Bioinformatics, 32(5), 2016, 705–712
doi: 10.1093/bioinformatics/btv657
Advance Access Publication Date: 9 November 2015
Original Paper

OXFORD

Gene expression

contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples

Qi Shen[1], Jiyuan Hu[1], Ning Jiang[1], Xiaohua Hu[1], Zewei Luo[2] and Hong Zhang[1,*]

[1]State Key Laboratory of Genetic Engineering and Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai 200433, People's Republic of China and [2]School of Biosciences, University of Birmingham, Birmingham B15 2TT, UK

[*]To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on June 11, 2015; revised on September 20, 2015; accepted on November 3, 2015

Abstract

Motivation: Accurate detection of differentially expressed genes between tumor and normal samples is a primary approach of cancer-related biomarker identification. Due to the infiltration of tumor surrounding normal cells, the expression data derived from tumor samples would always be contaminated with normal cells. Ignoring such cellular contamination would deflate the power of detecting DE genes and further confound the biological interpretation of the analysis results. For the time being, there does not exists any differential expression analysis approach for RNA-seq data in literature that can properly account for the contamination of tumor samples.

Results: Without appealing to any extra information, we develop a new method 'contamDE' based on a novel statistical model that associates RNA-seq expression levels with cell types. It is demonstrated through simulation studies that contamDE could be much more powerful than the existing methods that ignore the contamination. In the application to two cancer studies, contamDE uniquely found several potential therapy and prognostic biomarkers of prostate cancer and non-small cell lung cancer.

Availability and implementation: An R package contamDE is freely available at http://homepage.fudan.edu.cn/zhangh/softwares/.

Contact: zhanghfd@fudan.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The primary goal of the ordinary differential expression (DE) analysis methods is to distinguish biological variation between various conditions from technical effects and random noises (Anders and Huber, 2010; Robinson and Smyth, 2007; Robinson *et al.*, 2010; Zhou *et al.*, 2011). Since the clinical tumor samples obtained from patients are usually infiltrated with surrounding normal cells (de Ridder *et al.*, 2005; Liotta and Petricoin, 2000; Meyerson *et al.*, 2010; Palmer *et al.*, 2006), the traditional DE analysis cannot distinguish the expression variability resulting from the shifts in cell specific expression versus the cell proportion variation if the cellular contamination is ignored (Kuhn *et al.*, 2011). Ignoring the cellular contamination could also greatly lower the power of DE analysis (Shen-Orr *et al.*, 2010; Zhao and Simon, 2010). Several experimental methods aiming at dissecting multiple distinct cell types have been developed to eliminate the effects caused by contamination, such as laser capture microdissection and cell sorting. However, these methods are expensive, time consuming and may introduce new confounders (Okaty *et al.*, 2011). Computational approaches addressing the contamination issue could avoid such problems thus are much desired.

The following linear equation is commonly used to model the microarray expressions of contaminated tumor cells (Shen-Orr *et al.*, 2010):

$$T = wT_C + (1-w)T_N + \epsilon, \quad (1)$$

where T, T_C and T_N represent the expression profile of the contaminated tumor cells obtained by experiments and those of the underlying tumor component and normal component, respectively, w is the pure tumor cell proportion in each contaminated tumor sample shared by all genes, while ϵ is the random error.

RNA-seq data are inherently different from microarray data, and there is very limited work in literature focusing on the contamination issue for RNA-seq data, with a couple of exceptions (Gong and Szustakowski, 2013; Li and Xie, 2013). Li and Xie (2013) used an online expectation–maximization algorithm to quantify the expression levels of purified tumor transcripts in a probabilistic fashion for better understanding the abundance of the tumor transcripts. Gong and Szustakowski (2013) developed a quadratic programming method to obtain sub-cell proportions by utilizing significantly differentially expressed genes of homogeneous samples as a *prior*.

The above methods are subject to several limitations if they are used in DE analysis. First, it is hard to account for the variability of the purified expression profiles in the downstream DE analysis. Second, requiring extra information in the above methods greatly limits their applicability, since the extra information is expensive to obtain in practice. Third, additional bias could be introduced if the extra information is not accurate.

To our best knowledge, there is no statistical method in literature that can properly account for the contamination of tumor samples in DE analysis using RNA-seq data, and we aim to fill this gap. In this article, we propose a new method 'contamDE', which is based on a novel statistical model for associating gene expression profiles and cell types. Based on this model, a rigorous and efficient statistical method is developed for DE analysis using RNA-seq data from contaminated tumor samples and normal samples. In contamDE, the normalized proportions of tumor cells in contaminated tumor samples are estimated, and DE analysis can be consequently carried out.

The new method contamDE has several unique features. First, contamDE does not require any extra information that might be expensive to obtain in practice; second, the developed algorithm is computationally efficient; third, contamDE provides a unified procedure for DE analysis, which avoids the necessity of accounting for the variability of purified expression profiles required in the existing methods.

The rest of this article is organized as follows. The technical details of the new method contamDE are described in Section 2. Some simulation studies and five real data applications are provided in Section 3 and Section 4, respectively, to demonstrate the advantage of contamDE over the existing methods. Some concluding remarks and future research topics are presented in Section 5.

2 Methods

In this section, we describe the proposed method contamDE. We begin with the situation with a single tumor cell type, then extend the developed method to the situation with multiple tumor cell types.

2.1 Statistical models

First, we consider the situation where the normal samples and the tumor samples are unmatched. Suppose the read counts of G genes are obtained from I_0 normal cell samples and I tumor cell samples, respectively. Let N_{ij} be the number of the RNA-seq reads mapped to the jth gene ($j = 1, \ldots, G$) for the ith normal cell sample ($i = 1, 2, \ldots, I_0$); C_{ij} the number of the RNA-seq reads mapped to the jth ($j = 1, \ldots, G$) gene for the ith ($i = 1, 2, \ldots, I$) pure tumor cell sample. In order to distinguish the systematic expression difference between various conditions by controlling the biological and technical variations, it is often advantageous to use the two-parameter negative binomial distribution to model the read counts generated from RNA-seq experiments (Anders and Huber, 2010; Cameron and Trivedi, 2013; Robinson and Smyth, 2007; Robinson *et al.*, 2010). We assume that N_{ij} and C_{ij} follow negative binomial distributions $\mathrm{NB}(\kappa_i\mu_j, \phi_j)$ and $\mathrm{NB}(\kappa'_i\mu'_j, \phi_j)$, respectively. Here $\mathrm{NB}(\mu, \phi)$ denotes the negative binomial distribution with mean μ and dispersion parameter ϕ, whose variance is $\mu + \mu^2\phi$; κ_i and κ'_i are size factors accounting for library size effects, which can be estimated using any existing methods, e.g. the median normalization (Anders and Huber, 2010) and the quantile normalization (Bullard *et al.*, 2010). We use the median normalization because it is quite robust to outliers and simple to implement. Let $\delta_j = \mu'_j - \mu_j$ be the mean expression difference between tumors and normals. We are interested in identifying those genes with DE profiles between tumor and normal samples, i.e. those j's with $\delta_j = 0$.

In practice, the obtained tumor samples usually consist of both tumor cells and normal cells due to cellular contamination. Suppose that the ith contaminated tumor cell sample consists of a proportion of w_i tumor cells and a proportion of $1 - w_i$ normal cells. Let T_{ij} be the number of reads mapped to the jth gene of the ith contaminated tumor cell sample, which is equal to $C'_{ij} + N'_{ij}$, where C'_{ij} and N'_{ij} are the numbers of reads from pure cancer cells and pure normal cells, respectively. It is reasonable to assume that C'_{ij} and N'_{ij} follow negative binomial distributions $\mathrm{NB}(\kappa'_i(1-w_i)\mu_j, \phi_j)$ and $\mathrm{NB}(\kappa'_i w_i(\mu_j + \delta_j), \phi_j)$, respectively, where w_i and $1 - w_i$ are used to adjust the effects of cell proportions and κ'_i is a size factor, so that μ_j and $\mu_j + \delta_j$ are adjusted means. The expectation of T_{ij} is therefore $\kappa'_i(\mu_j + w_i\delta_j)$, but the distribution of T_{ij} would be rather complicated since N'_{ij} and C'_{ij} are correlated with each other. For the simplification purpose, we assume the following model:

$$N_{ij} \sim \mathrm{NB}(\kappa_i\mu_j, \phi_j) \text{ and } T_{ij} \sim \mathrm{NB}(\kappa'_i(\mu_j + w_i\delta_j), \phi_j). \quad (2)$$

Here $\mathrm{NB}(\kappa'_i(\mu_j + w_i\delta_j), \phi_j)$ is different from $(1-w_i)\mathrm{NB}(\kappa'_i\mu_j, \phi_j) + w_i\mathrm{NB}(\kappa'_i(\mu_j + \delta_j), \phi_j)$. The later distribution could be used to model the summation of two independent random variables. However, the independence assumption is violated in the current situation. It is unclear which of these two distributions fits T_{ij} better, so we use the former one because the corresponding estimation procedure is simpler.

Next, we consider the situation where each tumor sample is matched with a normal sample. The read counts N_{ij} and T_{ij} are matched if both tumor cell sample and normal cell sample are obtained from the same cancer patient, and their correlation should be accounted for. Actually, in a real prostate cancer study, we observed an evident positive correlation between the read counts from tumor sample and matched normal sample (Supplementary Material Fig. S1). Therefore, we use a fix effect to characterize such correlation. That is, we assume that N_{ij} and T_{ij} follow the negative binomial distributions with a common dispersion parameter ϕ_j and means $\kappa_i\mu_j e^{\alpha_{ij}}$ and $\kappa'_i\{(1-w_i)\mu_j + w_i(\mu_j + \delta_j)\}e^{\alpha_{ij}}$, respectively:

$$N_{ij} \sim \mathrm{NB}(\kappa_i\mu_j e^{\alpha_{ij}}, \phi_j) \text{ and } T_{ij} \sim \mathrm{NB}(\kappa'_i(\mu_j + w_i\delta_j)e^{\alpha_{ij}}, \phi_j). \quad (3)$$

The correlation between N_{ij} and T_{ij} is characterized by the shared fix effect α_{ij} as in McCarthy *et al.* (2012).

In both unmatched-sample model (2) and matched-sample model (3), the mean expression for a contaminated tumor cell sample follows a semi-additive pattern. In the absence of contamination (i.e. $w_1 = \cdots = w_I$), the two models reduce to the log-additive models commonly used in literature (Anders *et al.*, 2013). The semi-additive models (2) and (3) greatly facilitate the estimation of the proportions w_i, as will be shown in the next subsection.

Note that the proportions w_i are identifiable up to a scale parameter. We normalize the proportions such that their mean value is 1:

$$I^{-1}\sum_{i=1}^{I} w_i = 1. \tag{4}$$

Let the true proportions and mean expression differences be w_{i0} and δ_{j0}, respectively, then the values of w_i and δ_j satisfying constraint (4) are equal to

$$w_i = \frac{w_{i0}}{I^{-1}\sum_{i=1}^{I} w_{i0}} \text{ and } \delta_j = I^{-1}\sum_{i=1}^{I} w_{i0}\delta_{j0}, \tag{5}$$

respectively. As a result, for any $1 \le j \le G$, the null hypothesis of interest $H_j : \delta_{j0} = 0$ is equivalent to $H_j : \delta_j = 0$. We have the following proposition:

Proposition. Under constraint (4), the unknown parameters $\{\mu_1, \ldots, \mu_G, \delta_1, \ldots, \delta_G, \phi_1, \ldots, \phi_G, w_1, \ldots, w_I\}$ are identifiable in model (2) and $\{\mu_1, \ldots, \mu_G, \delta_1, \ldots, \delta_G, \phi_1, \ldots, \phi_G, w_1, \ldots, w_I, \alpha_{11}, \ldots, \alpha_{IG}\}$ are identifiable in model (3).

The proposition immediately follows from the fact that two distributions are the same implies that they have the same first two moments. The normal samples and the tumor samples have non-negative mean expression profiles, so μ_j and δ_j should satisfy the following additional constraints:

$$\mu_j \ge 0 \text{ and } \mu_j + w_i\delta_j \ge 0 \text{ for } i = 1, \ldots, I. \tag{6}$$

In the presence of contamination, the true proportions are smaller than 1. Therefore, the normalized proportion w_i given in (5) would be greater than w_{i0}, and some of them could be greater than 1. Similarly, the absolute value of δ_j is smaller than that of δ_{i0}. Therefore, under constraint (4), the tumor-versus-normal fold change defined by

$$\mathrm{FC}_j = \delta_j/\mu_j + 1 \tag{7}$$

is shrunk toward 1, compared with the true fold change. This is a tradeoff of not using extra information. The extent of shrinkage is decreasing in the magnitude of w_{i0}. In particular, FC_j is equal to the true fold change in the absence of contamination (i.e. $w_{10} = \cdots = w_{I0} = 1$).

Note that the likelihood ratio test we will use does not depend on the form of linear constraint since the likelihood function is invariant to the linear constraint.

2.2 Normalized proportion estimation

In this subsection, based on models (2) and (3), we construct appropriate estimators of w_i subject to constraint (4) in the matched sample situation and the unmatched sample situation, respectively, without appealing to extra information. In what follows, we propose an initial estimator of w_i, then update it in an iterative manner. The initial estimator could be much closer to the true one compared with the naive initial estimator (e.g. 1), which effectively save computational time, as confirmed in our simulation studies.

First, we consider the matched sample situation. In what follows, we construct an initial estimator of w_i and update the estimator based on a likelihood method.

Under model (3), the means of $\kappa_i^{-1}N_{ij}$ and $(\kappa_i')^{-1}T_{ij}$ are $\mu_j e^{\alpha_{ij}}$ and $(\mu_j + w_i\delta_j)e^{\alpha_{ij}}$, respectively. Therefore, we can use the following moment estimating equation:

$$\sum_{j\in\Omega_1}\left(\frac{T_{ij}}{\kappa_i'} - \frac{N_{ij}}{\kappa_i}\right) = w_i\sum_{j\in\Omega_1}\delta_j e^{\alpha_{ij}}. \tag{8}$$

Here Ω_1 is a set of DE genes ('working gene set' hereafter) used to remove the impact of non-DE genes since they are not informative in estimating w_i and can introduce a tremendous noise. We can include either up-regulated genes or down-regulated genes, but it is not appropriate to include both types of genes as the information for estimating w_i would be greatly obscured by doing so.

If $\sum_{j\in\Omega_1}\delta_j e^{\alpha_{ij}}$ is free of i, then it will not appear in the normalized w_i estimator derived from (8). This motivates an initial estimator of w_i given by

$$\hat{w}_i^{(1)} = \frac{I\sum_{j\in\Omega_1}(T_{ij}/\kappa_i' - N_{ij}/\kappa_i)}{\sum_{i=1}^{I}\sum_{j\in\Omega_1}(T_{ij}/\kappa_i' - N_{ij}/\kappa_i)}. \tag{9}$$

Obviously, $(\hat{w}_1^{(1)}, \ldots, \hat{w}_I^{(1)})$ satisfy constraint (4). In practice, we can choose the working gene set Ω_1 using any DE analysis tool such as edgeR by ignoring the contamination of the tumor samples. Specifically, we select those genes with DE analysis P-values smaller than a small threshold value (say 10^{-3}) and a negatively estimated log-fold change. Such a gene set results in virtually sound proportion estimators in both simulation studies and real data applications. We will give more details for obtaining such P-values in the next section.

Now we update the proportion estimators. Fixing w_i at $\hat{w}_i^{(1)}$ for $i = 1, \ldots, I$, we can obtain the pseudo-maximum likelihood estimators $(\hat{\mu}_j^{(1)}, \hat{\delta}_j^{(1)}, \hat{\alpha}_{1j}^{(1)}, \ldots, \hat{\alpha}_{Ij}^{(1)})$ of $(\mu_j, \delta_j, \alpha_{1j}, \ldots, \alpha_{Ij})$ using an algorithm described in the next subsection. Using the P-values of DE analysis based on the pseudo likelihood ratio tests described in the next subsection, an updated working gene set Ω_2 can be obtained based on the previous criterion (DE analysis P-value $< 10^{-3}$ and log-fold change < 0). More accurate estimators of w_i can be iteratively given by

$$\hat{w}_i^{(k)} = \frac{I\tilde{w}_i}{\sum_{i'=1}^{I}\tilde{w}_{i'}} \text{ with } \tilde{w}_i = \frac{\sum_{j\in\Omega_k}\{T_{ij}/\kappa_i' - N_{ij}/\kappa_i\}}{I^{-1}\sum_{i=1}^{I}\sum_{j\in\Omega_k}\hat{\delta}_j^{(k-1)}e^{\hat{\alpha}_{ij}^{(k-1)}}} \tag{10}$$

for $k = 2, 3, \ldots$ The iteration can be stopped if $\max_i|\hat{w}_i^{(k+1)} - \hat{w}_i^{(k)}| < \epsilon$ for some small ϵ. Denote by $\hat{w}_i^{(\infty)}$ the converged estimator of w_i.

Next, we consider the unmatched sample situation. Under model (2), we have the estimating equation

$$\sum_{j\in\Omega_1}\left(\frac{T_{ij}}{\kappa_i'} - \frac{1}{I_0}\sum_{i=1}^{I_0}\frac{N_{ij}}{\kappa_i}\right) = w_i\sum_{j\in\Omega_1}\delta_j \tag{11}$$

for some working gene set Ω_1. This motivates proportion estimators

$$\hat{w}_i^{(k)} = \frac{I\sum_{j\in\Omega_k}(T_{ij}/\kappa_i' - I_0^{-1}\sum_{i=1}^{I_0}N_{ij}/\kappa_i)}{\sum_{i=1}^{I}\sum_{j\in\Omega_k}(T_{ij}/\kappa_i' - I_0^{-1}\sum_{i=1}^{I_0}N_{ij}/\kappa_i)} \tag{12}$$

for $k = 1, 2, \ldots$ where $\Omega_1, \Omega_2, \ldots$ can be constructed similarly to the matched sample situation. Note that only the working gene set is updated in $\hat{w}_i^{(k)}$ $(k = 2, 3, \ldots)$ since $\sum_{j\in\Omega}\delta_j$ is free of i.

2.3 DE analysis procedure

In this subsection, we give an outline for estimating unknown parameters $(\phi_j, \mu_j, \delta_j)$ and $\boldsymbol{\alpha}_j := (\alpha_{1j}, \ldots, \alpha_{Ij})$, $j = 1, \ldots, G$, which is followed by a DE gene detection procedure. Note that $\boldsymbol{\alpha}_j$ needs to be estimated only in the matched sample situation. The detailed algorithm for estimating unknown parameters are given in Supplementary Material Methods.

First, we estimate the dispersion parameters ϕ_j and select a working gene set Ω_1. Since ϕ_j are nuisance parameters, for simplicity and other parsimonious reasons, we use an empirical Bayes strategy (McCarthy *et al.*, 2012) to estimate the dispersion parameters ϕ_j by ignoring the contamination of the tumor samples. Let $\hat{\phi}_j$ denote the resulting estimator of ϕ_j. DE analysis is then carried out using the likelihood ratio test, which results in a P-value for each gene. A working gene set Ω_1 can then be selected (P-value $< 10^{-3}$ and log-fold change < 0).

Next, we estimate μ_j, δ_j and $\boldsymbol{\alpha}_j$ with ϕ_j fixed at $\hat{\phi}_j$ and $\boldsymbol{w} := (w_1, \ldots, w_I)$ fixed at $\boldsymbol{w}\,(k) = (\hat{w}_1^{(k)}, \ldots, \hat{w}_I^{(k)})$ for $k = 1, 2, \ldots$, by maximizing a log-pseudo likelihood function. Denote by $(\hat{\mu}_j^{(k)}, \hat{\delta}_j^{(k)})$ and $(\hat{\mu}_j^{(k)}, \hat{\delta}_j^{(k)}, \hat{\boldsymbol{\alpha}}_j(k))$ the resulting pseudo-MLEs in the unmatched sample situation and matched sample situation, respectively. Denote by $(\tilde{\mu}_j^{(k)}, 0)$ and $(\tilde{\mu}_j^{(k)}, 0, \tilde{\boldsymbol{\alpha}}_j^{(k)})$ the corresponding pseudo-MLEs under the null hypothesis $H_j : \delta_j = 0$.

The fold change (7) can be estimated by

$$\mathrm{FC}_j^{(k)} = \hat{\delta}_j^{(k)} / \hat{\mu}_j^{(k)} + 1. \tag{13}$$

The pseudo-likelihood ratio test statistics for testing null hypotheses $H_j : \delta_j = 0$ can be constructed immediately. The asymptotic null distributions of the test statistics are the chi-square distribution on 1 degree of freedom, so the P-values for testing $H_j : \delta_j = 0$ easily follow. The algorithm for estimating all parameters is outlined in a flowchart (Supplementary Material Fig. S2).

2.4 Multiple tumor cell types

Some cancers have multiple tumor cell types. For example, some cancers have multiple subtypes (e.g. the main primary types of lung cancer include small-cell lung carcinoma and non-small-cell lung carcinoma), and some cancers undergo several stages with various tumor cell types. In Supplementary Material Methods, we describe a model and an outline of estimation and testing procedure for multiple tumor cell types.

2.5 Existing methods

DeMix (Ahn *et al.*, 2013) and UNDO (Wang *et al.*, 2015) are two methods designed for the deconvolution of array-based data without using extra information. Our preliminary numerical study indicated DeMix might not be applicable to RNA-seq data, so we only considered UNDO in the following numerical studies. We also considered a couple of benchmark DE analysis methods, which ignored the contamination of tumor samples and had been implemented in the R Bioconductor packages edgeR and DESeq2, respectively. Refer to Supplementary Material Methods for the descriptions of DeMix, UNDO, edgeR and DESeq2.

3 Simulation studies

We conducted simulations to evaluate the performance of contamDE. The data generation procedure and the corresponding simulation results in a two-condition situation (one tumor cell type plus one normal cell type) are described in the next three subsections. Some simulation results for multi-condition situation (two tumor cell types plus normal cell type) are reported in Supplementary Material Results and Supplementary Material Figures S3–S5.

3.1 Data generation

We generated genewise read counts ($G = 10,000$) for six pure normal samples and six contaminated tumor samples. In the unmatched sample situation, the read counts N_{ij} and C_{ij} for pure normal and tumor samples were generated from $\mathrm{NB}(\kappa_i\mu_j, \phi_j)$ and $\mathrm{NB}(\kappa_i'(\mu_j + \delta_j))$, respectively; in the matched sample situation, the read counts N_{ij} and C_{ij} for pure normal and tumor samples were generated from $\mathrm{NB}(\kappa_i\mu_j e^{\alpha_{ij}}, \phi_j)$ and $\mathrm{NB}(\kappa_i'(\mu_j + \delta_j)e^{\alpha_{ij}}, \phi_j)$, respectively. In both situations, the read count T_{ij} for contaminated tumor sample was $w_iC_{ij} + (1 - w_i)N_{ij}'$, where $N_{ji}' \sim \mathrm{NB}(\kappa_i\mu_j, \phi_j)$ in the unmatched sample situation and $N_{ji}{}' \sim \mathrm{NB}(\kappa_i'\mu_j e^{\alpha_{ij}}, \phi_j)$ in the matched sample situation, and N_{ij}, N_{ij}' and C_{ij} were independent of each other. To evaluate the performance of UNDO, we considered two scenarios, one used (N_{ij}, T_{ij}) as in contamDE; the other used (N_{ij}', T_{ij}). The proportions w_i were either homogeneous (i.e. $w_{10} = \cdots = w_{60} = 0.5$) or heterogeneous (i.e. $w_{i0} = i/6$). The description of the other parameters are given in Supplementary Material Methods. For each parameter combination, 100 datasets were generated.

3.2 Proportion estimation results

We evaluated three w_i estimators, namely $\hat{w}_i^{(1)}$, $\hat{w}_i^{(2)}$ and $\hat{w}_i^{(\infty)}$ described in the previous section. To obtain $\hat{w}_i^{(\infty)}$, we used the convergence criterion $\max_i|w_i^{(k+1)} - w_i^{(k)}| < 10^{-4}$. For comparison purpose, we rescaled the w_i estimators by 3.5/6 in the heterogeneous proportion situation and 0.5 in the homogeneous proportion situation, respectively.

In the heterogeneous proportion situation (i.e. $w_{i0} = i/6$ for $i = 1, \ldots, 6$), the boxplots of the proportion estimates for 100 simulated datasets are depicted in Figure 1 and Supplementary Material Figure S6. The initial estimator $\hat{w}_i^{(1)}$ was much closer to the true one than the naive initial estimator 1, so using $\hat{w}_i^{(1)}$ could effectively save computational time as claimed in Section 2.2. For example, in the unmatched sample situation, the average number of iterations with initial estimator $\hat{w}^{(1)}$ was around 2.7, and that with naive initial estimator 1 was 3.9. As shown in Supplementary Material Figure S7, the w_i estimation results did not depend on initial 'working gene set'. Furthermore, $\hat{w}_i^{(2)}$ and $\hat{w}_i^{(\infty)}$ were very close to each other, and their biases were much smaller than those of $\hat{w}_i^{(1)}$. For example, in the unmatched sample situation, the mean square error rates of $\hat{w}_i^{(1)}$'s, $\hat{w}_i^{(2)}$'s and $\hat{w}_i^{(\infty)}$'s were 1.0×10^{-3}, 0.7×10^{-3} and 0.7×10^{-3}, respectively. Therefore, we propose to use $\hat{w}_i^{(2)}$ instead of $\hat{w}_i^{(\infty)}$ for the sake of saving computational time. The saved number of iteration was particularly considerable in the matched sample situation, with an average value being 7.4.

UNDO performed well only when the two samples used for deconvolution were perfectly matched (i.e. N_{ij}' and T_{ij} were used), while its proportion estimates were very close to 1 when N_{ij} and T_{ij} were used. Noted that the perfect matching scenario was very unrealistic since N_{ij}' cannot be obtained in practice.

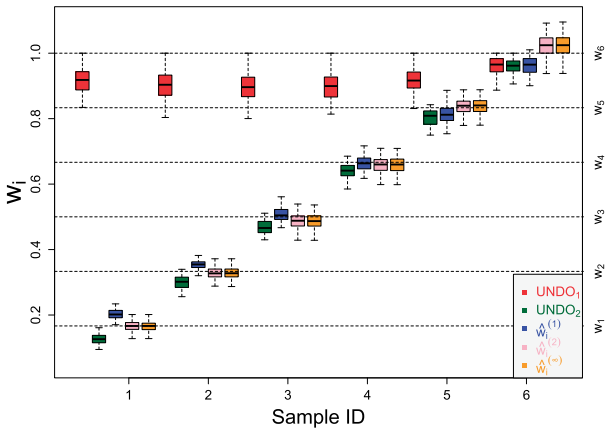

Fig. 1. Estimated proportions for contaminated tumor samples in the simulation study with two conditions and heterogeneous proportions (unmatched samples). $UNDO_1$ used (N_{ij}, T_{ij}) and $UNDO_2$ used (N'_{ij}, T_{ij})

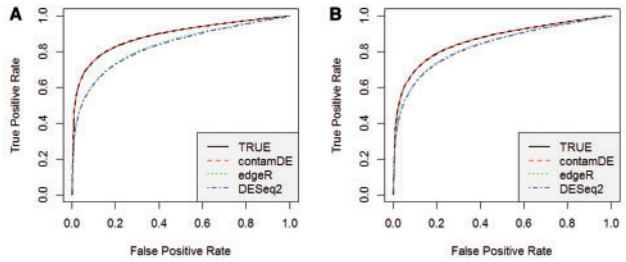

Fig. 2. ROC curves for detecting DE genes in the simulation study with two conditions and heterogenous proportions. (**A**) Unmatched samples; (**B**) matched samples

In the homogeneous proportion situation (i.e. $w_{i0} = 0.5$), contamDE was quite unbiased (Supplementary Material Fig. S8), while UNDO produced similar results as above (results not shown).

3.3 DE analysis results

In contamDE, we used $\hat{w}_i^{(2)}$ according to the previous simulation results. For comparison purpose, we also included a version of contamDE using true proportions (TRUE). The ability of DE analysis of TRUE, contamDE, edgeR and DESeq2 were evaluated through receive operating characteristic (ROC) curves, false positive rates (FPRs), true positive rates (TPRs) and false discovery rates (FDRs). We begin with the heterogeneous proportion situation.

For each of 100 simulated datasets, the DE analysis *P*-values of the DE genes (either up-regulated or down-regulated) and equally expressed genes were used to produce ROC curves. In the heterogeneity proportion situation, the average ROC curves based on 100 simulated datasets are displayed in Figure 2. Overall, TRUE and contamDE performed quite comparably, and they distinctly outperformed edgeR and DESeq2.

Next, we present the FPRs and TPRs (nominal level = 0.05) in Figure 3. Among the four methods, TRUE and contamDE had slightly lower FPRs and they performed comparably with each other. Also, TRUE and contamDE were more powerful than edgeR and DESeq2, with power gains varying from 0.08 to 0.13. Overall, TRUE and contamDE performed comparably, and they outperformed edgeR and DESeq2 in terms of ROC curves and TPRs. This demonstrates that including extra accurate information on proportions has little power improvement in DE analysis.

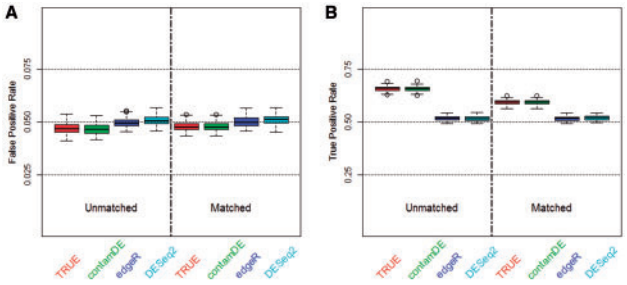

Fig. 3. False positive rates (**A**) and true positive rates (**B**) for detecting DE genes in the simulation study with two conditions and heterogenous proportions (nominal level = 0.05)

Finally, we compare the FDRs of the four methods. Among top N ($N = 200, 400, \ldots, 2000$) genes with smallest DE analysis *P*-values, the average empirical FDRs for detecting DE genes based on 100 simulated datasets are reported in Supplementary Material Figure S9. Evidently, TRUE and contamDE had much lower FDRs than edgeR and DESeq2. For each dataset, using the R function 'p.adjust', we also calculated the FDR-adjusted *P*-values (Benjamini and Hochberg, 1995), then obtained the empirical FDRs at various nominal levels. The average empirical FDRs are again reported in Supplementary Material Figure S9. TRUE and contamDE had much better control of FDRs than edgeR and DESeq, i.e, the empirical FDRs of TRUE and contamDE were much closer to the nominal levels.

On the other hand, the four methods performed comparably in the homogeneous proportion situation (Supplementary Material Figs S10–S12).

4 Real data applications

Five publicly available datasets were used to evaluate the performance of contamDE. The first one had both RNA-seq data and whole-exome sequence data. The second one had experimental mixtures and the third one had numerical mixtures. These three studies were mainly used to verify the performance of proportion estimation. The forth and fifth ones were used to evaluate the DE analysis performance. In these studies, a gene was claimed to be differentially expressed if the DE *P*-value after Bonferroni correction was smaller than 0.05.

4.1 Proportion estimation performance

First, a non-small cell lung cancer (NSCLC) study (Ju *et al.*, 2012) with both RNA-seq data and whole-exome sequence (WES) data were used to validate the proportion estimation performance. In this study, RNAs of tumor tissue from each of 165 NSCLC patients were sequenced on an Illumina Hiseq 2000 platform. WES data for the tumor tissues were available for two patients. We downloaded the RNA-seq data and the corresponding WES data for these two patients, and the RNA-seq data for three randomly selected normal samples (NCBI website http://www.ncbi.nlm.nih.gov, accession IDs: PRJEB3132). After some data processing, we obtained the RNA-seq read counts of 12 382 genes for each of the five samples (Supplementary Material Results). The DNA-seq reads from the WES data were aligned to human reference genome hg19 (UCSC release) using Bowtie2 (Langmead and Salzberg, 2012), and a java based SNP caller 'Virmid' (Kim *et al.*, 2013) designed for matched samples was used to identify contamination levels for these two samples, which could serve as golden standards. By (5), the true

proportions w_{i0}'s can be estimated by contamDE with some additional information (e.g. the average proportion). In detail, we estimated the mean proportion (i.e. $(w_{10}+w_{20})/2$) of the two samples using Virmid, then rescaled the normalized proportions estimated by contamDE to obtain the estimates of w_{0i}. As shown in Figure 4A, the resultant estimates by Virmid and contamDE were quite close to each other, which well verified the validity of contamDE. On the other hand, the proportion estimates produced by UNDO were close to 1, which coincided with UNDO_1 for the simulation data (Fig. 1).

Next, we applied contamDE to experimental mixtures of two lung adenocarcinoma cell lines (i.e. NCI-H1975 and HCC827) (Liu *et al.*, 2015). In this experimental mixture study, the sequence data were obtained for three NCI-H1975 samples, three HCC827 samples, and three biological replications of mixtrue samples at three mixture levels (NCI-H1975 versus HCC827 ratios were 1:3, 1:1 and 3:1, respectively). Therefore, the true proportions of NCI-H1975 in the three mixture samples were 0.25, 0.5, 0.75, respectively. We downloaded the summarized read count data for all the 15 samples from http://www.ncbi.nlm.nih.gov/geo (accession ID:GSE64098). The read counts for 23 669 genes were obtained. Using the pure samples of NCI-H1975 and HCC827, the proportions estimated by contamDE and UNDO were very both closed to 1 (Fig. 4B). We selected 12 combinations, each combination consisted of three pure samples and three matched mixture samples (Supplementary Materials Table S1). Since the average proportion of the mixture samples in each combination was 0.5, we rescaled the normalized proportion estimates of contamDE by multiplying 0.5. The mean square errors of the estimated proportions were 0.007 and 0.21 for contamDE and UNDO, respectively (Fig. 4C), showing that contamDE had a much higher precision than UNDO. Altogether, contamDE identified 4797 DE genes, compared with 4767 and 4087 by edgeR and DESeq2, respectively (Supplementary Material Fig. S13A). Since there was no ground truth, we treated the 3914 genes identified by all of the three methods as 'true' DE genes as suggested by Zhang *et al.* (2014). Overall, contamDE obtained many more 'true' DE genes than edgeR and DESeq2 (Supplementary Material Fig. S14).

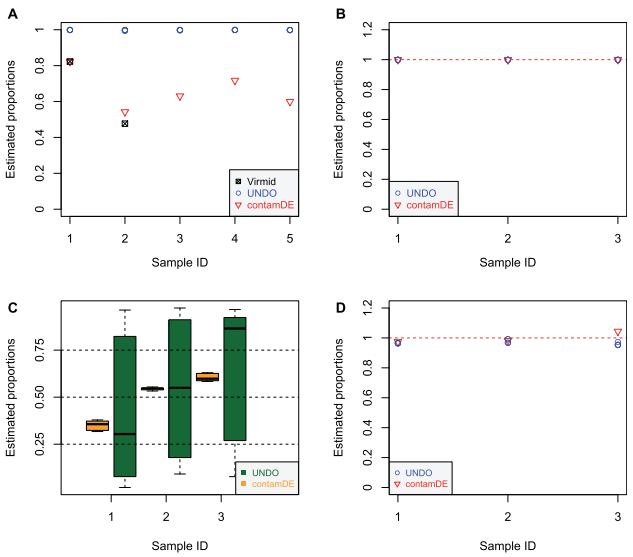

Fig. 4. Estimated proportions for contaminated tumor samples in real data applications. (**A**) A NSCLC study with both DNA-seq and RNA-seq data; (**B**) a study of lung adenocarcinoma cell lines (pure samples); (**C**) a study of lung adenocarcinoma cell lines (experimental mixture samples); (**D**) a *Drosophila melanogaster* study

Finally, we applied contamDE to a Drosophila melanogaster study (Brooks *et al.*, 2011). We downloaded the read count data summarized by Anders *et al.* (2013). In this study, there were seven RNA-seq samples of *Drosophila melanogaster* S2 cells, of which three samples were treated with siRNA targeting the splicing factor *pasilla* (CG1844) (referred to as cell type 'A') and four samples were untreated (referred to as cell type 'B'). The read counts for 7196 genes were obtained. As shown in Figure 4D, the estimated w_i's of contamDE were very close to 1. Since the original type 'A' cell samples were pure, the true proportions should be 1. These proportion estimates again demonstrated a pretty good performance of contamDE. The numbers of DE genes identified by contamDE, edgeR and DESeq2 were 187, 188 and 265, respectively (Supplementary Material Fig. S13B). The number of DE genes identified by both contamDE and edgeR was 184, showing a pretty high concordance between the results of contamDE and edgeR. Again, we treated the 169 genes identified by all the three methods as 'true' DE genes, and used artificial counts for contaminated tumor samples to validate the performance of the considered methods through the identified 'true' DE genes (Supplementary Material Results). With artificial count data, contamDE was found to be more accurate than UNDO in estimating contamination proportions, and contamDE identified many more 'true' DE genes than edgeR and DESeq2 (Supplementary Material Figs S15 and S16).

4.2 DE analysis performance

The prostate cancer data were downloaded from the EMBL-EBI website http://www.ebi.ac.uk/ (accession ID: E-MTAB-567). In this study, a prostate cancer cell sample and adjacent normal cell sample were provided by each of 14 patients from Shanghai Changhai Hospital (Ren *et al.*, 2012). After some data preprocessing, we obtained read counts for 12 698 genes (Supplementary Material Results).

Using contamDE, the maximal w_i estimate was 1.38 while the minimal one was 0.54, revealing an impressive heterogeneity of proportions across the contaminated tumor samples (Supplementary Material Fig. S17A). To verify the proportion estimation accuracy and the DE analysis power of contamDE, we adopted the same strategy used in *Drosophila melanogaster* study to generate artificial read counts for contaminated tumor samples, refer to Supplementary Material Results for details. Again, the proportion estimates were accurate enough and contamDE was more powerful than edgeR and DESeq2 (Supplementary Material Fig. S18).

The proposed method contamDE identified most DE genes, with a number of 1031, compared with 948 and 1022 by edgeR and DESeq2, respectively. Altogether, 810 genes were commonly identified by all the three methods, while 85 uniquely by contamDE, 15 uniquely by edgeR and 129 uniquely by DESeq2. See Supplementary Material Figure S17B for more details.

We further investigated the identified DE genes through database mining and functional analysis. Several genes uniquely identified by contamDE were found to be closely related to prostate cancers, i.e. *PDLIM4* and *RASL11A* were explored to be candidate biomarkers for prostate cancers. The uniquely identified GO terms enriched by contamDE were associated with gene functions of epithelial cell development, membrane and cellular proliferation that played important roles in tumor progression. Refer to Supplementary Results for more information.

In our NSCLC study, six NSCLC patients from Shanghai Changhai Hospital provided tumor samples and adjacent normal samples. The RNA samples were extracted and sequenced using Illumina Hiseq 2000 platform. Refer to Zhang *et al.* (2015) for details of this study. After the same data preprocessing as the prostate cancer study, we obtained the read counts for 13 312 genes for each of the 12 samples.

The normalized proportions w_i for tumor cell samples estimated by contamDE (Supplementary Material Fig. S19A) ranged from 0.57 to 1.47, indicating a considerable heterogeneity of proportions. As in the prostate cancer study, artificial read counts were generated to examine the accuracy of proportion estimates and DE detection power of contamDE, refer to Supplementary Material Results for detailed description. Again, contamDE was shown to be more powerful than edgeR and DESeq2 in detecting DE genes (Supplementary Material Fig. S20).

The numbers of the DE genes detected by contamDE, edgeR and DESeq2, respectively, are shown in Supplementary Material Figure S19B. Again, contamDE identified most DE genes, with a number of 477, compared with 454 and 405 by edgeR and DESeq2, respectively. Altogether 307 genes were commonly identified by all the three methods, while 75 uniquely by contamDE, 23 uniquely by edgeR and 53 uniquely by DESeq2. In these 75 genes, three were demonstrated in literature to be potential therapeutic drug targets in vivo for NSCLC. Moreover, three GO terms uniquely detected by contamDE were related to the gene functions of cytoskeletal part, cell-substrate junction and cytoskeleton, which had close relationship to tumor cell migration and invasion in non-small cell lung carcinomas. Refer to Supplementary Material Results for more information.

5 Discussion

In this article, we present a rigorous and efficient DE detection method contamDE using RNA-seq data from contaminated tumor samples, where the tumor samples could be either matched or unmatched with normal samples. Unlike most existing methods, contamDE does not require extra information. The normalized proportions reflect relative contamination intensities of the contaminated tumor samples, and the true proportions can be estimated if some pathology information or DNA sample is available. The proposed normalized proportion estimator was evaluated through both simulation studies and experimental/numerical mixture data. Through both simulation studies and real data applications, UNDO (an existing deconvolution algorithm originally designed for array-based data without using additional information) performed well only when the samples were pure (Fig. 4B and D) or the two samples used for deconvolution were perfectly matched ($UNDO_2$ in Fig. 1). In more practical situations, UNDO either greatly overestimated the proportions ($UNDO_1$ in Figs 1 and 4A) or had a much larger variance than contamDE (Fig. 4C).

In the presence of contamination, contamDE were shown through simulations to greatly outperformed the benchmarks edgeR and DESeq2 that ignored the contamination in terms of both powers and false discovery rates; in the absence of contamination, contamDE performed comparably with edgeR and DESeq2. In the application to a prostate cancer study and a lung cancer study, contamDE returned more biologically meaningful DE genes that were closely related to cancers, compared with edgeR and DESeq2.

An R package implementing contamDE can be downloaded from http://homepage.fudan.edu.cn/zhangh/softwares/, which can be used to estimate the normalized contamination proportions of contaminated tumor samples, to conduct pseudo-likelihood ratio tests for detecting DE genes between tumor samples and normal samples, and to estimate fold changes. Using a desktop computer with a 3.20 GHz CPU, it took contamDE no more than 10 minutes to analyze all considered real datasets, revealing that the computational burden of contamDE was quite acceptable.

Acknowledgements

We are grateful to three anonymous reviewers for their insightful comments.

Funding

This work was supported by the State Key Development Program for Basic Research of China (grant number 2012CB316505) and the National Natural Science Foundation of China (grant number 11371101).

Conflict of Interest: none declared.

References

Ahn,J. *et al.* (2013) DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, **29**, 1865–1871.

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

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using r and bioconductor. *Nat. Protoc.*, **8**, 1765–1786.

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

Brooks,A.N. *et al.* (2011) Conservation of an RNA regulatory map between drosophila and mammals. *Genome Res.*, **21**, 193–202.

Bullard,J.H. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94, doi:10.1186/1471-2105-11-94.

Cameron,A.C. and Trivedi,P. (2013) *Regression Analysis of Count Data*. Vol. 53. Cambridge University Press, New York.

de Ridder,D. *et al.* (2005) Purity for clarity: the need for purification of tumor cells in DNA microarray studies. *Leukemia*, **19**, 618–627.

Gong,T. and Szustakowski,J.D. (2013) DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, **29**, 1083–1085.

Ju,Y.S. *et al.* (2012) A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. *Genome Res.*, **22**, 436–445.

Kim,S. *et al.* (2013) Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome Biol.*, **14**, R90.

Kuhn,A. *et al.* (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods*, **8**, 945–947.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Li,Y. and Xie,X. (2013) A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics*, **14**, S11.

Liotta,L. and Petricoin,E. (2000) Molecular profiling of human cancer. *Nat. Rev. Genet.*, **1**, 48–56.

Liu,R. *et al.* (2015) Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res.*, gkv412.

McCarthy,D. *et al.* (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

Meyerson,M. *et al.* (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, **11**, 685–696.

Okaty,B.W. *et al.* (2011) A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. *PLoS One*, **6**, e16493.

Palmer,C. *et al*. (2006) Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics*, **7**, 115.

Ren,S. *et al*. (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.*, **22**, 806–821.

Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Shen-Orr,S.S. *et al*. (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

Wang,N. *et al*. (2015) UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics*, **31**, 137–139.

Zhang,H. *et al*. (2015) PLNseq: a multivariate poisson lognormal distribution for high-throughput matched rna-sequencing read count data. *Stat. Med.*, **34**, 1577–1589.

Zhang,Z.H. *et al*. (2014) A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. *PLoS ONE*, **9**: e103207. doi:10.1371/journal.pone.0103207.

Zhao,Y. and Simon,R. (2010) Gene expression deconvolution in clinical samples. *Genome Med.*, **2**, 93–93.

Zhou,Y.-H. *et al*. (2011) A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, **27**, 2672–2678.