frontiers in Genetics

ORIGINAL RESEARCH
published: 28 August 2018
doi: 10.3389/fgene.2018.00321

DECtp: Calling Differential Gene Expression Between Cancer and Normal Samples by Integrating Tumor Purity Information

Weiwei Zhang[1], *Haixia Long*[2], *Binsheng He*[3]* *and Jialiang Yang*[4,5]*

[1] School of Science, East China University of Technology, Nanchang, China, [2] Department of Information Science and Technology, Hainan Normal University, Haikou, China, [3] The First Affiliated Hosptial, Changsha Medical University, Changsha, China, [4] College of Information Engineering, Changsha Medical University, Changsha, China, [5] Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Identifying differentially expressed genes (DEGs) between tumor and normal samples is critical for studying tumorigenesis, and has been routinely applied to identify diagnostic, prognostic, and therapeutic biomarkers for many cancers. It is well-known that solid tumor tissue samples obtained from clinical settings are always mixtures of cancer and normal cells. However, the tumor purity information is more or less ignored in traditional differential expression analyses, which might decrease the power of differential gene identification or even bias the results. In this paper, we have developed a novel differential gene calling method called DECtp by integrating tumor purity information into a generalized least square procedure, followed by the Wald test. We compared DECtp with popular methods like *t*-test and limma on nine simulation datasets with different sample sizes and noise levels. DECtp achieved the highest area under curves (AUCs) for all the comparisons, suggesting that cancer purity information is critical for DEG calling between tumor and normal samples. In addition, we applied DECtp into cancer and normal samples of 14 tumor types collected from The Cancer Genome Atlas (TCGA) and compared the DEGs with those called by limma. As a result, DECtp achieved more sensitive, consistent, and biologically meaningful results and identified a few novel DEGs for further experimental validation.

Keywords: differentially expressed genes, tumor purity, generalized least square, the Wald test, generalized least square

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Yuannyu Zhang,
The University of Texas at Dallas, United States
Minxian Wang,
Broad Institute, United States
Cheng Guo,
Columbia University, United States

***Correspondence:**
Binsheng He
hbscsmu@163.com
Jialiang Yang
jialiang.yang@mssm.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *14 June 2018*
Accepted: *30 July 2018*
Published: *28 August 2018*

Citation:
Zhang W, Long H, He B and Yang J (2018) DECtp: Calling Differential Gene Expression Between Cancer and Normal Samples by Integrating Tumor Purity Information. Front. Genet. 9:321. doi: 10.3389/fgene.2018.00321

INTRODUCTION

Nowadays, RNA sequencing (RNA-Seq) has become a routine for measuring RNA expression levels (Mortazavi et al., 2008; Wang et al., 2009). Due to continuous improvements on sequencing accuracy and reduction on costs, this technology has revolutionized most fields in life sciences especially clinical medicine (Berger et al., 2010). Among many goals of RNA-Seq study, identifying differentially expressed genes (DEGs) between usually two conditions is probably the most common (Ritchie et al., 2015). Generally speaking, DEG analysis performs statistical analysis to discover significant gene expression changes between the experimental and control groups, which are critical for explaining transcriptomic changes incurred by experimental conditions. For instance, DEGs between normal and tumor samples help to study tumorigenesis, and have been routinely applied to identify diagnostic, prognostic, and therapeutic biomarkers for many cancers (Wu et al., 2013).

Over the past years, a number of statistical methods and softwares have been developed for identifying DEGs by considering the distributions of gene transcript abundance measured by read counts, Fragments Per Kilobase of transcript per Million (FPKM) (Trapnell et al., 2012), RNA-Seq by Expectation Maximization (RSEM) (Li and Dewey, 2011), and so on. Gene read counts usually follow a multinomial distribution, which can be approximated by a Poisson distribution, if they are independently sampled from a population with fixed fractions of genes. Consequently, the Poisson distribution has been widely assumed to test for differential expressions (Marioni et al., 2008; Wang et al., 2010). However, there is only one single parameter in the Poisson distribution, so the resulting statistical test does not control for the type-I error (Robinson and Smyth, 2007). To solve this so-called over-dispersion problem, the negative binomial (NB) distribution has been proposed to model count data (Anders and Huber, 2010; Zhou et al., 2011; McCarthy et al., 2012; Wu et al., 2013). Alternatively, the read counts can be converted to log2 transformed counts per million, for which the Bayes moderated Student's *t*-test and linear modeling methods like limma can be used. For instance, limma used a linear model to assess differential expression from microarray or RNA-Seq technologies by using multifactor designed experiments. It has a few advantages include stable on even small sample sizes and good in complex experiments with a variety of experimental conditions and predictors (Ritchie et al., 2015).

Differential expression analyses have been widely performed in cancer (Liang and Pardee, 2003). It is known that clinical tumor samples contain not only tumor cells but also tumor-associated normal epithelial and stromal cells, immune cells, and vascular cells (Joyce and Pollard, 2009), which play important roles in tumor growth, disease progression, and drug resistance (Hanahan and Weinberg, 2011; Junttila and de Sauvage, 2013). As a result, tumor purity, i.e., the percentages of cancer cells in solid tumor samples, is critical in genomic, transcriptomic, and methylation analyses in cancer (Aran et al., 2015; Zheng et al., 2017). For example, we recently developed InfiniumPurify by integrating tumor purity into differential methylation (DM) analysis, which significantly improved the accuracy of the DM identification (Zheng et al., 2017). In addition, we developed a rigorous statistical method InfiniumClust to perform sample clustering on DNA methylation data using tumor purity, which also exhibited superior accuracy (Zhang et al., 2017). There are also a few attempts to account for tumor purity in differential expression analysis (Wang et al., 2015; Shen et al., 2016) by adding it as an additive or semi-additive covariate in linear models (Aran et al., 2015). For example, contamDE proposed a few statistical models to call differential genes between unmatched or matched normal and tumor samples, in which the mean expression for a "contaminated" tumor cell sample follows a semi-additive pattern (Shen et al., 2016). Briefly, let w_i be the proportion of tumor cells in the ith tumor sample. For the jth gene, contamDE models the distribution of reads from normal cell samples as $N_{ij} \sim NB(k_i\mu_j, \phi_j)$ and those from "contaminated" tumor samples as $T_{ij} \sim NB(k_i{}' (\mu_j + w_i\delta_j), \phi_j)$, where NB denotes the negative binomial distribution, k_i and $k_i^{'}$ are normalization size factors for normal and tumor samples, μ_j and $\mu_j + w_i\delta_j$ are the adjusted means for normal and tumor samples, and ϕ_j is the dispersion. The DE is obtained by testing if δ_j is 0. UNDO is designed for deconvoluting array-based gene expression data of tumor samples (Wang et al., 2015), which models the mixing proportion of pure tumor and stroma cells as latent variables. However, tumor purity has multiplicative effects on gene expression, which might not be additive (Zheng et al., 2017). Thus, it is inadequate to simply treat tumor purity as an additive or semi-additive covariate in computational models.

To solve this problem, we have developed a novel method called *D*ifferential *E*xpression *C*aller by combining *t*umor *p*urity information (DECtp) to identify DEGs between tumor and normal samples. DECtp models expression profiles of tumor samples as a mixed Gaussian distribution, where the mixing proportion is tumor purity. With known or estimated tumor purity, differential expressions are then called based on a generalized least square procedure followed by the Wald test. We performed analyses on extensive simulated data with different sample sizes and noise levels and TCGA data of various cancers. DECtp achieves more accurate, consistent, and biologically meaningful results than those from other state of the art methods, such as limma (Ritchie et al., 2015).

MATERIALS AND METHODS

Supposing that the input data consists of expression profiles of N genes on n_0 normal and n_1 cancer samples, we first transform the expression values on each sample group (by log2 transformation, quantile normalization, and so on) such that they will follow a Gaussian distribution. This transformation allows for the introduction of a linear model with Gaussian noise in subsequent steps.

Specifically, for any gene i, let X_i be its transformed expressions on all normal samples. We assume that $X_i \sim N(m_i, \sigma_i^2)$, where m_i and σ_i^2 represent the mean and variance of X_i. Similarly, let Y_i be the transformed expressions on "pure" cancer samples for gene i, which also admits a normal distribution. Without loss of generality, we assume $Y_i = X_i + \delta_i$, where δ_i represents the difference between cancer and normal samples. Clearly, δ_i is a random variable following a normal distribution with mean μ_i and variance τ_i^2, i.e., $\delta_i \sim N(\mu_i, \tau_i^2)$. Thus, differential genes could be inferred by the hypothesis test: $H_0 : \mu_i = 0$. However in practice, the expression profile of "pure" cancer sample Y_i is not observed. Instead, the observed expressions of solid tumor samples are always a mixture of expressions on cancer and normal cells.

Let $Y_i^{'}$ be the expression profile of gene i on observed tumor samples. For a tumor sample with known purity λ_s estimated by existing methods, we use $Y_{is}^{'}$ to denote the expression of gene i on sample s. Then $Y_{is}^{'}$ can be modeled by a linear formula: $Y_{is}^{'} = (1-\lambda_s)X_{is} + \lambda_s Y_{is} = (1-\lambda_s)X_{is} + \lambda_s(X_{is} + \delta_{is}) = X_{is} + \lambda_s\delta_{is}$, so $Y_{is}{}' \sim N(m_i + \lambda_s\mu_i, \sigma_i^2 + \lambda_s^2\tau_i^2)$. Clearly, the gene expression

variance of tumor samples are greater than or equal to that of normal samples since $\sigma_i^2 + \lambda_s^2\tau_i^2 \geq \sigma_i^2$, and bias can arise when directly testing the mean difference between X_{is} and $Y_{is}^{'}$ due to the influence of tumor purity. It is worth noting that tumor purity has multiplicative (instead of additive) effect (Zheng et al., 2017) on differential expression under this assumption. So previous DEG calling method modeling tumor purity as an additive covariate might be inappropriate (Aran et al., 2015).

To solve this problem, we propose a simple linear model and a generalized least square procedure by taking X_{is} and $Y_{is}^{'}$ as input data. Specifically for gene i, the linear regression model is trained as follows: $Z_i = W\beta_i + \epsilon_i$, where

$$Z_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{in_0} \\ Y_{i1}^{'} \\ Y_{i2}^{'} \\ \vdots \\ Y_{in_1}^{'} \end{bmatrix}, W = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & \lambda_1 \\ 1 & \lambda_2 \\ \vdots & \vdots \\ 1 & \lambda_{n_1} \end{bmatrix}, \beta_i = \begin{bmatrix} m_i \\ \mu_i \end{bmatrix}, and\, \epsilon_i = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n_0} \\ \epsilon_{n_0+1} \\ \epsilon_{n_0+2} \\ \vdots \\ \epsilon_{n_0+n_1} \end{bmatrix}.$$

Here, the $(n_0 + n_1) \times 1$ vector Z_i represents expressions from normal and tumor samples with the first n_0 entries from normal samples, and the last n_1 entries from tumor samples. In addition, W is a matrix of dimensionality $(n_0 + n_1) \times 2$ with the first column consisting of all 1 s and the second column consisting of n_0 0 s and n_1 tumor purities (i.e., $\lambda_1, \lambda_2, \ldots, \lambda_{n_1}$) for respective tumor samples. β_i is the linear model parameter to be determined, and ϵ_i is the random error. The objective is to test $H_0 : \mu_i = 0$.

The parameters can be fitted by a least square procedure to minimize $\left|Z_i - (W\beta_i + \epsilon_i)\right|_2^2$. As a result, $\hat{\beta}_i = (W^TW)^{-1}W^TZ_i \triangleq HZ_i$ where $H = (W^TW)^{-1}W^T$, and $var\left(\hat{\beta}_i\right) = Hvar\,(Z_i)H^T$. The variance of Z_i is $\begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma^{'} \end{bmatrix}$, where $\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}_{n_0 \times n_0}$ and $\Sigma^{'} = \begin{bmatrix} \sigma^{'2} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^{'2} \end{bmatrix}_{n_1 \times n_1}$.

So, $var\left(\hat{\beta}_i\right) = Hvar\,(Z_i)\,H^T = [H_1\ H_2]\begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma^{'} \end{bmatrix}\begin{bmatrix} H_1^T \\ H_2^T \end{bmatrix} = H_1\Sigma H_1^T + H_2\Sigma' H_2^T$, then $var(\hat{\beta}_i)$ can be obtained with σ^2 and $\sigma^{'2}$, the residual variances from normal and cancer groups respectively. Given estimated $\hat{\beta}_i$, regression residuals are now $\hat{\epsilon} = Z_i - W\hat{\beta}_i$, and the residual variances from normal and cancer groups are obtained as $\sigma^2 = \frac{\sum_{i=1}^{n_0}\hat{\epsilon}_i^2}{n_0-2}$, $\sigma^{'2} = \frac{\sum_{i=n_0+1}^{n_0+n_1}\hat{\epsilon}_i^2}{n_1-2}$. We apply a shrinkage estimator similar to Cui et al. (2005) on the estimated cancer/normal variances, and obtained $\tilde{\sigma}^2$ and $\tilde{\sigma}^{'2}$. The procedure shrinks all residual variances to the genometeric mean and stabilizes the estimates. After getting $\hat{\beta}_i$ and $var\left(\hat{\beta}_i\right)$, the Wald test statistics for testing $H_0 : \mu_i = 0$ is calculated as $t_i = \frac{\hat{\beta}_{i[2]}}{\sqrt{var(\hat{\beta}_i)_{[2,2]}}}$, where $\hat{\beta}_{i[2]}$ is the second item of $\hat{\beta}_i$ and $\sqrt{var(\hat{\beta}_i)_{[2,2]}}$ is the element of the matrix $\sqrt{var(\hat{\beta}_i)}$ at indices [2,2]. Finally, we assume the Wald test follow a t distribution with $n_0 + n_1 - 2$ degrees of freedom, and the p-values can be obtained accordingly. False discovery rate (FDR) can be estimated using established procedures such as the Benjamini-Hochberg method (Benjamini et al., 2001).

RESULTS

We applied and compared DECtp with canonical DEG calling algorithms like limma on a few simulated datasets and cancer datasets downloaded from The Cancer Genome Atlas (https://cancergenome.nih.gov/). Before stepping into detailed analyses, it is insightful to first examine the relationship between gene expression and tumor purity.

Correlation Between Gene Expression and Tumor Purity

Through extensive analyses of the TCGA data, we discovered that the expression levels of many genes have strong correlation with tumor purity in cancer and the correlation increases with the difference of gene expressions between cancer and normal samples. Specifically, the tumor purities were downloaded from https://zenodo.org/record/253193, which were calculated by InfiniumPurify (Zhang et al., 2015; Zheng et al., 2017). InfiniumPurify for purity estimation is based on an important observation from the Illumina Infinium 450 k methylation data: the number of probes with intermediate methylation level is significantly greater in tumor samples than that in normal samples. InfiniumPurify first identifies a number of informative differentially methylated CpG sites (iDMCs) from cancer-normal comparison by using a non-parametric Wilcoxon Rank-Sum test and ANOVA analysis for each probe, and then estimates purity from the probability density of methylation levels of iDMCs.

Expression Levels of Many Genes Have Strong Correlation With Tumor Purity

We used Prostate adenocarcinoma (PRAD) in TCGA as an example to illustrate the correlation between gene expression and tumor purity. Specifically, after quantile-normalizing the expression profiles (quantified by RSEM, Li and Dewey, 2011) for tumor samples, the purity value of each sample was estimated by InfiniumPurify (Zheng et al., 2017). For each gene, we computed the Spearman correlation between expression levels and tumor purities across tumor samples (termed as "Observed" in **Figure 1A**). From there we obtained 20440 correlation values, each for a gene. As a comparison, we also randomly shuffled the purities of all tumor samples, and used the shuffled tumor purities as input to compute the correlation (termed as "Random" in **Figure 1A**). As can be seen from **Figure 1A**, the distribution of observed correlations has a longer right tail, demonstrating that there are much more genes with high correlation with tumor purity than by random. In particular, we identified 1252 genes

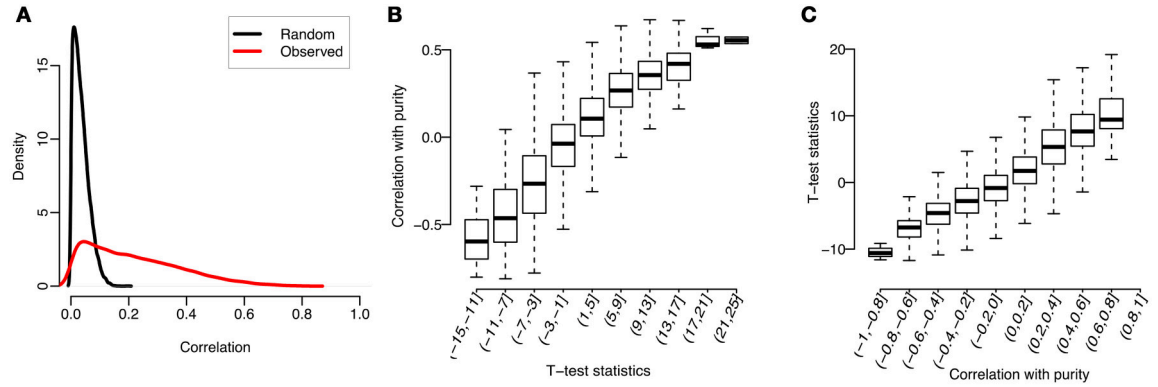

FIGURE 1 | Correlations between tumor purities and gene expressions for PRAD: **(A)** Distributions of Spearman correlations between gene expression and observed or randomly shuffled tumor purities across tumor samples; **(B)** Observed correlations, grouped by *t*-test statistics; **(C)** *t*-test statistics, grouped by observed correlations.

with absolute observed correlation over 0.5 (accounting for 6.2% of all genes), while this number is close to 0 by random.

Correlation Between Gene Expression and Tumor Purity Increases With the Difference of Gene Expressions Between Cancer and Normal Samples

We identified genes highly correlated with tumor purity. What are these genes? To answer this question, we studied the relationship between previously calculated correlations and gene expression changes between tumor and normal samples. Specifically, we first conducted a *t*-test on the normalized expression profiles of each gene between tumor and normal samples, and then divided all genes into 10 subsets by the test statistics. We then plotted in **Figure 1B** the distribution of observed correlations (between tumor purity and gene expression) in each group. As can be seen, the mean observed correlation in each group increases with the *t*-test statistics (measuring the extent of gene expression difference between tumor and normal samples). Similarly, we also classified the genes into 10 subgroups according to their correlations with tumor purity and observed a positive correlation between the *t*-test statistics and group labels (see **Figure 1C**).

We conducted the above analyses across 14 cancer types with sufficient normal tissues (each cancer type with over 10 normal samples) including Bladder Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Esophageal Carcinoma (ESCA), Head-Neck Squamous Cell Carcinoma (HNSC), Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), Cervical Kidney renal papillary cell carcinoma (KIRP), Liver Hepatocellular Carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), PRAD, Stomach Adenocarcinoma (STAD), Thyroid Cancer (THCA), and Uterine Corpus Endometrial Carcinoma (UCEC). The top 1000 genes with the largest correlations for each cancer type were shown in **Supplementary Table S1**. The results were similar for all cancers, which could be well explained by our linear regression model on gene expression (see Materials and Methods). When there are significant differences between tumor and normal samples (i.e., δ_{is} is big), the gene expressions are more correlated with purities. However, when there is no difference between tumor and normal samples (i.e., δ_{is} is close to 0), the gene expressions will have a low correlation with purities. These results revealed that tumor purity will bias differential expression analysis if not correctly accounted for, and our method was motivated from this observation.

Analyses on Simulated Data

To evaluate DECtp and compare it with other methods, we simulated a few datasets resembling true biological scenarios with different sample sizes and noise levels.

Simulated Datasets

We first downloaded from TCGA the LUAD gene expression data (in RSEM values) consisting of 517 tumor and 59 matched normal samples. Each RSEM value was transformed to log2 (RSEM + min), where min is the minimum non-zero RSEM value. The log2-transofmred data was quantile normalized, which was then used to generate simulation data.

It is worth mentioning that our purpose is to call DEGs between pure normal and pure tumor samples. However, both kinds of samples are infeasible to retrieve in reality, thus we adopted a compromised strategy as follows:

(1) For each gene i, we simulated expression profile of "pure" normal sample j as $X_{ij} \sim N(m_i, \sigma_i^2)$, where m_i is the mean expression of gene i across all 59 LUAD normal samples, and σ_i^2 is their variance.

(2) Similarly, we simulated expression profile of "pure" tumor sample j as $Y_{ij} \sim N(m_i^{'}, \sigma_i^{'2})$, where $m_i^{'}$ is mean expression across 517 LUAD tumor samples and $\sigma_i^{'2}$ is the variance. Since the two expression profiles ("pure" normal and "pure" tumor) are normally distributed, we assumed that gene i is a true DEG if $\left|m_i - m_i^{'}\right| \geq \delta$, where δ is a predefined threshold.

(3) We generated tumor purity values λ_j uniformly from [0.05, 0.95]. Plugging in X_{ij}, Y_{ij} and λ_j into the formula $Y'_{ij} = \lambda_j Y_{ij} + (1 - \lambda_j)X_{ij}$, we simulated Y'_{ij} as the observed expression profile of sample j at gene i, which is a mixture of expression profile from "pure" tumor and "pure" normal samples.

We then called DEGs between simulated pure normal (e.g., X_{ij}) and mixed (e.g., Y'_{ij}) samples and compared them with the underlying true DEGs to assess accuracy. Because the true mean expression levels are known, we can construct a gold standard for comparison. For a gene, if the absolute difference of the true expression profiles between normal and pure tumor samples is greater than a threshold, it is defined as a DEG. The simulations were repeated for $\delta = 1, 2, 3$, which roughly provides proportions of DEGs at 38%, 16%, and 8% of total number of genes. We also tested the performance of the algorithms with varied sample sizes from 10, 50, and 100, respectively.

DECtp Outperforms Other Methods in Simulated Datasets

We performed DEG calling on the 9 simulated datasets using DECtp and a few popular methods including *t*-test, limma. The receiver operating characteristic (ROC) curve analysis

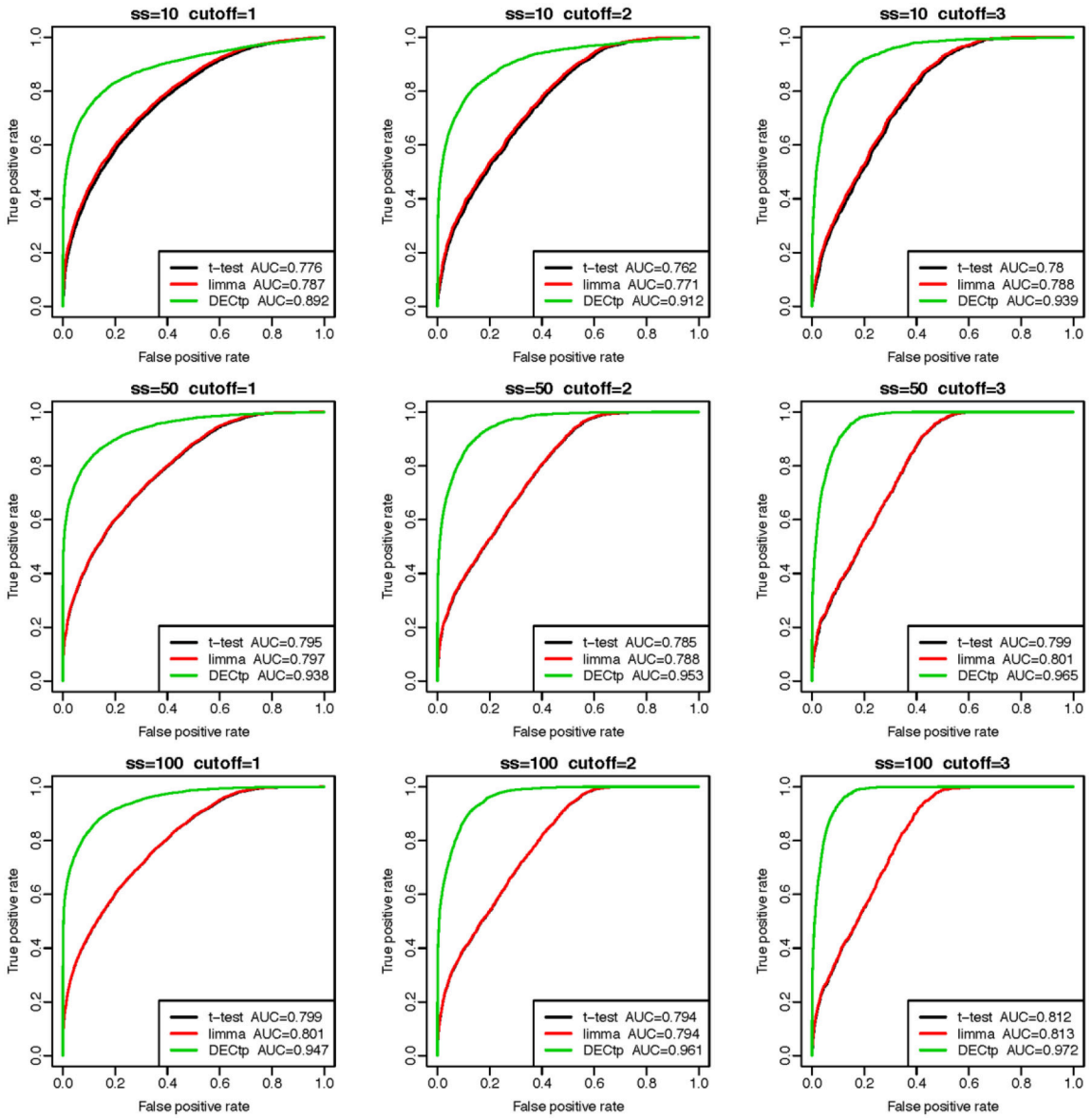

FIGURE 2 | Comparison of DE detection accuracies of the three methods including *t*-test, limma and DECtp on 9 simulated datasets with sample sizes 10, 50, and 100 and cutoffs (δ) 1, 2, and 3.

(Davis and Goadrich, 2006) using truth DEGs as a gold standard was performed to compare the performances of the methods (see **Figure 2**). Compared with traditional DEG calling methods, DECtp takes purity as an experimental design factor in a linear model. So we added to tumor purities a noise of the Gaussian distribution with mean 0 and standard deviations 0.1 to test the robustness of our method against purity estimation. It is clear that DECtp achieved the best AUCs in all simulated datasets even if estimated tumor purities are biased. In addition, limma and *t*-test have very similar performances, which is not surprising since it is known that they are similar for normal distributed data (Murie et al., 2009). Moreover, the performances of all methods became better when the thresholds (δ) or sample sizes increase as expected. Overall, these real data-based simulation results demonstrate the robustness and accuracy of DECtp in DE detection when tumor purity is a confounding factor.

Analyses on Real Data

With the success of DECtp on simulated data, we next tested DECtp on real TCGA tumor data on 14 cancer types including BLCA, BRCA, ESCA, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, STAD, THCA, and UCEC respectively. There are overall 6289 tumor and 632 normal samples. For all cancers, we estimated tumor purities by InfiniumPurify (Zheng et al., 2017).

The Top Differential Genes Identified by DECtp Is More Associated With Tumor Purity Than Those of Limma

To study the correlation between tumor purity and top ranked differential genes, we first ranked genes by their false discovery rate calculated by DECtp or limma. We then calculated the average absolute correlation between tumor purity and top n ranked genes. In **Figure 3A**, we plotted the average absolute correlation against n ($0 \leq n \leq 20000$) for BLCA and PRAD. Similar to previous findings, we found that top differentially expressed genes are more correlated with purity than other genes for both DECtp and limma. The trend is clearer for DECtp, indicating that it is better in identifying tumor purity-associated differential genes. The observation holds for all 14 cancer types (see **Supplementary Figure S1**).

We also examined the overlaps of DEGs (at FDR 0.001) called from the *t*-test, limma and DECtp. **Figure 3B** shows the overlapping Venn diagrams for BLCA and PRAD respectively. For BLCA, the *t*-test identified 5,689 DEGs, among which 4,231 (74%) are overlapped with those identified by DECtp. limma identified 5,393 DEGs, among which 4180 (78%) are overlapped with those identified by DECtp. Similarly for PRAD, the *t*-test identified 9,223 DEGs, among which 8696 (94%) are overlapped with those identified by DECtp. Limma identified 8682 DEGs, among which 8271 (95%) are overlapped with DECtp. The overlaps of DEGs for other cancer types were

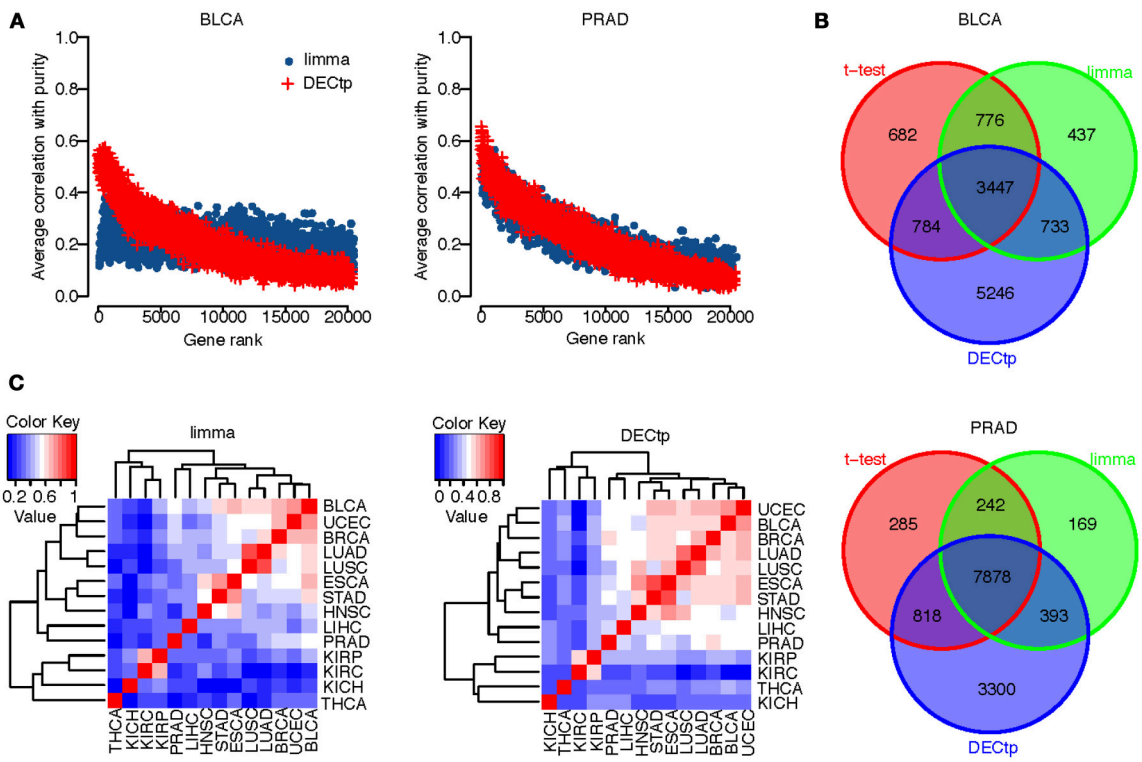

FIGURE 3 | (A) Scatter plot of the number of top-ranked genes versus the average absolute correlations for BLCA and PRAD; **(B)** Overlaps of DEGs called from *t*-test, limma and DECtp for BLCA and PRAD; **(C)** Inter-cancer correlations of test statistics by limma and DECtp.

shown in **Supplementary Figure S2**. In summary for all tested cancer types, there are 114842 DEGs overlapped between DECtp (with an overall of 151327 DEGs) and *t*-test (with an overall of 136918 DEGs), 107621 DEGs overlapped between DECtp and limma (with an overall of 121378 DEGs), 112772 DEGs overlapped between *t*-test and limma, suggesting that the three methods are generally consistent. We also have downloaded RNA-seq count data of six cancer types from TCGA, including BLCA, BRCA, HNSC, LUAD, LUSC, and PRAD to investigate the overlaps of DEGs called from DECtp, limma and edgeR. To have a fair comparison, we selected the same tumor and normal samples from the two different data type (count vs. RSEM value) when using DECtp and edgeR (332 normal samples versus 2858 tumor samples). The overlaps of DEGs for the three methods were shown in **Supplementary Figure S3**. It is shown that DEGs called from the three methods have rather significant overlap for the six cancer types. To be specific, for the six cancer types, limma identified 55593 DEGs, edgeR identified 59860 DEGs, and DECtp identified 71115 DEGs, and 44532 DEGs (accounting for 62.6%) in DECtp are overlapped with those identified by limma and edgeR.

Next, we examined the Pearson correlation among test statistics for different cancer types. Even though different cancer types have distinct etiologies, they might still share many genomic and transcriptomic features. We plotted in **Figure 3C**

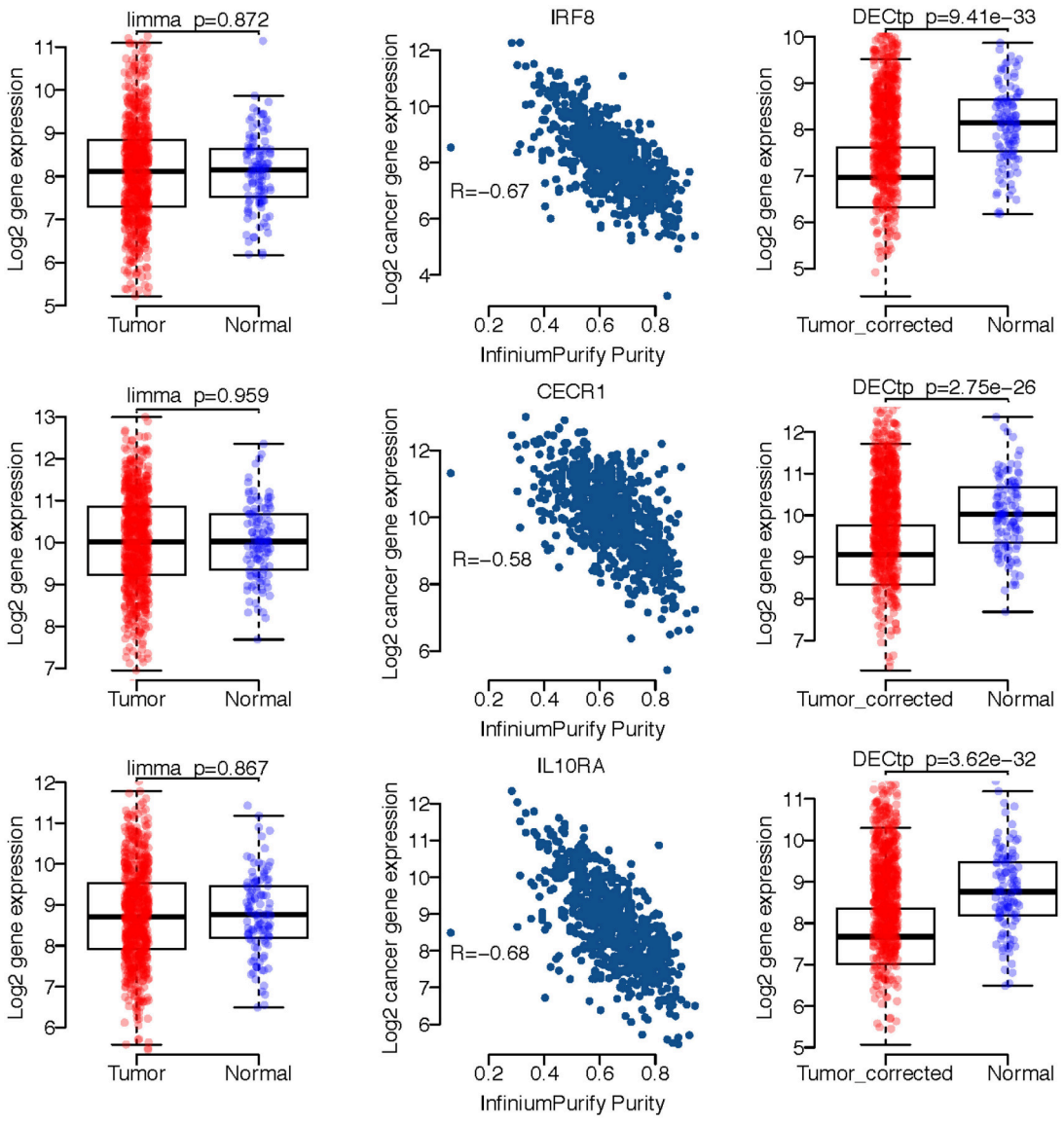

FIGURE 4 | A few exemplified DEGs only detected by DECtp in BRCA. **Left** panel shows log2 expression profiles distributions for tumor and normal samples. **Middle** panel shows Spearman correlations between log2 expression profiles and tumor purities. **Right** panel shows log2 expression profiles for tumor and normal samples after adjusting tumor purities.

the correlation of test statistics among 14 cancer types using both DECtp and limma. Overall the correlations for DECtp are higher than those of limma.

DECtp Identifies New Biological Meaningful Differential Genes

We selected several gene expression profiles from BRCA to demonstrate the confounding effect of tumor purity on differential expression analysis. As shown in **Figure 4**, the left panel displays the boxplots of three genes expression profiles including *IRF8*, *CECR1* and *IL10RA* for tumor and normal samples. It is clear that the *p*-values are not statistically significant for limma, i.e., the *p*-value is 0.872 for *IRF8*, 0.959 for *CECR1*, and 0.867 for *IL10RA*. The middle panel shows the scatter plot of expression profiles versus InfiniumPurify purities, in which the correlations are all very high, especially, −0.68 for *IL10RA*. The high correlation indicates that the large within group variance of cancer samples is mostly caused by variation in purities for different samples, which dilutes the signals of DEGs. And thus, after removing the effect of tumor purity, we could observe significant difference on expressions of these genes between normal and tumor groups. Indeed, there are many studies linking these 3 genes to breast cancer (Heinonen et al., 2008; Takaoka et al., 2008; Pavlides et al., 2010). We also selected the differentially expressed genes detected only by DECtp for the David enrichment analysis (at FDR < 0.05). **Supplementary Table S2** shows the enrichment of DE genes for the 10 cancer types. We have obtained a lot of biological functions. For example, GO:0006955∼immune response is the most enriched Go term for BLCA and PRAD with FDR being 6.161542e-29 and 1.45e-12, respectively. Thus, by considering tumor purity, DECtp could identify new biological meaningful DEGs for further experimental validation.

DECtp Is More Consistent and Identifies More Biological Meaningful Differential Genes Than Limma

It is known to us all that consistency is a very important criteria to evaluate DE calling methods on real data. Generally speaking, a robust method should obtain consistent results on technical or biological replicates. To compare the consistency of DECtp with that of limma, we randomly divided tumor samples in each cancer into two groups, and then detected DEGs by comparing the two tumor groups with normal samples, respectively. This process was repeated 50 times. **Figure 5A** shows the overlapping odd ratios of the top 500 DE genes for all 14 cancers. Clearly, DECtp detected more overlapped DE genes than those of limma in most cancer types, which suggests that it is more consistent. We then examined the biological implications of the DE calling results. To have a fair comparison, we selected top 4,000 differential genes by the two methods, and tested their enrichments with "PATHWAYS_IN_CANCER" from KEGG (Kanehisa and Goto, 2000), which contains 328 biologically meaningful genes. DECtp detects 110 genes compared to 80 genes by limma in UCEC. **Figure 5B** shows the −log10 of the *p*-values for the enrichment of DEGs in "PATHWAYS_IN_CANCER" by using the Chi-square test. As can be seen, DECtp shows much smaller *p*-values compared to limma in most cancer types, especially in UCEC and BRCA. Overall, these results suggest that DECtp can detect more enriched DEGs in "PATHWAYS_IN_CANCER" than limma.

DISCUSSIONS

In this work, we systematically investigated the impact of tumor purity as a confounding factor in differential expression analysis (Aran et al., 2015; Wang et al., 2015; Shen et al., 2016), and proposed a novel statistical model to adjust for tumor purity in

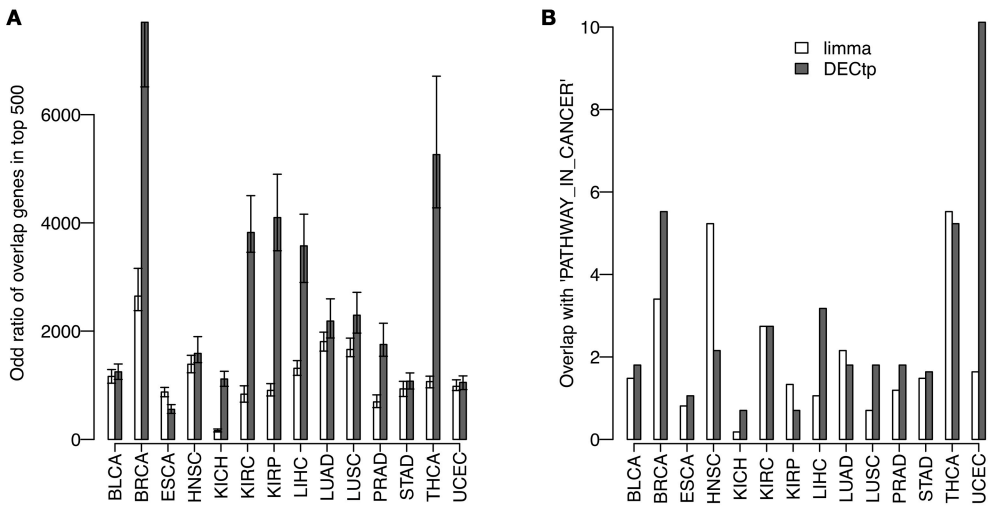

FIGURE 5 | Comparing the DEGs called by limma and DECtp: **(A)** The overlapping odd ratios of the top 500 DEGs between biological replicates for all 14 cancers; **(B)** Enrichment *p*-values for top 4,000 differentially expressed genes within "PATHWAY-IN-CANCER" from KEGG.

DE calling. We first examined the correlations between cancer expression profiles and tumor purity, and found that DE genes have high correlations with tumor purity. It is known that tumor purity has multiplicative effect on gene expression, instead of additive, so traditional DE calling methods ignoring tumor purity or modeling it as an additive covariate may present biased results. To solve this problem, we proposed DECtp, in which gene expression profiles from tumor samples are modeled as mixed Gaussian distributions, where the mixing proportion is tumor purity. DECtp achieved more robust and accurate DEGs in both simulation and real data studies compared with canonical methods like limma, which reinforces our previous claim that tumor purity may confound genomic analyses if not correctly accounted for (Zhang et al., 2017; Zheng et al., 2017).

DECtp is specifically developed to identify DEGs for gene expression profiles admitting normal distributions. However, RNA-sequencing technology has led to a rapid increase in gene expression data in the form of counts. The counts data are usually modeled by the negative binominal (NB) models, thus DECtp cannot be directly applied. In the future, it will be interesting to develop similar models using the NB distributions incorporating tumor purity information.

Finally, we would like to point out that DECtp may have a few further applications. Similar to differential gene analysis, differential protein and differential methylation analyses have also been widely performed between cancer and normal samples. In principle, DECtp could be applied to any differential analysis between cancer and normal samples given the data is Gaussian. In addition, Aran et al. found that identifying co-expression networks from genomics data without accounting for tumor purity is problematic (Aran et al., 2015). So we believe that similar principals proposed in this work can be applied to analyzing gene co-expression. Moreover, tumor purity information might be useful in identifying cancer associated expression quantitative trait loci (eQTLs). However, it is out of the scope of this study.

AUTHOR CONTRIBUTIONS

WZ and JY conceived the concept of the work. WZ, BH, and HL performed the experiments. WZ, JY, and BH wrote the paper.

FUNDING

This work was supported by the Hainan Provincial Natural Science Foundation of China (Grant No. 618MS057), National Natural Science Foundation of China (Grant Nos. 61762034 and 11661003), the Natural Science Foundation of Hunan, China (Grant No. 2018JJ2461), the research grant (Grant No. GJJ170445) from Science and Technology Project of Education Department of Jiangxi Province, the Key Program of Hunan Provincial Education Department (Grant No. 15A026), and the General Program of Hunan Provincial Philosophy and Social Science Planning Fund office (Grant No. 15YBA035).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00321/full#supplementary-material

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* 6, 8971. doi: 10.1038/ncomms9971

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* 125, 279–284. doi: 10.1016/S0166-4328(01)00297-2

Berger, M. F., Levin, J. Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., et al. (2010). Integrative analysis of the melanoma transcriptome. *Genome Res.* 20, 413–427. doi: 10.1101/gr.103697.109

Cui, X., Hwang, J. T., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 6, 59–75. doi: 10.1093/biostatistics/kxh018

Davis, J., and Goadrich, M. (2006). "The relationship between Precision-Recall and ROC curves," in *International Conference on Machine Learning* (Pittsburgh, PA), 233–240.

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Heinonen, H., Nieminen, A., Saarela, M., Kallioniemi, A., Klefström, J., Hautaniemi, S., et al. (2008). Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. *BMC Genomics* 9:348. doi: 10.1186/1471-2164-9-348

Joyce, J. A., and Pollard, J. W. (2009). Microenvironmental regulation of metastasis. *Nat. Rev. Cancer* 9, 239–252. doi: 10.1038/nrc2618

Junttila, M. R., and de Sauvage, F. J. (2013). Influence of tumour microenvironment heterogeneity on therapeutic response. *Nature* 501, 346–354. doi: 10.1038/nature12626

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Liang, P., and Pardee, A. B. (2003). Analysing differential gene expression in cancer. *Nat. Rev. Cancer* 3, 869–876. doi: 10.1038/nrc1214

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517. doi: 10.1101/gr.079558.108

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Murie, C., Woody, O., Lee, A. Y., and Nadon, R. (2009). Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics* 10:45. doi: 10.1186/1471-2105-10-45

Pavlides, S., Tsirigos, A., Vera, I., Flomenberg, N., Frank, P. G., Casimiro, M. C., et al. (2010). Transcriptional evidence for the "Reverse Warburg Effect" in human breast cancer tumor stroma and metastasis: similarities with oxidative stress, inflammation, Alzheimer's disease, and "Neuron-Glia Metabolic Coupling". *Aging (Albany NY)* 2, 185–199. doi: 10.18632/aging.100134

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Robinson, M. D., and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23, 2881–2887. doi: 10.1093/bioinformatics/btm453

Shen, Q., Hu, J., Jiang, N., Hu, X., Luo, Z., and Zhang, H. (2016). contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples. *Bioinformatics* 32, 705–712. doi: 10.1093/bioinformatics/btv657

Takaoka, A., Tamura, T., and Taniguchi, T. (2008). Interferon regulatory factor family of transcription factors and regulation of oncogenesis. *Cancer Sci.* 99, 467–478. doi: 10.1111/j.1349-7006.2007.00720.x

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138. doi: 10.1093/bioinformatics/btp612

Wang, N., Gong, T., Clarke, R., Chen, L., Shih, IeM., Zhang, Z., et al. (2015). UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics* 31, 137–139. doi: 10.1093/bioinformatics/btu607

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 14, 232–243. doi: 10.1093/biostatistics/kxs033

Zhang, N., Wu, H. J., Zhang, W., Wang, J., Wu, H., and Zheng, X. (2015). Predicting tumor purity from methylation microarray data. *Bioinformatics* 31, 3401–3405. doi: 10.1093/bioinformatics/btv370

Zhang, W., Feng, H., Wu, H., and Zheng, X. (2017). Accounting for tumor purity improves cancer subtype classification from DNA methylation data. *Bioinformatics* 33, 2651–2657. doi: 10.1093/bioinformatics/btx303

Zheng, X., Zhang, N., Wu, H. J., and Wu, H. (2017). Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.* 18, 17. doi: 10.1186/s13059-016-1143-5

Zhou, Y. H., Xia, K., and Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* 27, 2672–2678. doi: 10.1093/bioinformatics/btr449

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.