H to $(G-H)$ go to t_1; (3) (Alternative paths) Let $d_+(u)$ denote the number of outgoing edges from the vertex u and $d_-(u)$ the number of incoming edges of u. Then $d_+(s_1) > 1$ and $d_-(t_1) > 1$; (4) (Minimality) There does not exist a vertex $v \in V_H$, such that v post-dominates s_1 or pre-dominates t_1 in $H(s, t)$. Moreover, an ASM $H_1(t_1, s_1)$ can be a subgraph of another ASM $H_2(t_2, s_2)$. If there is no ASM that contains H_1 and is contained by H_2, $H_1(s_1, t_1)$ is said to be a child ASM of H_2 and H_2 is the parent ASM of H_1. The *abstraction* of an ASM $H_2(s_2, t_2)$ is a graph obtained by replacing every child ASM $H_1(s_1, t_1)$ of $H_2(s_2, t_2)$ with an artificial edge (s_1, t_1). An ASM path is a path from s_2 to t_2 in the abstraction of an ASM $H_2(s_2, t_2)$.

2.1 Discovery of ASMs

The algorithm proposed in DiffSplice for discovering ASMs is not used here because we have found a counterexample to its correctness (see the Supplementary Materials for a detailed discussion). In our new ASM discovery algorithm, every ASM is discovered before its parent and then shrunk into an artificial edge immediately. All vertices of the input splice graph G are sorted by topological sort (Cormen *et al.*, 2001). Let β be the topological order of vertices and $\beta(u)$ the index of vertex u in the order. If there is a path from vertex u to vertex v, then the index of vertex u is greater than the index of vertex v, *i.e.,* $\beta(u) > \beta(v)$. Assume that an ASM $H_2(s_2, t_2)$ has a child ASM $H_1(s_1, t_1)$. If all vertices are traversed in the order of β, the exit t_1 of H_1 must be traversed before the exit t_2 of its parent H_2, *i.e.*, $\beta(t_1) < \beta(t_2)$. Let $\bar{\beta}$ be the reverse of the topological order β. Similarly, s_1 is always traversed before s_2, *i.e.*, $\bar{\beta}(s_1) < \bar{\beta}(s_2)$. Moreover, for any ASM $H(s_1, t_1)$, $\beta(s_1) < \beta(t_1)$. To discover ASMs, every vertex v with the in-degree $d_+(v) > 1$ (or the out-degree $d_+(v) > 1$) is selected as a candidate of an exit (or an entry, respectively.) of ASMs. We firstly traverse all candidates of the entry in the order of $\bar{\beta}$. When a candidate of the entry u is visited, for every candidate exit v such that $\beta(v) > \beta(u)$, v is chosen in the order of β to pair up with u as a candidate entry-exit pair (u, v). The union of all paths bounded by (u, v) is checked if it is an ASM or not. If the union is an ASM, we replace it by an artificial edge. In this order of enumerating candidate entry-exit pairs, for every ASM $H_1(s_1, t_1)$ and its parent $H_2(s_2, t_2)$, the candidate entry-exit pair (s_1, t_1) is always tested before (s_2, t_2). Thus, every ASM is guaranteed to be identified before its parent. The time complexity of identifying ASMs from a splicing graph G is $O(|V|^3 + |V|^2|E|)$. Please see the Supplementary Materials for a pseudocode of the algorithm for discovering ASMs and the derivation of the time complexity.

2.2 Estimation of Expression Features

In SDEAP, the expression features of a gene are the abundance values of all ASM paths in the ASMs of the gene by default. However, when the number of ASM paths in an ASM increases, estimating the abundance values of ASM paths is less accurate due to non-identifability as discussed in DiffSplice (Hu *et al.*, 2013). Hence, if the number of paths in an ASM is greater than 3, instead of the abundance values of the ASM paths, we use the abundance values of the expressed segments and junctions in the ASM as its expression features. The abundance values are measured in terms of the average RNA-Seq fragment coverage per kilobases per million fragments (FPKM). For an expressed segment, the FPKM is the number of fragments mapped to the expressed segment divided by the length of the segments in kilo bases and the size of the RNA-Seq fragment library in million bases. For a junction, because the length of mapped reads is the length of the region where each junction read spans, the FPKM of a junction is the number of mapped reads divided by the read length and the size of the library. Given an ASM $H(u, v)$, let the ASM paths of $H(u, v)$ be $P = \{p_1, p_2, ..., p_N\}$ such that all expressed segments and junctions covered by the paths can be represented as a binary matrix $A_{M\times N} = (a_{i,j}), 1 \leq i \leq M$ and $1 \leq N \leq j$, where each of the M rows represents an expressed segment or a junction and each of the N columns represents a path. If the path p_j includes an expressed segment or junction i, $a_{i,j} = 1$. Otherwise, $a_{i,j} = 0$. Every expressed segment or junction i is associated with a value of effective length l_i. If the row i represents an expressed segment, l_i, is the length of the expressed segment. Otherwise, l_i is set as the length of the RNA-Seq reads. Note that the first and last vertices and artificial edges of each path are not included in the rows of $A_{M\times N}$. Let the abundance values (FPKMs) of the paths be $X = \{x_1, x_2, ..., x_N\}$. All mapped reads are assumed to be evenly distributed on each of the paths. Let the observed number of reads falling into the i-th expressed segment or junction be r_i. The expression levels of the paths, $X = \{x_1, ..., x_N\}$, can then be determined by using the abundance values X^* that minimizes the following residual sum of squares as done in IsoInfer (Feng *et al.*, 2011):

$$X^* = \arg\max_X \sum_{i=1}^{M} (\frac{r_i}{l_i} - \sum_{j=1}^{N} a_{i,j}x_j)^2 \quad (1)$$

with respect to the constraints that $x_j \geq 0$ for all $1 \leq j \leq N$. In the implementation of SDEAP, an R package opt is used to solve the above quadratic optimization problem by the L-BFGS-B algorithm (Byrd *et al.*, 1995).

2.3 Selecting Informative Expression Features

When there are n RNA samples, each expression feature f has n instances, $F = \{f_1, f_2, ..., f_n\}$, where f_i represents the abundance value of f in the i-th RNA-Seq sample. In the literature (Anders and Huber, 2010; Robinson *et al.*, 2010b), the observed variance of the instances is postulated as due to technical noise and biological variation. To control the number of false positives, only the expression features with variance of the instances significantly higher than the variance due to technical noise are considered as informative features and used in the DTE analysis. The expected variance ρ due to technical noise is usually modeled as a quadratic function of the observed mean μ_f of f such that $\rho(\mu_f) = \mu_f + \phi{\mu_f}^2$, where the parameter ϕ is a parameter to be estimated by regression using all expression features in the samples of the same condition. However, since the biological conditions are not given *a priori* in our case, ϕ is approximated by using all input samples as done in the literature (Brennecke *et al.*, 2013). Let the observed variance of the instances of an expression feature f be $\hat{\rho}_f$. An expression feature f is selected as an informative feature if $\hat{\rho}_f/\rho(\mu_f) > \gamma$, where γ is a user defined threshold, as employed in (Anders and Huber, 2010; Buettner *et al.*, 2015).

2.4 Testing Differential Transcript Expression

The Dirichlet infinite mixture model used in SDEAP is a Gaussian mixture model that allows us to determine the number of components from data automatically. To illustrate the Dirichlet infinite mixture model, we start from a finite mixture model of fixed k components. Let F be the n instances

of an informative expression feature f and $C = \{c_1, c_2, ..., c_n\}, c_i \in \{1, 2, ..., k\}$ be the set of component indices such that each index c_i indicates which component f_i belongs to. In a finite mixture model, the joint probability of C and F can be written as:

$$Pr(C, F|\pi, \theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} I(c_i = j)\pi_j N(f_i|\theta_j), \quad (2)$$

where $\pi = \{\pi_1, \pi_2, ..., \pi_k\}$ such that π_j is the probability of an instance f_i belonging to component j, I is an indicator function and the base distribution $N(f_i|\theta_j)$ is the distribution of instances in the component j given the parameters $\theta = \{\theta_1, \theta_2, ..., \theta_k\}$ of the k components.

In a Dirichlet infinite mixture model, each $\pi \geq 0$, is assigned a Dirichlet prior with k concentration parameters such that $\pi \sim \text{Dirichlet}(\alpha/k, ..., \alpha/k)$ where α is a hyper parameter. At the same time, the number of components, k, is allowed to go to infinity. By integrating the mixing proportion π, the conditional prior probability of f_i belonging to component j, *i.e.*, $c_i = j$, is

$$Pr(c_i = j|c_1, ..., c_{i-1}) \rightarrow \frac{n_{i,j}}{i - 1 + \alpha}, \quad (3)$$

where $n_{i,j}$ is the number of components i' with $c_{i'} = j, i' < i$. By combining the likelihood function of Eq. (2) with Eq. (3), the conditional posterior probability functions for $c_i = c_j$ with the given model parameters, μ and ρ, and observed feature values F are

$$Pr(c_i = j|C_{-i}, F, \mu, \rho) \propto b\frac{n_{-i,j}}{n - 1 + \alpha} N(f_i|\theta_j), \quad (4)$$

where b is a constant for normalization, C_{-i}=$C - \{c_i\}$ and $n_{-i,j}$ is the number of $c_{i'} \in C_{-i}$ such that $c_{i'} = j$.. The conditional posterior probability functions for $c_i \neq c_j$, for all $j \neq i$ are

$$Pr(c_i \neq c_j, i \neq j|C_{-i}, F, \mu, \rho) \propto b\frac{\alpha}{n - 1 + \alpha} \int N(f_i|\theta_j) dG_0, \quad (5)$$

where G_0 is the prior probability of θ. More detailed derivation of the prior and posterior probabilities of the parameters is discussed in (Neal, 2007).

The component indicators C that maximize the joint probability given in Eq. (2) can be determined by sampling from a Markov chain of the posterior probabilities with Eq. (4) and Eq. (5) as its equilibrium distribution. To further improve the execution time of the sampling process, an algorithm based on variational inference has been proposed in the literature (Blei and Jordan, 2006). Here, we assume that reads from a transcript are uniformly distributed among the transcript. As justified in the literature (Feng *et al.*, 2011), the distribution of the FPKM values can be approximated by a Gaussian distribution based on the assumption. Hence, the Dirichlet infinite mixture model is appropriate to fit the observed FPKM values from RNA-Seq data. The implementation of the variational inference algorithm in the Python package scikit-learn is used in SDEAP (Pedregosa *et al.*, 2011). After the clustering of the instances F, a one-way ANOVA test is performed to test if the clusters are indeed significantly different (Fisher, 1958). The p-values from the ANOVA test are adjusted for multiple comparisons and the adjusted p-values are called the false discovery rates (FDRs) as done in (Benjamini and Hochberg, 1995). If the FDR of an informative feature f is smaller than a given threshold, *e.g.*, 0.1, we conclude that the informative feature f differentially expressed across the input samples.

2.5 Evaluation Metrics

All our experimental results are evaluated in terms of precision (PRE), $\text{PRE} = \text{TP}/(\text{TP} + \text{FP})$, and recall (REC), $\text{REC} = \text{TP}/(\text{TP} + \text{FN})$, where TP is the number of true positives, FP the number of false positives and FN the number of false negatives. Because the number of equally or differentially expressed transcripts in DTE analysis is usually unbalanced, a recently published study suggests that precision-recall curves are more informative than ROC curves on unbalanced data (Saito and Rehmsmeier, 2015). Hence, the area under the precision-recall curve (or AUC_{pr}) is used as a measure of the overall performance of a prediction method in our tests. In this paper, an R package PRROC is used to calculate the PRE, REC and AUC_{pr} scores (Grau1 and Keilwagen, 2015). To measure the similarity between clusters of RNA-Seq samples and real biological conditions, a widely used assessment, Jaccard index, is calculated as in the literature (Sneath, 1957).

3 Experimental Results

SDEAP and DEXUS are tested on both simulated and real datasets. In our simulation study, several realistic configurations of real RNA-Seq data are considered. In the first simulation, bimodal RNA-Seq data, as assumed in DEXUS, are simulated, while data are generated from three or more overlapping groups in the second simulation. Noise unique to single-cell RNA-Seq is introduced in the third simulation to test the robustness of the methods in dealing with data with high background variance. The simulation study performed in (Kimes *et al.*, 2014) is repeated as the forth simulation experiment to assess the performance of SDEAP in calling DS genes. The results of SDEAP are then compared with those of SigFuge reported in (Kimes *et al.*, 2014). Moreover, the performance of SDEAP is also compared with that of three popular DTE analysis methods for data with predefined biological conditions, namely DESeq2, Cuffdiff 2 and edgeR. In our experiments on real data, we assume that if the DTE genes are correctly predicted, the expression features of the predicted DTE genes can then be used to reconstruct the biological conditions of the input samples. Hence, given the expression features from the predicted DTE genes, all samples are clustered by the hierarchical clustering package MADE4 which is widely used in gene expression analysis (Culhane *et al.*, 2005). Hierarchical clustering is performed using the average linkage clustering algorithm in MADE4 while the measure of similarity is the Pearson correlation of expression features. Note that, the partitions (or clusterings) made from the Dirichlet model are for calling DTE genes and may be coarser than the partition corresponding to the biological conditions. Ideally, the biological conditions of samples should represent a refinement of the partition obtained on each individual expression features, as illustrated by an example in Supplementary Fig. S1. To avoid biases due to the sizes of genes, if a predicted DTE gene has more than one differentially expressed feature, the feature with the greatest significance measurement, *i.e.*, FDRs in SDEAP or I/NI scores in DEXUS, is selected as the representative expression feature for the DTE gene. Although our real data analysis will include two experiments on single-cell RNA-Seq data, the extended protocol of DESeq2 for single-cell RNA-Seq data is not compared in the experiments because it requires spike-in ERCC information which is not provided in our single-cell RNA-Seq datasets. Note that DEXUS provides the I/NI scores as the output which cannot be compared with the FDR scores from SDEAP. Hence, we set a widely used FDR value of 0.1 as the threshold to call DTE genes in SDEAP. DEXUS is compared by using the same number of top-ranked genes in its prediction.

3.1 Experiments on Simulated Data

3.1.1 Simulation of Regular RNA-Seq Reads

In our simulation studies, we simulate RNA-Seq reads from the Ensembl GRCh38 (hg38) reference genome as the input data to test the above-mentioned DTE analysis methods. The simulated datasets consist of various numbers of samples from two or more predefined conditions according to different experimental designs. The simulation process for each datasets consists of three steps. The first step is to create an expression profile, which includes the FPKM values of all transcripts annotated in the hg38 reference genome, for each of the predefined conditions. Then, the observed numbers of reads from the transcripts in every sample of a predefined condition are randomly drawn from the negative binomial distribution

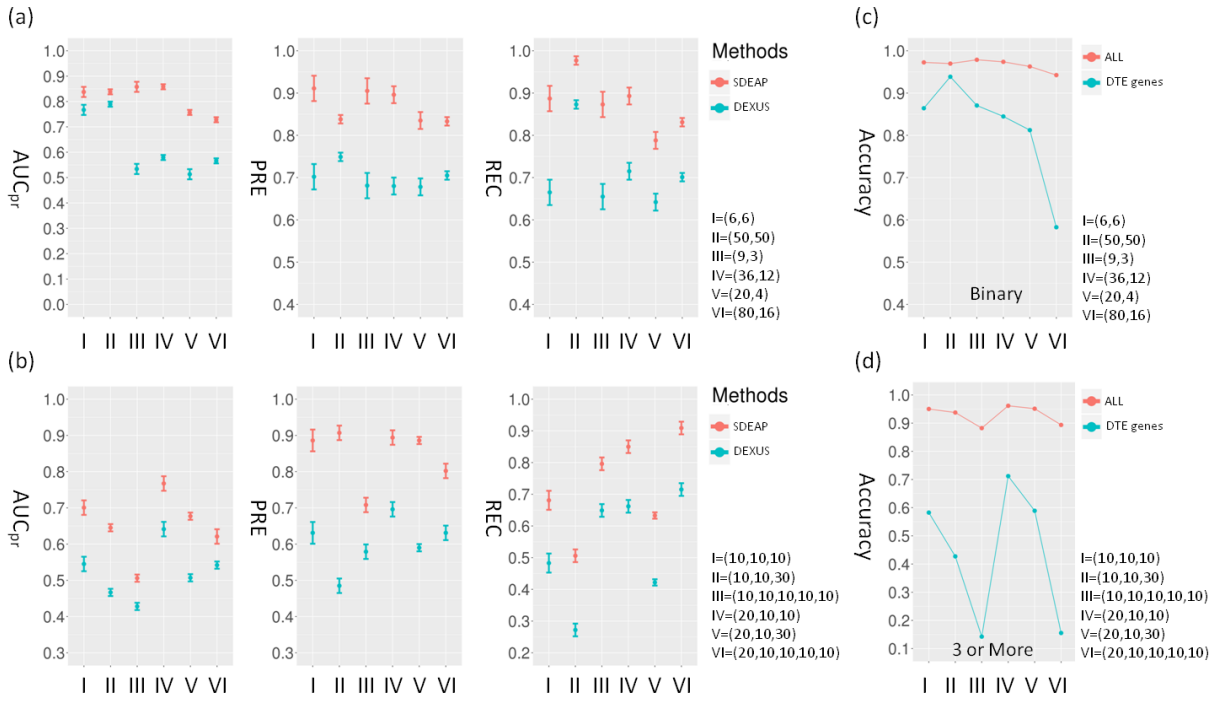

Fig. 1. The performance of the two DTE analysis methods, SDEAP and DEXUS, on simulated regular RNA-Seq data. Plot (a) compares the AUC_{pr}, PRE and REC scores achieved by the two methods on simulated RNA-Seq samples of binary conditions, while Plot (b) provides the scores on samples of three or more conditions. The X-axis shows different combinations of group sizes, where $(n_1, n_2, \ldots)$ indicates the number n_i of replicates in condition i. The Y-axis represents the AUC_{pr}, PRE and REC scores averaged over 10 replicates. The error bars in each plot demonstrate the standard deviation of the scores. Plots (c) and (d) show the proportions of genes, where the numbers of conditions were correctly estimated, on the simulated RNA-Seq samples of binary and three or more conditions, respectively. In Plots (c) and (d), the pink lines indicate the proportions over all genes while the blue lines represent the proportions over the true DTE genes. The X-axis in these plots shows the different combinations of condition sizes as in Plots (a) and (b).

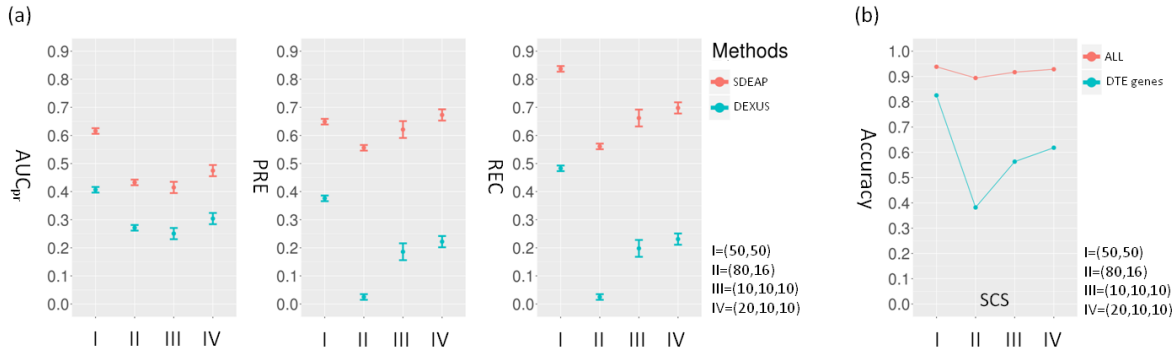

Fig. 2. The performance of SDEAP and DEXUS on simulated SC RNA-Seq data. Plot (a) presents the AUC_{pr}, PRE and REC scores on simulated SC RNA-Seq samples with different configurations of group sizes as considered in Fig. 1(a) and 1(b). Again, the X-axis shows different combinations of group sizes and the Y-axis illustrates the three scores. Plot (b) shows the accuracy of estimating the numbers of conditions on simulated SC RNA-Seq data, which is demonstrated in the same way as in Fig. 1(c) and 1(d).

parameterized by the FPKM values. Finally, paired-end reads are synthesized from the annotated transcripts according the observed numbers. The details of the three steps are given below.

We describe first how the two expression profiles for binary conditions are created. The construction of the profiles for three or more conditions will be introduced later in Section 3.1.3. In the first of the two expression profiles, the FPKM value of each transcript is randomly drawn from a log-normal distribution as done in the literature (Li and Jiang, 2012). Here, only genes that have multiple transcripts are considered. A transcript is said to be *detectable* if its FPKM value is greater than 0.1. A gene with multiple transcripts is discarded if the sum of the FPKM values of its transcripts is less than 1.0 or none of its transcripts is detectable according to the first expression profile. Hence, our simulated datasets are comprised of 3089 genes. To create the second expression profile, among the 3089 genes, 308 ($\sim$10%) genes are chosen as DTE genes. All the DTE genes are evenly divided into three categories: up-regulated, down-regulated and differentially spliced. For each up-regulated gene, a detectable transcript is randomly selected and its abundance is increased by a factor of at least 4, a widely used threshold to define differential expression in the literature (Yang *et al.*, 2013; Bullard *et al.*, 2010). Similarly, for each down-regulated gene, the abundance of a randomly selected (detectable) transcript is decreased by a factor of at least 4. For each differentially spliced gene, the maximum

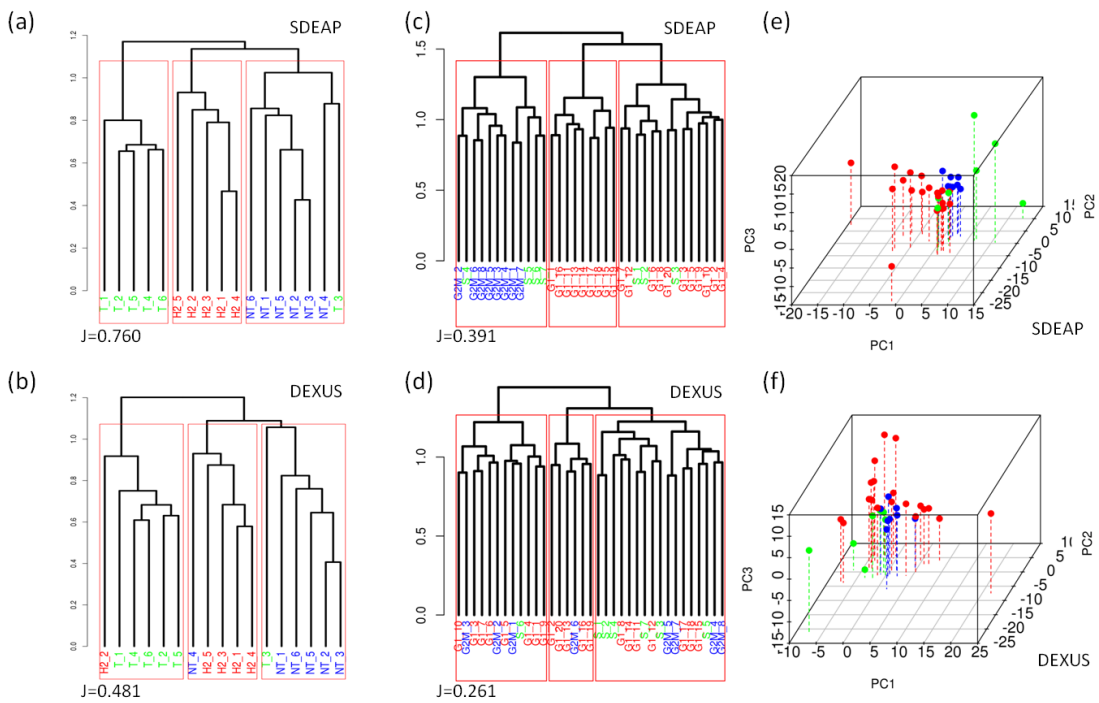

Fig. 3. The clustering results on two real RNA-Seq datasets. Plot (a) and (b) present the dendrograms of the hierarchical clustering by SDEAP and DEXUS on the BC dataset. The Y-axis is the measurement of similarity between the samples and X-axis are the labels of each sample. The HER2 samples are colored red, the TNBC samples green and the non-TNBC samples blue. The three red boxes in each dendrogram illustrate three clusters obtained by the corresponding method. Similarly, Plot (c) and (d) demonstrate the hierarchical clustering by SDEAP and DEXUS on the SC samples of different cell-cycle phases. Every red label is a cell in the G1 cell-cycle phase and every blue label a cell in the G2/M phase. Cells in the S phase are represented by green labels. The three red boxes in each dendrogram illustrate three clusters. The expression profiles obtained from SDEAP and DEXUS on the samples in the SC datasets are further visualized by the PCA analysis in Plot (e) and (f) respectively. The dots are colored in the same away as in Plot (c) and (d).

and minimum abundance values of its transcripts are swapped. Different from up- and down-regulated genes, total amount of reads mapped to differentially spliced genes may not have significant difference between the samples from the binary conditions. All three types of DTE genes are included and tested in the following simulation studies. The other 2781 genes are selected as equally expressed (EE) genes and the FPKM values of their transcripts remain the same in both expression profiles.

Given an FPKM value $x_{t,c}$ of transcript t in the expression profile for a predefined condition c, the number of observed RNA-Seq reads, $r_{t,i}$, from the transcript t in a sample i of the condition c follows the negative binomial distribution $NB(\theta_{t,c}, \phi)$, where $\theta_{t,c}$ is the mean value of the read counts and ϕ is the dispersion rates. The variance of $r_{t,i}$ modeled by the distribution is $\theta_{t,c} + \phi\theta_{t,c}^2$. The mean value $\theta_{t,c} = x_{t,c} \times l_t \times \hat{s}$, where l_t is the length of t in kilo bps and $\hat{s}$ is the size of the RNA-Seq library in millions. We simulate RNA-Seq libraries of moderate sizes with $\hat{s}$ set to 40M. The dispersion rate ϕ is set to be 0.179 as done in the literature (Kimes *et al.*, 2014). In real RNA-Seq data, the observed variance of the read counts is significantly greater than the sample variance modeled by the distribution (Oberg *et al.*, 2012) due to outliers. Two studies based on real RNA-Seq data show that approximately 5% genes from the same biological condition have significantly higher variance in transcript expression than expected due to outliers (Oberg *et al.*, 2012; Gierlinski *et al.*, 2015). To simulate datasets that reflect real RNA-Seq data as much as possible, 5% genes are selected as genes that contain outliers in the simulated samples. Extreme high values of transcript expression are usually detected in approximately 10% of real RNA-Seq samples (Gierlinski *et al.*, 2015). To simulate such extreme high expression values, we allow the randomly drawn read counts $r_{t,i}$ for transcript t from the selected outlier genes to have a 10% probability of being amplified from 5 to 10 times in sample i as done in (Zhou *et al.*, 2014). In addition to the existence of extremely highly expressed transcripts, a study shows that the exons of lowly expressed transcripts could be ubiquitously missing in every one of two technical or biological replicates (McIntyre *et al.*, 2011). To account for such missing-value events, among the 5% selected outlier genes, we allow the randomly drawn read count $r_{t,i}$ for a transcript t with abundance lower than 1.0 to have a 30% to 50% probability of being assigned zero in sample i.

Given the randomly drawn read counts $r_{t,i}$, $r_{t,i}$ paired-end RNA-Seq reads of 50 bps each are obtained from both ends of $r_{t,i}/2$ cDNA fragments synthesized from the genomic region of the transcript t in the hg38 reference genome. The density distribution of the synthesized cDNA fragments along the genomic region of the transcript t follows a positional profile that reflects positional biases due to complementary DNA fragmentation (Li and Jiang, 2012). To avoid biased assessment of prediction accuracy due to random sampling, the simulation experiment on each configuration is repeated 10 times and the average performance of each method is reported in the following discussion.

3.1.2 Performance on RNA-Seq Data from Two Conditions

Both SDEAP and DEXUS are tested on several simulated datasets with binary conditions. The combinations of group sizes, n_1 and n_2, in the simulated datasets are from the literature (Kimes *et al.*, 2014; Klambauer *et al.*, 2013). If $n_1 > n_2$, the n_1 RNA-Seq samples are called the major group and the n_2 samples are called the minor group. The performance results of both methods on all group size configurations are summarized in Fig. 1(c). The accuracy in estimating the number of conditions, as demonstrated in Fig. 1(b), is measured by the proportion of genes that have at least one expression feature resulting in correctly estimated number

of conditions. To evaluate the ability of false positive control, the false positive rates at the REC score (*i.e.*, sensitivity) 0.9 on the 2781 EE genes are illustrated in Supplementary Fig. S2(a).

A comparison of the overall performance (the AUC_{pr} scores) of the methods shows that SDEAP clearly outperforms DEXUS. The average AUC_{pr} of SDEAP over all configurations is 0.809 and the average AUC_{pr} of DEXUS is only 0.624. SDEAP outperforms DEXUS by at least 0.09 and 0.1 in the PRE and REC scores, respectively. In general, increasing the number of samples benefits the REC scores of both methods. The performance of DEXUS is generally lower than what was reported before in the literature (Klambauer *et al.*, 2013). This is because outliers frequently observed in real data were not included in the experiments performed in (Klambauer *et al.*, 2013). The false positive rates of SDEAP are significantly lower than those of DEXUS, as shown in Supplementary Fig. S2(a). This shows that SDEAP controls false positives better due to a robust background variance analysis where most of the outliers could be filtered out due to their variance which is generally lower than that of the DTE genes. As summarized in Fig. 1(c), SDEAP is able to identify the correct number of conditions for at least 94.2% of the genes. We notice that its accuracy in predicting the correct number of conditions on DTE genes drops significantly on the configuration (80,16). This is because the clustering algorithm of SDEAP on the minor group is more sensitive to the outliers when the proportion of the minor group decreases. Clearly, the accuracy in estimating the number of conditions is correlated with the quality of predicted DTE genes. Notably, the prediction accuracy of DEXUS is very sensitive to changes in the proportion of the minor group such that its AUC_{pr} score drops drastically from 0.789 on the configuration (50,50) to 0.513 on the configuration (9,3). The performance of SDEAP is more robust with respect to the decrease in the proportion of the minor group until it drops down to 16.6% in the last two experiments with the configurations (20,4) and (80,16). A possible explanation for the deteriorated performance of SDEAP is that, on these two imbalanced configurations, the observed variance of the expression features in some DTE genes is close to the background variance due to the outliers and these DTE genes may be filtered out in the background variance analyses. The other reason could be due to less accurate clustering results on the DTE genes.

3.1.3 Performance on RNA-Seq Data from Three or More Conditions
To simulate datasets from three conditions, we start with the two expression profiles for binary conditions described in the previous simulation experiment and define the third expression profile as the average of the first and second profiles. For example, if the FPKM values $x_{t,a}$ and $x_{t,b}$, of a transcript t are 1.0 and 4.0 in the first and second profiles for two conditions a and b, respectively, then the FPKM value $x_{t,c}$ in the third profile for the condition c is 2.5. These configurations are designed in order to reflect the reality in some challenging practical applications, *e.g.*, RNA-Seq data sampled during cell development (Äijö *et al.*, 2014). To further test the capabilities of SDEAP and DEXUS, we include simulated RNA-Seq samples from five conditions as follows. Given the profiles for the previous three conditions, the fourth expression profile is the average of the first and the third profile, while the fifth expression profile is the average of the second and the third profile. For example, if the FPKM values $x_{t,a}$, $x_{t,b}$ and $x_{t,c}$, of a transcript t are 1.0, 4.0 and 2.5, respectively, then $x_{t,d}$ and $x_{t,e}$ are 1.75 and 3.25 in the profiles for the fourth and fifth conditions d and e. Since DEXUS does not provide any tool to estimate the correct number of conditions when clustering instances of every expression feature, the default number (two) of conditions is still used for DEXUS. The performance of SDEAP and DEXUS on various configurations with three or more conditions is reported in Fig. 1(b). The accuracy in estimating the numbers of conditions by SDEAP is summarized in Fig. 1(d). The false positive rates at the REC score 0.9 of the two methods is illustrated in Supplementary Fig. S2(b).

Again, SDEAP significantly outperforms DEXUS overall. The AUC_{pr} scores of SDEAP are at least 0.126 higher than that of DEXUS in every experimental setting. This could be due to the incorrect assumptions on the number of clusters by DEXUS. When the correct numbers of conditions, three or five, are provided to DEXUS as a part of the input, the performance of DEXUS can be improved as illustrated in Supplementary Tab. S1. This demonstrates the importance of being able to know the correct number of conditions in DTE analysis. The AUC_{pr} scores of SDEAP drastically decrease on the configuration (10,10,10,10,10). This is because the expression features of DTE genes have low observed variance so close to the background variance that the features are very likely considered as non-informative in the background variance analysis. In general, the expression features of DTE genes present higher variance on the configuration (20,10,10,10,10) than on (10,10,10,10,10) and are thus less likely considered as non-informative. This explains why SDEAP performs better on the configuration (20,10,10,10,10). The comparisons in terms of precision scores and false positive rates suggest again that SDEAP controls false positives better than DEXUS. SDEAP is able to predict the correct number of conditions for at least 88.2% of the 3089 genes on average. We notice that the accuracy in estimating the correct number of clusters for the 308 DTE genes drops down to 0.142 and 0.155 on the configurations (10,10,10,10,10) and (20,10,10,10,10), respectively. This is due to the fact that the separation of expression features becomes more subtle on these two configurations.

3.1.4 Robustness on Simulated SC Data
Single-cell RNA-Seq serves as a fundamental tool to measure the expression of transcripts in individual cells and has numerous applications in biological research. However, due to the low abundance of transcripts in an individual cell, the technical noise in single-cell (SC) RNA-Seq data is much higher than that in regular RNA-Seq data (Buettner *et al.*, 2015). Moreover, some transcripts of genes with moderate or high abundance in one cell may not be detected in another cell (Buettner *et al.*, 2015). In our simulated SC RNA-Seq data, outliers unique to SC RNA-Seq data, as described in the literature, are included in the simulation to test the robustness of the two DTE analysis methods on noisy RNA-Seq data. See the Supplementary Materials for the protocol used to simulate our SC RNA-Seq datasets. We reuse the group sizes in the above experiments on regular RNA-Seq data to study the prediction accuracy of SDEAP and DEXUS on balanced and unbalanced SC data. The assessment of the performance of the two methods on four simulated SC dataset is summarized in Fig. 2(a) and the accuracy in estimating the numbers of conditions by SDEAP is summarized in Fig. 2(b). The false positive rates of the two methods are shown in Supplementary Fig. S2(c).

Similar to the results on the simulated regular RNA-Seq data, SDEAP significantly improves the precision and recall scores of DEXUS by at least 0.273 and thus achieves much better overall performance score AUC_{pr}. This shows that SDEAP is more robust with respect to the increased background noise. Note that among the four configurations, DEXUS has very low prediction accuracy, at most 0.222 and 0.231 in precision and recall, respectively, except on the balanced binary configuration (50,50). This suggests that DEXUS may not be suitable for treating SC data. The conclusion is consistent with the results of our later experiments on real SC data. Due to the high background noise, SDEAP achieves significantly lower false positive rates than those obtained by DEXUS as illustrated in Supplementary Fig. S2(c). However, the false positive rates of SDEAP on the SC datasets are much higher than those on the regular RNA-Seq data. At the same time, the accuracy in estimating the number of conditions is lower than that on the regular RNA-Seq samples in general as shown Fig. 2(b).

3.1.5 Detecting Changes in the Relative Abundance of Transcripts
To compare the performance of SDEAP, DEXUS and SigFuge in DS analysis, we repeat the simulation experiments in (Kimes *et al.*, 2014) as follows.

Two hypothetical gene models concerning two transcripts, isoform t_1 and isoform t_2, are considered as illustrated in Supplementary Fig. S3. The first model has a cassette exon excluded from isoform t_1 but retained in isoform t_2. The second model contains mutually exclusive cassette exons. For each configuration, RNA-Seq samples of binary conditions are simulated. Let the relative abundance values of the two transcripts t_1 and t_2 be ψ_1 and ψ_2, respectively. In the first configuration, the first gene model is used where the relative abundance values, (ψ_1, ψ_2), are set as (0.5, 0.5) in order to evaluate the number of false positives in prediction. In the second configuration, (ψ_1, ψ_2) is set as (0.75, 0.25) and (0.25, 0.75) for the binary conditions, in order to evaluate the number of true positives. In the third configuration, the same abundance values in the second configuration are applied to the second gene model. In each configuration, two combinations of group sizes (n_1, n_2), (50, 50) and (75, 25), are considered. The detailed simulation of RNA-Seq reads is described in the Supplementary Materials. In each configuration, 100 individual genes of each exon model are tested as done in (Kimes *et al.*, 2014) and the DS analysis results are summarized in Supplementary Tab. S2. SDEAP consistently achieves high sensitivity and precision with zero false positives in all three experimental settings. SigFuge also achieves a high sensitivity in the second configuration. However, its sensitivity drops in the third configuration when the sizes of the groups become unbalanced.

3.1.6 DTE Analyses with or without Predefined Conditions

In this subsection, we evaluate how much the known conditions of samples may contribute to the accuracy of a DTE analysis by comparing the prediction results of SDEAP with those of three state-of-the-art DTE methods, Cuffdiff2, DESeq2 and edgeR, which require predefined conditions. The latest implementations of the three methods, Cufflinks 2.2.1, DESeq2 1.12.2 and edgeR 3.14.0, are downloaded and run on the simulated datasets of binary conditions. The biological conditions of the samples are provided to Cuffdiff2, DESeq2 and edgeR as a part of the input. A FDR value 0.1 is set as the threshold to call DTE genes in SDEAP. The PRE and REC scores of the three methods are calculated by using the same number of top-ranked genes in their predictions. The assessment of the performance on the simulated datasets is summarized in Supplementary Fig. S4.

Among the three DTE methods requiring known conditions, DESeq2 achieves the best overall performance, *i.e.,* AUC scores. The AUC scores of SDEAP, ranged from 0.858 to 0.837, are comparable to those of DESeq2 on the first four configurations. However, on the configurations (20,4) and (80,16), where group sizes are highly imbalanced, the AUC scores of SDEAP drop to 0.726 and 0.737 on the two configurations, respectively. As discussed in Section 3.1.2, the AUC scores of SDEAP may decrease when the proportion of the minor group becomes very small. At the same time, Cuffdiff 2, edgeR and DESeq2 provide consistent predictions with similar accuracies across all configurations. The results show that the conditions of samples, if available, can significantly contribute to the accuracy of a DTE analysis, especially when group sizes are highly imbalanced. When the group sizes are balanced, DTE analysis methods like SDEAP that require no prior knowledge of the conditions can do just as well as those that assume the conditions are given.

3.2 Experiments on Real Data

3.2.1 SDEAP Detects Different Cancer Subtypes

In this experiment, SDEAP and DEXUS are applied to a recently published RNA-Seq dataset including 17 individual human samples belonging to three subtypes of BC: Triple-Negative, Non-Triple-Negative and HER2-positive (Eswaran *et al.*, 2012). The RNA-Seq reads of the BC samples are aligned against the Ensembl GRCh37.62 B (hg19) reference genome using TopHat 1.4.1 with the default parameters, *i.e.*, the maximum number of mismatches allowed per aligned read is 2, the maximum insertion/deletion length is 3 and so on, as defined in (Trapnell *et al.*, 2009). The downloaded reads do not contain any adapter. All reads that are mapped to multiple locations or have the low mapping quality score (MAPQ <10) are removed by SAMtools 1.2 (Li *et al.*, 2009). With the FDR threshold 0.1, SDEAP predicts 1366 DTE genes. These genes are compared with the top-ranked 1366 predicted DTE genes by DEXUS and used to hierarchically cluster the 17 samples as illustrated in Fig. 3(a) and 3(b). The 17 samples in each dendrogram are then partitioned into three clusters and compared with the three subtypes of BC. In the clustering by SDEAP, only one of the 17 samples is misclassified while there are three misclassified samples in the clustering by DEXUS. The Jaccard index of SDEAP's clustering is 0.760, which is significantly higher than that of DEXUS (0.481). Moreover, in this BC dataset, six differentially expressed (DE) genes were validated experimentally by qPCR with fold change rates greater than 5.0. Three of the six validated DE genes are predicted by SDEAP while two are among the predicted DTE genes by DEXUS.

3.2.2 SDEAP Identifies More Validated Marker Genes Specific to Cell Types

In this experiment, an SC RNA-Seq dataset of two cell types, 12 mouse ES cells and 12 primitive endoderm (PrE) cells, is downloaded from the NCBI GEO database with accession code GSE42268. The downloaded reads do not contain any adapter and are mapped to the mouse reference genome (mm9) by TopHat 1.4.1 with the default parameters. The same data-preprocessing protocol in the Section 3.2.1 is applied to this dataset. In this dataset, there are 17 manually selected biomarker genes reported in the literature (Sasagawa *et al.*, 2013). The ranks of the 17 biomarker genes in the ranked lists of DTE genes predicted by SDEAP and DEXUS are summarized in Supplementary Tab. S3. Using the FDR threshold 0.1, SDEAP predicts 1614 DTE genes. All 17 biomarker genes are included in these predicted DTE genes. Among the 1614 top-ranked DTE genes predicted by DEXUS, only 13 of the 17 biomarkers are included. 8 of the 17 biomarkers were further validated by qPCR in (Sasagawa *et al.*, 2013). While all of these 8 validated biomarkers are in the prediction of the DTE genes by SDEAP, 2 of them are missed in the prediction by DEXUS. The better coverage of the biologically meaningful biomarkers by SDEAP suggests that it can provide a more comprehensive picture of transcript expression. On the other hand, the clustering results based on the predicted DTE genes by SDEAP and DEXUS are equally well as shown in Supplementary Fig. S5. This can perhaps be explained by the fact that the two cell types have many redundant biomarkers in the sense that even if some of them are missed, the rest are still able to separate the cell types.

3.2.3 SDEAP Is Better at Separating Cell-Cycle Phases

The SC RNA-Seq dataset of 35 samples with the accession code GSE42268 is downloaded from the NCBI GEO database where the cell-cycle phases of each cell is known *a priori*. The downloaded reads do not contain any adapter. Among the 35 samples, there are 20 cells in the Growth 1 phase (G1), 8 in the pre-mitotic/mitotic (G2/M) phase and 7 in the synthesis (S) phase. All sequenced reads of each RNA-Seq sample are aligned against the Ensembl GRCh37.62 B (mm9) reference genome using TopHat 1.4.1 with the default parameters as done in the previous subsections. With the same FDR threshold of 0.1, the 532 top-ranked genes in the predictions by SDEAP and DEXUS are used to hierarchically cluster the samples and the clustering results are illustrated in Fig 3(c) and 3(d), respectively. The cells in the dendrograms are partitioned into three clusters, as shown by the red boxes, and clusters are then compared with the three cell-cycle phases. In the clustering by SDEAP, some S cells are clustered together with G1 cells while the other S cells are with G2/M cells. This makes some sense because the S cell-cycle phase is between the G1 and G2/M phases in the cell-cycle and hence the expression profiles of some S cells are closer to those of G1 cells while the other S cells might be closer to G2/M cells. In general, the G1 cells and G2/M cells are well separated by SDEAP. However, the clustering by DEXUS fails to provide a reasonable partition consistent

with the cell-cycle phases. As a result, the Jaccard index of the DEXUS clustering (0.261) is much lower than that of the SDEAP clustering (0.391). The similarity of the 35 SC samples encoded by the expression features of the predicted DTE genes is further visualized in the 3D space by principal component analysis (PCA), as shown in Fig. 3(e) and 3(f). In the PCA transformation using the DTE genes predicted by SDEAP, although some S cells are mixed with G1 and G2/M cells, the cells of the three cell-cycle phases are still visually separable. However, in the PCA transformation using the DTE prediction of DEXUS, all cells of the three cell-cycle phases are mixed together such that the separation between the cell-cycle phases becomes more subtle. In our simulation experiments, we concluded that SDEAP is less sensitive to outliers in SC RNA-Seq data than DEXUS and is able to discover more true DTE genes that characterize the biological conditions in a population. The above clustering results on real SC RNA-Seq data support these claims. Note that since this dataset does not offer any qPCR validated DTE genes, we are unable to compare true DTE genes predicted by the methods as in the previous two experiments.

4 Conclusion

We have introduced SDEAP, an algorithm to identify DTE genes for a population of RNA-Seq samples with unknown conditions based on the splice graph data structure. SDEAP takes advantages of an accurate graph modular decomposition algorithm for discovering ASMs, efficient feature extraction for reducing the impact of technical noise, and a robust Dirichlet mixture model for inferring the groups in a population without assuming the number of biological conditions. These features make SDEAP more suitable for many practical applications. As shown in our simulation and real data experiments, the DTE features identified by SDEAP suffice to separate the subtypes of cancer, detect cell types and classify cell-cycle phases. We expect that SDEAP will serve as a useful differential expression/splicing analysis tool for RNA-Seq data in population studies with unknown biological conditions.

Acknowledgements

The research was partially supported by National Science Foundation grants DBI-1262107 and IIS-16121312.

References

Äijö, T., Butty, V., Chen, Z., Salo, V., Tripathi, S., Burge, C. B., Lahesmaa, R., and Lähdesmäki, H. (2014). Methods for time series analysis of rna-seq data with application to human th17 cell differentiation. *Bioinformatics*, **30**(12), i113–20.

Anders, S. *et al.* (2012). Detecting differential usage of exons from rna-seq data. *Genome Res.*, **22**, 2008–2017.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*, **11**(10), R106.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc.*, **B 57**, 289–300.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Bayesian Anal.*, **1**, 121–143.

Bonnal, S. *et al.* (2012). The spliceosome as a target of novel antitumour drugs. *Nat. Rev. Drug Discov.*, **11**, 847–859.

Brennecke, P. *et al.* (2013). Accounting for technical noise in single-cell rna-seq experiments. *Nat. Meth.*, **10**(11), 1093–1095.

Buettner, F. *et al.* (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotech.*, **33**(2).

Bullard *et al.* (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11**(1), 94.

Byrd, R. H. *et al.* (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**(5), 1190–1208.

Cormen, T. H. *et al.* (2001). *Intro. to Algorithms*. McGraw-Hill Higher Education, 2nd edition.

Culhane, A. C. *et al.* (2005). Made4: an r package for multivariate analysis of gene expression data. *Bioinformatics*, **21**(11), 2789–2790.

Eswaran, J. *et al.* (2012). Transcriptomic landscape of breast cancers through mRNA sequencing. *Scientific Reports*, **2**, 264.

Feng, J. *et al.* (2011). Inference of isoforms from short sequence reads. *J. Comput. Biol.*, **18(3)**, 305–321.

Fisher, R. A. (1958). *Statistical methods for research workers; 13th ed.* Oliver and Boyd, Edinburgh.

Gierlinski, M. *et al.* (2015). Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, pages 1–6.

Grau1, J., G. I. and Keilwagen, J. (2015). Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, **31(15)**, 2595–2597.

Griffith, M. *et al.* (2010). Alternative expression analysis by rna sequencing. *Nat. Meth.*, **7**(10), 843–847.

Hu, Y. *et al.* (2013). Diffsplice: the genome-wide detection of differential splicing events with rna-seq. *Nucleic Acids Res.*, **41**(2), e39.

Katz, Y. *et al.* (2010). Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat. Meth.*, **7**(12), 1009–15.

Kimes, P. K. *et al.* (2014). SigFuge: single gene clustering of RNA-seq reveals differential isoform usage among cancer samples. *Nucleic Acids Res.*, **42**(14), e113.

Klambauer, G. *et al.* (2013). Dexus: identifying differential expression in rna-seq studies with unknown conditions. *Nucleic Acids Res.*, **41**(21), e198.

Lehmann, B. D. *et al.* (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, **121**(7), 2750.

Li, H. *et al.* (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25(16)**, 2078.

Li, W. and Jiang, T. (2012). Transcriptome assembly and isoform expression level estimation from biased rna-seq reads. *Bioinformatics*, **28**(22), 2914–2921.

Love, M. *et al.* (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, **15(12)**, 550.

McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., and Nuzhdin, S. V. (2011). RNA-seq: technical variability and sampling. *BMC genomics*, **12**(1), 293.

Neal, R. M. (2007). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *J. Comp. Graph. Stat.*, **9**(2), 249–265.

Oberg, A. L. *et al.* (2012). Technical and biological variance structure in mrna-seq data: life in the real world. *BMC Genomics*, **13**(1), 304.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Robinson, M. *et al.* (2010a). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 1.

Robinson, M. *et al.* (2010b). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, **10(3)**, e0118432.

Sasagawa, Y. *et al.* (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.*, **14**(4), R31.

Shen, S. *et al.* (2012). Mats: a bayesian framework for flexible detection of differential alternative splicing from rna-seq data. *Nucleic Acids Res.*, **40**(8), e61.

Singh, D. *et al.* (2011). Fdm: a graph-based statistical method to detect differential transcription using rna-seq data. *Bioinformatics*, **27**(19), 2633–2640.

Sneath, P. (1957). Some thoughts on bacterial classification. *J. Gen. Microbiol.*, **18**, 184–200.

Tong, P. *et al.* (2013). Siber: systematic identification of bimodally expressed genes using rnaseq data. *Bioinformatics*, **29**(5), 605–613.

Trapnell *et al.* (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nat. Biotech.*, **31**(1), 46–53.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.*, pages 1491–1498.

Trapnell, C. *et al.* (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**(9), 1105–1111.

Trapnell, C. *et al.* (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.*, **28**(5), 511–515.

Yang, E. *et al.* (2013). Differential gene expression analysis using coexpression and rna-seq data. *Bioinformatics*, **29 (17)**, 2153–2161.

Zhou, X. *et al.* (2014). Robustly detecting differential expression in rna sequencing data using observation weights. *Nucleic Acids Res.*, **42**(11), e91.