

## A novel feature selection for RNA-seq analysis

Henry Han

*Department of Computer and Information Science Fordham University, Lincoln Center, New York, NY 10023*

---

### Abstract

RNA-seq data are challenging existing omics data analytics for its volume and complexity. Although quite a few computational models were proposed from different standing points to conduct differential expression (D.E.) analysis, almost all these methods do not provide a rigorous feature selection for high-dimensional RNA-seq count data. Instead, most or even all genes are invited into differential calls no matter they have real contributions to data variations or not. Thus, it would inevitably affect the robustness of D.E. analysis and lead to the increase of false positive ratios.

In this study, we presented a novel feature selection method: nonnegative singular value approximation (NSVA) to enhance RNA-seq differential expression analysis by taking advantage of RNA-seq count data's non-negativity. As a variance-based feature selection method, it selects genes according to its contribution to the first singular value direction of input data in a data-driven approach. It demonstrates robustness to depth bias and gene length bias in feature selection in comparison with its five peer methods. Combining with state-of-the-art RNA-seq differential expression analysis, it contributes to enhancing differential expression analysis by lowering false discovery rates caused by the biases. Furthermore, we demonstrated the effectiveness of the proposed feature selection by proposing a data-driven differential expression analysis: NSVA-seq, besides conducting network marker discovery.

*Keywords:* RNA-seq, Feature selection, differential expression analysis

*2010 MSC:* 00-01, 99-00

---

*Email address:* xhan9@fordham.edu (Henry Han)

## 1. Introduction

RNA-seq provides a revolutionary way to unveil transcription by using ultra-high-throughput sequencing technologies to generate hundreds of million short reads from RNA molecules [1, 2, 3]. As raw RNA-seq data, the short reads usually ask several to  
5 even hundreds of Gigabytes storage. The short reads are further assembled or aligned against a reference genome (e.g. human genome) to produce a transcriptome by using assembly or alignment tools such as Bowtie, SOAPdenovo-Trans, SOAP3, or HTSeq [4, 5, 6]. As a genome level transcription map, the transcriptome consist of the expres-  
10 sion levels of all genes in transcription and each gene's expression is represented as the number of short reads mapped to the gene in the alignment or assembly [7]. In fact, the terminology gene refers to more general biological features in transcription such as a gene, exon, or transcript [8, 9].

The transcription map can be represented by a nonnegative integer read count matrix  $X^{m \times n} \geq 0$  by collecting all read counts mapped to each gene, where each row  
15 and column represent a gene and sample respectively. According to different sources, a sample can be classified as a biological or technical replicate. The former is an alternative sequencing of a same biological sample, and the latter is the direct sequencing of an independent biological sample. For the convenience of description, we also use RNA-seq data to refer to the read count matrix  $X$  of the original RNA-seq data in this  
20 study.

RNA-seq data are actually biased data subject to sequencing depth and gene length by the nature of RNA-seq sequencing even after normalization [10, 11]. The sequencing depth bias means that those genes from a sample with a high sequencing depth, usually have higher expression levels (high counts) than the same genes from a sample  
25 with normal or low sequencing depth; The gene length bias refers to that long genes have more counts in RNA-sequencing than short genes. The biases bring challenges in normalization, differential expression (D.E.) analysis, and feature selection [10, 11].

However, they share the same characteristic with traditional omics data (e.g. microarray). That is, they both are high-dimensional data, i.e. the number of variables  
30 (genes) is generally much greater than the number of samples (observations) in a RNA-

seq dataset  $X \in \mathbb{Z}^{n \times p}$ , i.e.  $n \gg p$ . Unlike the traditional omics data, RNA-seq data are generally modeled by a poisson or negative binomial (NB) distribution instead of a normal or log-normal distribution [7, 8, 10].

An essential issue in RNA-seq analysis is to answer the query: 'given a read count  
35 matrix, how to robustly determine whether the observed difference for a gene across two or more conditions is statistically significant? '. Quite a few differential expression (D.E.) analysis methods were proposed to answer it from different perspectives [8, 9, 11, 12, 13, 14, 15].

They can be categorized as either parametric or non-parametric methods according  
40 to whether they rely on statistical parameter estimation modeling approaches or not in D.E. analysis. The former assumes that RNA-seq data are subject to a well defined distribution and estimates corresponding parameters for the distribution before conducting hypothesis tests. For example, *DESeq* and *edgeR* methods both model RNA-seq data by the negative binomial (NB) distribution and estimate genes' mean  $\mu$  and variance  
45  $\sigma^2$  parameters [8, 13]; The latter does not assume data are subject to any distribution. Instead, its differential expression call can be based on an empirical distribution of some statistics derived from the input data. For example, *NOISEq* finds differentially expressed (DE) genes by relying on two derived statistics: absolute-expression and log-fold change between conditions [11].

50 However, all the D.E. analysis methods usually invite almost all genes into D.E. analysis without conducting a rigorous feature selection for high-dimensional RNA-seq data. Although they usually employ some simple filtering techniques to exclude those genes with zero or low counts before normalization, such a low-count filtering is by no means a desirable feature selection for high-dimensional data. Especially, such  
55 a count-based filtering may have a risk to affect the following D.E. analysis by feeding it with many high-count genes due to sequencing depth and gene length biases. As a result, it not only would increase false positives in the D.E. analysis, but also easily lead to a misleading result that those genes are differentially expressed simply because they have higher coverages or long gene length.

60 As typical high-dimensional data, RNA-seq data call for a rigorous feature selection algorithm, which should be robust to the depth and gene-length biases, to select poten-

tially D.E. genes for the sake of D.E. analysis, rather than its number of read counts only. Such a feature selection should overcome the weakness of the naive count-based filtering by removing redundant and noise-contained genes, instead of only low-count  
65 genes.

In fact, those high-count genes with very low variance values can be viewed as the redundant genes because they may not have real contributions to data variations. Furthermore, some low count genes with several exceptional high count peaks on few replicates in a same condition can be viewed as noise-contained genes, because the  
70 peaks could be generated from depth-based oversampling. The genes will be falsely identified as D.E. genes in differential calls, even if the observed differences between conditions are actually due to the artifacts of over-sequencing or library preparations instead of real reactions to a treatment.

On the other hand, a rigorous feature selection will strengthen RNA-seq normaliza-  
75 tion efforts to alleviate the effects of the depth and gene length bias factors by removing those genes not totally 'corrected' in normalization, in addition to lowering the computing complexities in the following D.E. analysis. As such, there needs a rigorous feature selection to glean meaningful genes to achieve a more targeted and accurate differential expression analysis.

In this study, we present a novel feature selection method: nonnegative singular  
80 value approximation (NSVA) to enhance RNA-seq analysis by taking advantages of RNA-seq data's built-in non-negativity. The non-negativity is an important characteristic of RNA-seq data, but it is ignored in most feature selection methods. Nonnegativity based analysis can contribute to enhancing data locality and capturing latent data be-  
85 havior [20]. As a data driven feature selection, NSVA does not assume any priori distribution for RNA-seq data; As a variance-based feature selection, it selects genes according to its contribution to the first singular value direction of input data.

We compare the proposed feature selection method with its five peers in state-of-  
the-art RNA-seq differential expression analysis. It demonstrates robustness to depth  
90 bias and gene length bias in feature selection and contributes to more efficient D.E. analysis. To further explore NSVA's effectiveness, we propose a data-driven differential expression analysis: NSVA-seq that is a novel nonparametric D.E. analysis without



M-D odd ratio comparisons[11]. More importantly, it overcomes the weakness of existing D.E. models for input data with a small number of samples, and demonstrates  
95 a better sensitivity in D.E. analysis than its peers for the datasets with a few samples [8]. Finally, we demonstrate that the proposed feature selection method can also lead to meaningful network marker discovery for complex diseases.

## 2. Methods

Various feature selection algorithms are widely available for traditional omics data  
100 via different statistical tests [16]. However, most of these statistical tests based methods (e.g. *t-test*) can not apply to RNA-seq data directly, because they usually assume population data are normally distributed [17]. Some nonparametric statistical tests proposed for microarray data are available, but they are not widely employed in RNA-seq analysis probably because of their different generation mechanisms or differential analysis  
105 approaches [7, 12, 18].

On the other hand, traditional transform-based feature selection methods such as principal component analysis (PCA), nonnegative matrix factorization (NMF) or their variants can apply to RNA-seq data directly due to their purely data-driven characteristics, in which no distributions are assumed for input data [19, 20, 21, 22, 23, 24]. In fact,  
110 they transform input data into a subspace generated by principal components, or non-negative bases to seek meaningful linear combinations of features (genes). However, they face difficulties in ranking each gene because it is technically hard to distinguish an individual gene's contribution to all genes' linear combinations due to the nature of the linear or nonlinear transforms.

As such, it is believed that a desirable feature selection for RNA-seq data should  
115 satisfy the following criteria. First, it should be a data-driven method without prior data distribution assumption to prevent possible probabilistic modeling biases. Second, it should avoid evaluating each gene's significance from the linear combinations of all genes in a subspace directly. Third, it should take consideration of the nonnegative  
120 characteristic of RNA-seq data instead of treating them as generic data to maintain locality in data analysis. Fourth, it should overcome the weakness of naive count-

based filtering and contribute to following D.E. analysis by demonstrating robustness to depth and gene-length biases. That is, the feature selection algorithm should be more efficient than its other peers in identifying potential differentially expressed (DE) genes, rather than picking high-count genes or long genes only in feature selection.

We present a novel data-driven feature selection method: nonnegative singular value approximation (NSVA) that satisfies the criteria. It can be viewed as a special singular value decomposition (SVD) built upon Perron-Frobenius theorem, which is widely used in Google webpage ranking, to disclose novel findings for nonnegative data [25, 26, 27].

*Perron-Frobenius theorem.* Given matrix  $A \in \mathbb{R}^{n \times n}$ ,  $A > 0$ , its largest eigenvalue  $\lambda_{max} > 0$  is always positive and its associative eigenvector  $v$  is always positive, i.e.  $v > 0$ . For any  $Ax = \lambda x$ ,  $\lambda_{max} \neq \lambda$ , there exists at least one entry  $x_j \in x$ ,  $x_j < 0$ .

### 2.1. Nonnegative singular value approximation (NSVA)

Given a nonnegative matrix  $A \in \mathbb{R}^{n \times p}$ ,  $A \geq 0$  with a rank  $r = \min(n, p)$ , and its SVD decomposition  $A = \sum_{i=1}^r s_i u_i v_i^T$ , where  $u_i \in \mathbb{R}^n$  and  $v_i \in \mathbb{R}^p$ ,  $i = 1, 2 \dots r$ , then, we have the following results,

1. Both  $u_1$  and  $v_1$  have only nonnegative entries, i.e.,  $u_1 \geq 0$ ,  $v_1 \geq 0$ .
2. The vectors  $u_j$  and  $v_k$  contain at least one negative entry when  $j \geq 2$ ,  $j = 1, 2 \dots n$ , and  $k \geq 2$ ,  $k = 1, 2 \dots r$
3. Matrix A has the following first level singular value approximation:

$$A \sim s_1 u_1 v_1^T = \sum_{i=1}^n \sum_{j=1}^p s_1 u_{i1} v_{j1} \quad (1)$$

by dropping all  $u_i$  and  $v_j$ , when  $i, j \geq 2$ .

To prove nonnegative singular value approximation, we prove the following Perron-Frobenius extension theorem at first, which extends the results of the original theorem to nonnegative data.

145 *2.1.1. Lemma: Perron-Frobenius extension theorem*

Given matrix  $A \in \mathfrak{R}^{n \times n}$ ,  $A \geq 0$ , its largest eigenvalue  $\lambda_{max}(A) \geq 0$  and its associative eigenvector  $v$  is nonnegative, i.e.  $v_{max} \geq 0$ .  $v > 0$ . For any For any  $Ax = \lambda x$ ,  $\lambda_{max} \neq \lambda$ , there exists at least one entry  $x_j \in x$ ,  $x_j < 0$ .

Proof. We approximate  $A$  as a sequence of positive matrix  $A_n$  such as  $\lim_{n \rightarrow \infty} A_n =$   
 150  $A$ . For example, if  $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ , then it can be approximated as a sequence posi-  
 tive matrix  $A_n = \begin{pmatrix} 1 & 1 \\ 1 & \frac{1}{n} \end{pmatrix}$ , i.e.  $\lim_{n \rightarrow \infty} A_n = A$ .

It is clear that the characteristic equation of  $A_n$ :  $\det(\lambda I - A_n)$  will also approximate the characteristic equation of  $A$ :  $\det(\lambda I - A)$  when  $n \rightarrow \infty$ . As such,

$$\lim_{n \rightarrow \infty} \lambda_{max}(A_n) = \lambda_{max}(A) \quad (2)$$

Since all  $\lambda_{max}(A_n)$  are positive, its limit should be nonnegative by the compactness of the sequence convergence, that is  $\lambda_{max}(A) \geq 0$ . Similarly, we normalize corresponding eigenvector  $v_n$  of  $\lambda_{max}(A_n)$  such that  $\|v_n\|_1 = 1$ , i.e.  $\vec{1}^T v_n = 1$ , where  $\vec{1}$  is a  
 155 vector with all entries as 1. Thus,  $\lim_{n \rightarrow \infty} v_n = v$ .

Do limit for the following equation for the positive sequences,

$$A_n v_n = \lambda_{max}(A_n) v_n \quad (3)$$

we have  $Av = \lambda_{max}(A)v \geq 0$ . That is,  $v \geq 0$ . Proceeding in the similar way, we can prove  $v > 0$ . For any  $Ax = \lambda x$ ,  $\lambda_{max} \neq \lambda$ , there exists at least one entry  $x_j \in x$ ,  $x_j < 0$ .

*2.1.2. The proof of nonnegative singular value approximation.*

160 Suppose we have a nonnegative matrix  $A \in \mathfrak{R}^{n \times p}$ ,  $A \geq 0$ , we assume  $n \geq p$  without loss of generality. Then,  $AA^T \in \mathfrak{R}^{n \times n}$ ,  $A^T A \in \mathfrak{R}^{p \times p}$ , both are nonnegative semi-positive definite matrices.

By singular value decomposition (SVD):  $A = \sum_{i=1}^p s_i u_i v_i^T$ , where  $s_i$  is the  $i^{th}$  singular value of  $A$ , it is easy to know  $u_1$  and  $v_1$  are the first eigenvectors of  $AA^T$   
 165 and  $A^T A$  respectively. That is, their corresponding eigenvalues are the first (largest)

eigenvalue of  $AA^T$  and  $A^T A$ . The vectors  $u_j, j = 2 \cdots n$ , and  $v_k, k = 2 \cdots p$  are the other eigenvectors of  $AA^T$  and  $A^T A$  respectively.

Applying the Perron-Frobenius extension theorem, we have  $u_1 \geq 0, v_1 \geq 0$ . The other vectors  $u_j$  and  $v_k$ , contain at least one negative entry. We only use the  $u_1$  and  $v_1$  to decompose  $A$  and drop  $u_j$  and  $v_k$ , for  $j, k \geq 2$ , we have  $A \sim s_1 u_1 v_1^T = \sum_{i=1}^n \sum_{j=1}^p s_1 u_{i1} v_{j1}$ .

### 2.1.3. The biological meaning of nonnegative singular value approximation in RNA-seq analysis

It is noted that NSVA guarantees a purely additive decomposition of a nonnegative matrix along the first singular value direction  $v_1$ .  $A \sim s_1 u_1 v_1^T = \sum_{i=1}^n \sum_{j=1}^p s_1 u_1^{(i)} v_1^{(j)}$ . In fact, each nonnegative entry  $u_1^{(i)}$  in  $u_1$  can be viewed as a corresponding coefficient of the row  $A_i^T$ , which represents the  $i^{th}$  gene of input RNA-seq data, in the one-dimensional “meta-sample space” spanned by all entries of  $v_1$ , i.e.  $S = span(v_1^{(1)}, v_1^{(2)} \cdots v_1^{(p)})$  with a weight  $s_1$ . Thus, from a single gene viewpoint, NSVA implies that each gene is approximated by the projection of its corresponding entry in vector  $u_1$  on the singular value direction  $v_1$ , i.e.  $A_i^T \sim s_1 u_1^{(i)} v_1^T, i = 1, 2, \cdots n$ .

Such an approximation makes it possible to rank each gene by using its coefficient in the meta-sample space, where each  $v_1^{(j)}$  is the meta-sample corresponding to the original  $j^{th}$  sample, and  $u_1^{(i)}$  indicates the  $i^{th}$  gene  $A_i^T$ 's contribution to all the meta-samples. It answers the following question: what's a gene's contribution to all meta-samples along with the first singular value direction? As such, it is natural to define  $u_1^{(i)}$  as a gene contribution score (GCS) to quantify its contribution to all meta-samples.

### 2.1.4. Gene contribution scores (GCS)

A gene contribution score (GCS) measures a gene's contribution to all samples of a RNA-seq dataset  $A \in \mathbb{R}^{n \times p}$  by evaluating its contribution to all meta-samples in a low dimensional space. The gene contribution score of the  $i^{th}$  gene to all samples is defined as  $u_1^{(i)} \sim s_1^{-1} \sum_{j=1}^p a_{ij} v_1^{(j)} = s_1^{-1} A_i^T v_1$  by applying nonnegative singular value approximation. It is clear that filtering genes according to their gene contribution scores is equivalent to filtering genes by their count variances by the nature of the GCS.

195 In fact, NSVA feature selection consists of two major steps. The first step conducts NSVA for input data and calculates GCS for each gene. The second step ranks all genes by their GCSs and selects the genes with large GCSs for the following D.E. analysis.

It is worthwhile to point out that the first singular value  $s_1$  is usually quite large for a RNA-seq dataset compared with the other singular values. we define a data variation  
200 explanation ratio as

$$\rho = \frac{s_1}{\sum_{i=1}^p s_i} \quad (4)$$

It is the ratio between the first singular value and the sum of singular values, to evaluate that the percentage of information can be represented in NSVA. The ratio actually represents the percentage of the data variances along the first singular vector direction. In fact, we have found that the ratio usually reaches at least  $\rho \geq 60\%$  or even  
205 90% high for most RNA-seq datasets. For examples, the data variation explanation ratios of the three datasets in this study are 60.49%, 85.60% and 90.16% for the *Marioni*, *Prostate* and *Fly embryos* datasets respectively. In fact, the high data variation explanation ratios demonstrated by RNA-seq data guarantee the effectiveness of the first level singular value data approximation and the feasibility of the proposed feature selection  
210 algorithm.

## 2.2. NSVA-Seq: a data-driven differential expression analysis method

We propose a data-driven differential expression analysis method: NSVA-seq that employs NSVA to collect potential DE genes and compares each gene's expression  
215 with those of remaining genes under a modified-fisher-exact-test (mFET) by computing exact p-values. Unlike other methods (e.g. DESeq), NSVA-seq avoids parameter estimation and tuning. Moreover, its average expression based hypothesis query under a contingency table can somewhat avoid the limitations of the existing D.E. analysis methods such that data with few number of replicates will not be 'discriminated' in  
220 D.E. analysis for its built-in disadvantage in parameter estimation or M-D odd ratio estimations[8, 11].

The NSVA-seq can be simply described as: given a normalized library, NSVA-seq applies the modified fisher exact test (mFET) to a set of genes selected by NSVA. Our modified fisher exact test (mFET) is described as follows.

225

**Table 1.** Modified fisher exact test (*mFET*)

Gene	Control	Treatment	Total
Query gene $g$	$a$	$b$	$a + b$
Remaining genes	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

230 The original fisher exact test is used to determine if there are non-random associations between two categorical variables [28]. In the modified fisher exact test, we query if a gene is differentially expressed by comparing it with all the remaining genes. Table 1 illustrates the mFET's initial state: parameters  $a$ ,  $b$  are the expression levels of a given gene  $g$  under the control and treatment conditions respectively. Similarly, parameters  $c$  and  $d$  are the expression levels of all remaining genes under the two conditions. All of these parameters are all non-negative integers initially.

The null hypothesis  $\mathbf{H}_0$  in context is to query whether a gene  $g$  has the same level expression as all the other remaining genes. The p-value of this hypothesis test can be calculated by a hypergeometric distribution as

$$p = \frac{(a + b)!(c + d)!(a + c)!(b + d)!}{a!b!c!d!n!} \quad (5)$$

235 , where  $n = a + b + c + d$ .

To avoid large computing complexities from the large or even huge values of gene count data, we apply a log transform to the equation and employ gamma function  $\Gamma(n + 1)$  to valuate  $n!$ . That is,

$$p = e^{(\sum_{i=1}^4 \ln(t_i+1) - \sum_{i=1}^5 \ln(s_i+1))} \quad (6)$$

where  $t_1 = a + b$ ,  $t_2 = c + d$ ,  $t_3 = a + c$ ,  $t_4 = b + d$ ,  $s_1 = a$ ,  $s_2 = b$ ,  $s_3 = c$ ,  $s_4 = d$ , and  $s_5 = n$ . In fact, we are able to extend the calculations to any nonnegative data instead of only nonnegative integers by the nature of the transformation. As such, the modified fisher exact test (mFET) can be reformulated as follows.

Given a RNA-seq data  $X \in \mathfrak{R}^{N \times p}$ , where each gene has two conditions control (C) and treatment (T), we have the following parameter specifications in the proposed mFET method:

$$a = \frac{1}{m_c} \sum_{j:\tau(j)=C} X_{ij} \quad (7)$$

$$b = \frac{1}{m_t} \sum_{j:\tau(j)=T} X_{ij} \quad (8)$$

where  $X_{ij}$  is the expression level of the  $i^{th}$  gene of the sample  $j$ ,  $\tau(j)$  is the condition of the  $j^{th}$  sample, and  $m_c$  and  $m_t$  are number of samples in the control and treatment conditions respectively. Similarly, we have

$$c = \frac{1}{m_c(N-1)} \sum_{k \neq i} \sum_{j:\tau(j)=C} X_{kj} \quad (9)$$

$$d = \frac{1}{m_t(N-1)} \sum_{k \neq i} \sum_{j:\tau(j)=T} X_{kj} \quad (10)$$

240 The proposed NSVA-seq provides more freedom in D.E. analysis than existing D.E. models. It can not only work well for normalized data, but also raw read count integer data. The proposed modified fisher exact test not only extends the input data domain of the original fisher exact test, but also generalizes its differential expression test for each gene by using the whole remaining gene expressions. As a nonparametric method,  
 245 NSVA-seq does not need a parameter estimation process to find the mean and variance parameters of a specified distribution. As such, it somewhat provides a 'fair' D.E. analysis environment for those datasets consisting of few samples. It can be essentially important for clinical D.E. analysis, in which no enough samples are generally available [9, 13, 15]. On the other hand, it does not need to do M-D odd ratio comparisons as  
 250 NOISeq for its more transparent D.E. analysis mechanism [11, 29].

### 3. Result

#### 3.1. Datasets

We include three benchmark RNA-seq datasets in our experiments and their details are described as follows.

255 **Marioni data** originally consist of 32,000 genes across 14 samples after Illumina-supplied alignment algorithm ELAND. The samples are composed of two groups: the seven technical replicates from a kidney sample and another seven technical replicates from a liver sample, both of which are from a single human male [7]. This dataset is an important benchmark in normalization and D.E. analysis: it includes important gene  
260 length information for each gene compared to other RNA-seq datasets.

**Prostate data** consist of 17 million short reads and they were sequenced under the Illumina technology for two types of samples: four prostate cancer cells treated with androgen/DHT (DHT-treated), and three prostate cancer LNCap cells without DHT treatment (Mock-treated) [34]. We employed *Bowtie* and *SAMtools* to align its raw  
265 data with respect to the the human genome indexes (NCBI version 37), and obtained a nonnegative integer matrix with 4 DHT-treated and 3 Mock-treated samples across 23,068 genes [5, 30].

**Fly embryos data** consist of 17,605 genes across four samples. The four samples are composed of two biological replicates at conditions “A” (treated) and “B” (control) respectively. This dataset consists of only four samples and usually demonstrate  
270 some disadvantage in the existing D.E. models that require relatively more samples to complete parameter estimation [8, 13, 15].

### 3.2. *DESeq analysis with nonnegative singular value approximation (NSVA) feature selection*

275 To verify the effectiveness of the proposed feature selection, we firstly combine it with DESeq model, which is a typical parametric D.E. analysis model, to answer the query: ‘what will happen to DESeq analysis when NSVA feature selection is applied to input data?’ Figure 1 evaluates the performance of three NSVA-selected gene-sets consisting of 2,000, 3,000, and 5,000 genes and original data under DESeq. The false  
280 discovery ratio (FDR) cutoff was uniformly chosen as 0.001 in our experiments. Each horizontal and vertical axis in the subplots represent the mean expression of each gene and its corresponding  $\log_2$  fold changes under two different conditions respectively. The differentially expressed (DE) and non-DE genes are represented by red and blue markers respectively.



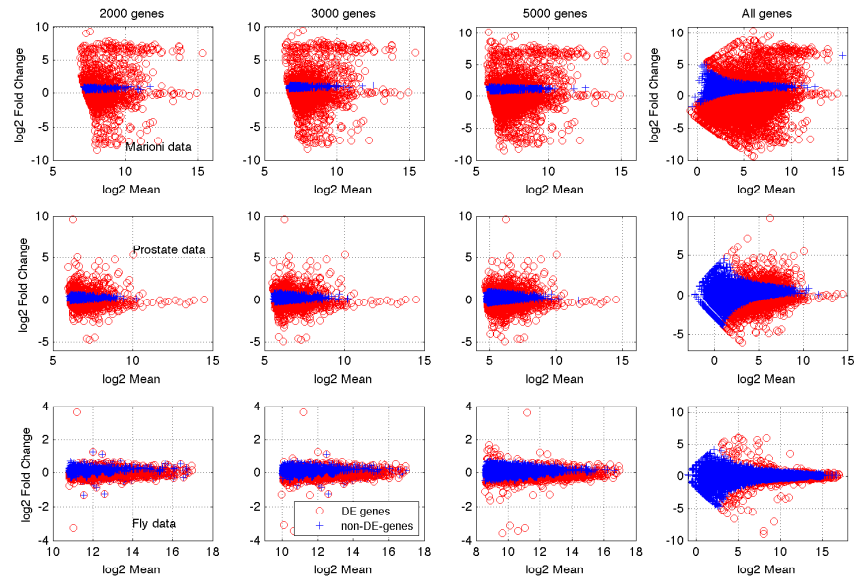


Figure 1: The scatter plots of normalized data mean versus  $\log_2$  fold change for original data and different NSVA-selected gene sets under DESeq on Marioni and Prostate datasets. The D.E. and non-D.E. genes are represented by red and gray markers respectively. The non-D.E. genes dropped remarkably in DESeq analysis when NSVA is applied to filter more genes.

285 Interestingly, The non-D.E. genes seem to drop remarkably under DESeq when  
NSVA is applied to each dataset. It indicates that the proposed NSVA feature  
selection demonstrates a good sensitivity to filter those non-differentially expressed (non-DE)  
genes for each dataset by picking the genes with large gene contribution scores (GCS).  
In other words, NSVA seems to be able to select more potential DE genes, which have  
290 more contributions to variance on behalf of the first singular value direction. Such a  
feature selection makes the following DESeq analysis more focused on the potential  
DE genes and contributes to decreasing false positives in D.E. analysis. Such a re-  
sult suggests that the proposed feature selection can enhance D.E. analysis by picking  
meaningful genes.

295 *3.2.1. The impacts of NSVA feature selection on the DESeq model*

In addition to comparing NSVA-DESeq with DESeq from a performance stand point, we have the following findings about NSVA's impacts on the DESeq model itself.

300 First, we have found that NSVA feature selection will contribute to a better size factor estimation in normalization because of filtering outliers (e.g., genes with very low counts). The size factor  $s_j$  of a sample  $j$  in DESeq model is a normalization factor to make the sample, which may be subject to different sequencing depth, comparable with the others. However, the size factors actually relies on a pseudo-reference sample, which is a virtual sample consisting of geometric means of all genes [8]. Filtering the  
305 outliers will prevent their geometric means from being entries of the pseudo-reference sample, which will cause the size factor estimation to be closer to the 'truth' and mitigate the bias caused by the sequencing depth.

Second, NSVA feature selection makes the parameter estimations of  $u_{ij}$  and  $\sigma_{ij}^2$ , which are mean and variance parameters of gene  $x_{ij}$  (the  $i^{th}$  gene on the  $j^{th}$  sample),  
310 under a negative binomial (NB) distribution, more robust. This was partially because both parameter estimations were strongly dependent on the estimation of the size factor  $s_j$  [8]. More interestingly, we found that the smooth function used by the DESeq method to model the dependence of the raw variance on the mean was fitted much better using NSVA-selected genes than the all genes in the local regression.

315 Figure 2 illustrated the means and raw variances of seven liver samples in the *Marioni* data, and the fitting of the raw variances with respect to the means (red lines) using all 15,514 genes and only 2000 NSVA-selected genes, in the NW and NE plots respectively. The similar results can be also found for the *Fly embryos* data with 4 samples but more than 17000 genes, in the SW and SE plots of Figure 2. It is clear that the  
320 fits under the NSVA-selected genes are much better than those under all genes. More importantly, such a good fitting contributes to more accurate variance parameter estimation in DESeq, which will enhance the accuracy of the following hypothesis test in the differential expression call.

Last, NSVA feature selection contributed to decreasing the complexity of the hy-

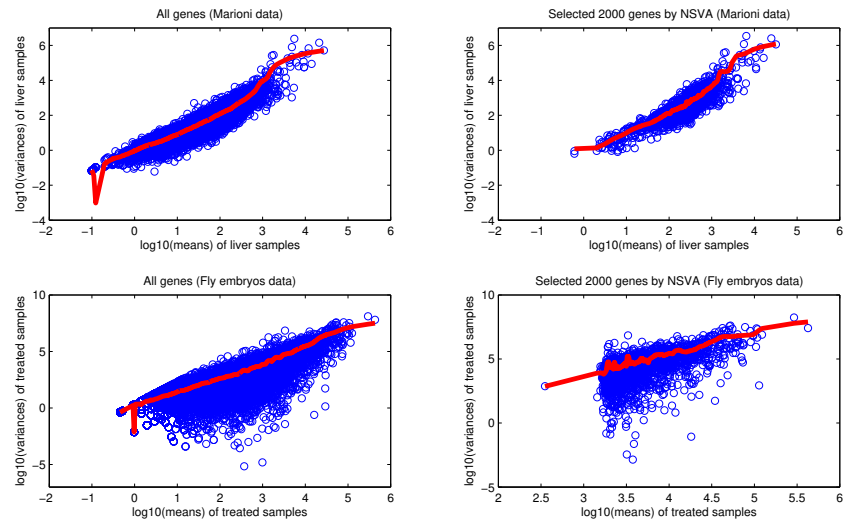


Figure 2: The plots of the sample variances with respect to the means of 7 *Liver* samples of the *Marioni* dataset under all 15,514 genes and NSVA-selected 2000 genes and the similar plots for the *Fly embryos* data. The red lines represent the fits of raw variances with respect to the means of the *Liver* (*Fly embryos*) samples respectively.

325 pothesis test in DESeq due to the fact that a lot of genes were filtered by NSVA, which  
actually avoids quite a lot computing burden because calculating the p-value for each  
gene in DESeq requires to enumerate all possible count sum combinations of two con-  
ditions (e.g., treated vs untreated) from a given total count sum [8].

### 3.3. Compare nonnegative singular value approximation with peer methods

330 We further compare our NSVA feature selection with other peer methods to demon-  
strate its advantage in picking potential DE genes. These methods include count-based  
naive feature selection (NFS), principal component analysis (PCA), nonnegative ma-  
trix factorization (NMF), signal-noise-ratio (SNR), and geometric signal-noise-ratio  
(GSNR) [20, 19, 31, 32, 33]. All the five comparison methods are data-driven methods  
335 as NSVA.

In fact, the count-based naive feature selection (NFS) is just the widely used low-  
count filtering method. PCA and NMF both belong to variance-based feature selection

methods as NSVA though they use different variance metrics in feature selection. SNR and GSNR belong to signal-noise based feature selection method that rank each gene  
340 via comparing signal-noise ratios.

Count-based naive feature selection (NFS). As a coverage-based feature selection method, NFS filters the genes with low counts and keeps those with high counts before D.E. analysis. It selects all genes  $\geq$  the median gene count of the input data, sorts all genes according to its coverage, i.e., the sum of its counts, and picks the top-ranked genes  
345 before D.E. analysis [8, 13].

### 3.3.1. Principal component analysis (PCA).

As a variance-based feature selection method, PCA ranks each gene by using the 2-norm of its projection in the subspace spanned by the first three principal components (PCs) by measuring the gene's contribution to the three most significant PCs [19, 32].  
350 It is noted that the three major PCs usually have a quite high variance explained ratio (e.g., 99%) for each dataset. The PCA feature selection consists of the following three steps.

The first step conducts PCA for input data  $X \in \mathbb{R}^{n \times p}$  to obtain the principal component (PC) matrix:  $U \leftarrow \text{princomp}(X)$ , and projected data to the first three PCs, i.e.  
355  $P \leftarrow X * U(:, 1 : 3)$ ,  $X^* = X - \frac{1}{n} \times \vec{1} \times (\vec{1})^T \times X$ , where  $\vec{1} \in \mathbb{R}^n$  is a vector with all entries '1'. The second step calculates the 2-norm for the projection data of each gene in the subspace spanned by the three PCs:  $\tau_i = (\sum_{j=1}^3 p_{ij}^2)^{1/2}$ , where  $p_{ij}$  is the  $i^{\text{th}}$  row of the projection matrix  $P \in \mathbb{R}^{n \times 3}$ ,  $i = 1, 2 \dots n$ . Finally, the third step sorts the genes according to  $\tau_i$  and selects the top-ranked genes.

### 360 3.3.2. Nonnegative matrix factorization (NMF)

Similar to PCA, NMF is a variance-based feature selection but requires the non-negativity of input data [20]. Given an input RNA-seq data  $X \geq 0 \in \mathbb{R}^{n \times p}$ , NMF conducted the following decomposition:  $X \sim WH$  at rank  $p - 1$  firstly, where  $W = [w_1, w_2 \dots w_n]^T$ ,  $w_i \in \mathbb{R}^{p-1}$ ,  $i = 1, 2, \dots n$ , and  $H \in \mathbb{R}^{(p-1) \times p}$ . Then,  $\|w_i\|_2$   
365 is used to rank the  $i^{\text{th}}$  gene's contribution to the whole data variance. Finally, the top-ranked genes were selected by sorting the values of  $\|w_i\|_2$ .

### 3.3.3. Signal-Noise ratio (SNR) and geometric signal-noise ratio (GSNR)

This data-derive feature selection method ranks each gene by the ratio of gene mean and standard deviation:  $\tau_i = \frac{\mu_i}{\sigma_i}$ , where  $\mu_i$  and  $\sigma_i$  are estimated as  $(\sum_{j=1}^p X_{ij})/p$  and  $\frac{\sum_{j=1}^p (X_{ij} - \mu_i)^2}{p-1}$  respectively for given  $X \in \mathfrak{R}^{n \times p}$ . These genes with large SNR values are believed to be more meaningful genes. It is noted that the infinite SNR values are set as zeros automatically in our feature selection [31].

Different to the SNR feature selection, this method ranks each gene by using the geometric signal-noise ratio to rank each gene. GSNR is defined as the ratio of the geometric mean and geometric standard deviation as  $\tau_i^{(g)} = \frac{\mu_i^{(g)}}{\sigma_i^{(g)}}$ , where  $\mu_i^{(g)} = (X_{i1}X_{i2} \cdots X_{ip})^{1/p}$  and  $\sigma_i^{(g)} = \exp(\sqrt{\frac{\sum_{i=1}^p (\ln A_i / \mu_i^{(g)})^2}{p}})$  respectively for given  $X \in \mathfrak{R}^{n \times p}$ . These genes with large  $\tau_i^{(g)}$  values are believed to be more meaningful genes in feature selection [33].

### 3.3.4. NSVA is robust to depth and gene-length biases

We need to answer the following two questions: 1) Is NSVA more efficient than its peers in identifying potential DE genes? 2) Is NSVA a depth-dependent feature selection method, where high-count genes are more likely to be identified as DE genes?

To answer the queries, we compare proposed NSVA with its peers on two measures under DESeq analysis: *DE ratios* and *DE gene median counts*. The DE ratio refers to the ratio between the number of DE genes identified by a differential expression analysis  $\Omega$ , which is employed as DESeq analysis in this context, and the total number of genes:  $N$  of the input data, where  $\varepsilon$  is a significant level cutoff (e.g. 0.01) and  $\theta$  is a feature selection method employed before differential expression analysis, namely,

$$\eta = \frac{|\{g : (g.pvalue|\Omega, \theta) < \varepsilon\}|}{N} \quad (11)$$

It measures the efficiency of a feature selection method. An efficient feature selection method  $\theta$  should produce a high DE ratio for a dataset under a specified DE analysis  $\Omega$ . The DE gene median count  $\tau$  is the read count median among all DE genes

$$\tau = median\{g.count : (g.pvalue|\Omega, \theta) < \varepsilon\} \quad (12)$$

under  $\Omega$  and  $\theta$ . A feature selection method would be a depth-dependent method, provided it had high median counts for DE genes.

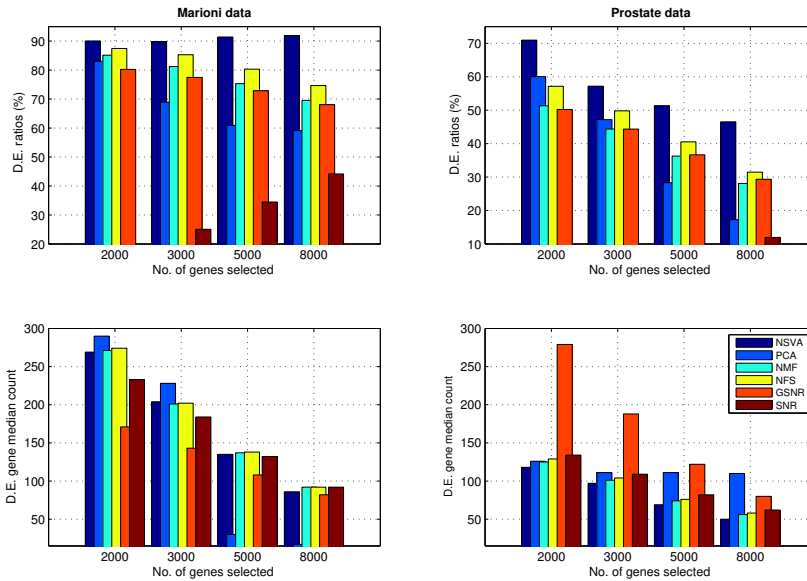


Figure 3: The comparisons of DE ratios and DE gene median counts for NSVA and its peers under DESeq analysis on two datasets. The proposed NSVA feature selection demonstrates strong advantages in selecting potential DE genes than its competing methods by producing highest DE ratios. The DE gene median counts of NSVA are generally lower than those of other peers for the two datasets.

Figure 3 demonstrates the DE ratios and DE gene median counts of proposed NSVA and its five peers on the *Marioni* and *Prostate* data, when 2000, 3000, 5000, and 8000 genes are selected in feature selection [7, 34]. Interestingly, the results suggest that NSVA is a more efficient method compared with its peers. It achieves the highest DE ratios for each case among all feature selection methods. The NFS feature selection performs a little bit better than NMF, PCA and GSNR. It indicates that complicate transform-based feature selection methods (e.g. NMF) may not contribute to DE analysis. SNR has the worst DE ratios. It indicates that simple feature selection methods like SNR can not contribute to DE analysis either.

In addition, NSVA has the shortest median DE gene count values than all the other methods for *Prostate* data. On the other hand, the DE gene median counts of NSVA are generally lower than those of PCA and NFS, equivalent to that of NMF, and higher than those of GSNR and SNR for the *Marioni* data. Those results strongly suggest that  
405 NSVA should not be a depth-dependent feature selection method like NFS.

*Does NSVA only pick long genes in feature selection?* That is, NSVA can contribute to avoiding gene-length bias in RNA-seq analysis? To answer this query, Figure 4 compares the gene length medians of the genes selected by the six feature selection methods and the DE genes among the selected ones for the *Marioni* data. The other  
410 two datasets have no gene length information available.

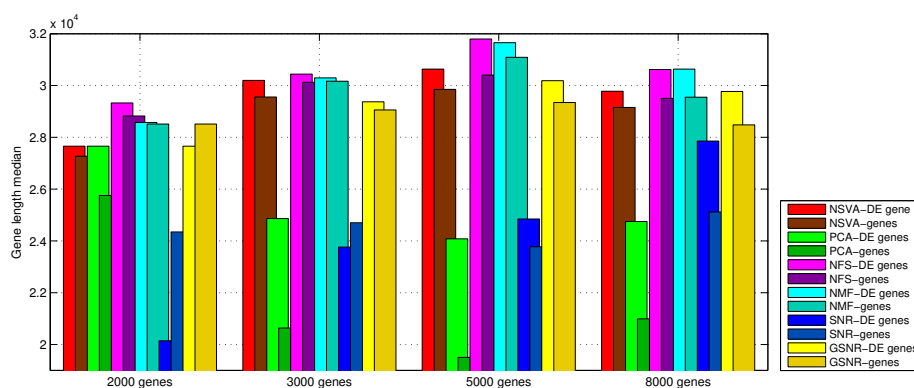


Figure 4: The comparisons of the gene length medians of the genes selected by different feature selection methods and DE genes among the selected genes for the *Marioni* data. The DE genes generally have longer gene length than those selected genes from almost all methods. The NSVA-selected genes and their DE genes are shorter than those from their peers like NFS, and NMF.

It is interesting to see that the DE genes generally have longer gene length than those selected genes from almost all methods except SNR. Such a result is consistent with the consensus that long genes are more likely to be selected as DE genes in RNA-seq D.E. analysis. Furthermore, NSVA-selected genes and its DE genes are shorter than  
415 those from NFS and NMF, but longer than those from PCA and SNR. For example, The median gene lengths of NSVA-selected genes in all the four gene-selection cases are

higher than the DE gene median length (26,445 bp) of all genes for the *Marioni* data [7]. Furthermore, its DE gene median length has reached 27,659 bp on the 2,000 gene selection case, which is much lower than that of NFS (29,328 bp) and NMF (28, 513 bp). GSNR has an almost same level gene median length as NSVA, but it has relatively lower DE ratios than NSVA.

Table 2 compares the six methods on behalf of DE ratios, DE gene count median, gene-median-length and DE gene-median-length. The gene-median-length and DE gene-median-length refer to the gene median length of NSVA-selected genes and DE genes among the NSVA-selected genes respectively. It is clear that NSVA is more efficient than its peers in identifying potential DE genes for its highest DE ratios. NSVA also demonstrates it is not a depth-dependent feature selection method, where high-count genes are more likely to be identified as DE genes for its low DE gene-count-medians. Furthermore, NSVA selects 'median-long' genes instead of only long genes or short genes from the gene median length of the NSVA-selected genes and DE genes among them.

As such, NSVA seems to be the best one among the six feature selection methods for its robustness in depth bias and gene length biases. It not only avoids only picking those long genes or high-count genes like NFS/NMF, but also the short genes or low-count genes as PCA/SNR by considering its high DE ratios and low DE gene medians counts. That is, it can contribute to picking potentially DE genes and decreasing the false positives in DE analysis.

**Table 2** The comparisons of six feature selection methods

Methods	DE ratios	DE gene count median	gene-median-length	DE gene-median-length
<i>NSVA</i>	highest	low	middle	middle
<i>PCA</i>	low	low	short	short
<i>SNR</i>	lowest	low	short	short
<i>GSNR</i>	high	instable	long	long
<i>NFS</i>	high	low	long	long
<i>NMF</i>	middle	instable	long	long



### 3.4. Nonnegative singular value approximation for non-parametric D.E. models

We further apply NSVA to non-parametric D.E. method NOISEq to demonstrate its effectiveness in differential expression analysis. The NOISEq employs two statistics:  $M$  and  $D$  to compare these to the noise distribution to determine whether the expression is statistically significant. The  $M$  and  $D$  values measure the  $\log_2$  fold change and  $\log_2$  absolute expression difference between conditions. From this comparison, NOISEq produces the probability value of their odd ratio that, when compared to a threshold number ( $q$ ), which is set as 0.8 in our experiment, determines whether the gene is actually differentially expressed [11].

*Applying NSVA to NOISEq.* Like DESeq, NOISEq demonstrates the increase of DE ratios in D.E. analysis when using NSVA feature selection [8, 11]. Figure 5 illustrates the scatterplots of plotted ( $M, D$ ) values ( $M$ - $D$  plots) produced from our NOISEq method for all the datasets. The  $M$ - $D$  plot is essentially to impose an  $M$ -plot on a  $D$ -plot, which is similar to volcano plot in traditional microarray analysis [35].

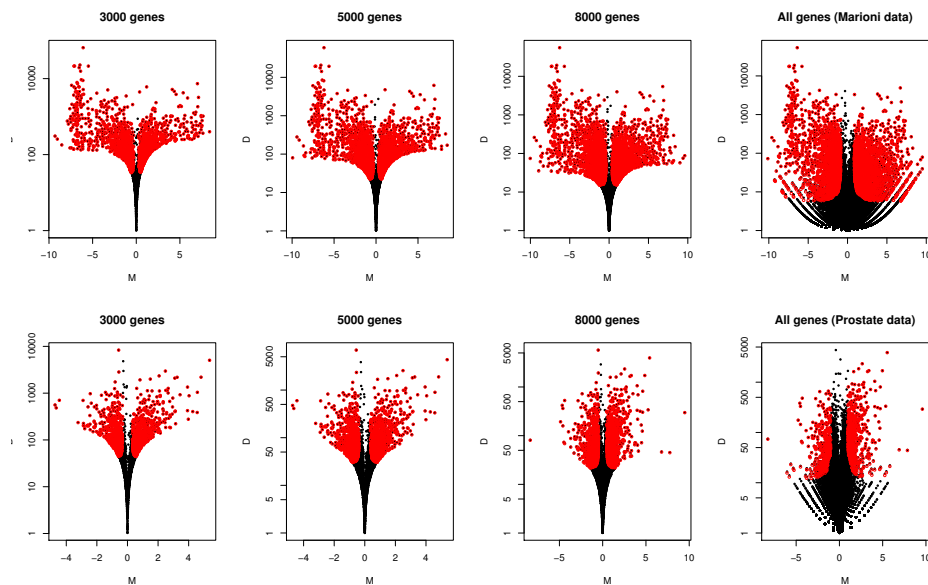


Figure 5: The comparisons of the  $M$ - $D$  plots of NSVA-selected genes and the original *Marioni* and *Prostate* data in *NOISEq* analysis, where DE and non-DE genes are indicated by red and black dots respectively. The non-DE genes drop remarkably in *NOISEq* when NSVA is applied to filter more genes

The plots compare the M-D plots of the top 3000, 5000, and 8000 NSVA-selected genes, along with the original dataset under NOISeq respectively. The red and black dots represent differentially expressed genes, and non-differentially expressed genes. It is clear that non-DE genes drop remarkably when more genes are filtered by NSVA for the two datasets. In other words, the corresponding DE ratio would increase for each selected gene set under such a feature selection. Obviously, such a result is consistent to the previous results from applying NSVA to the parametric model: DESeq [8]. It further indicates that such a feature selection can enhance D.E. analysis by picking meaningful genes for both parametric and non-parametric D.E. analysis models.

**Table 3** DE ratios of NSVA-selected datasets and original data under NOISeq

Selected genes	DE ratios ( <i>Fly</i> data)	DE ratios ( <i>Prostate</i> data)	DE ratios ( <i>Marioni</i> data)
2000 genes	2.4%	54.45%	81.90%
3000 genes	1.87%	47.27%	78.27%
5000 genes	1.04%	37.40%	72.56%
Total data	0.32%	22.92%	17.52%

Table 3 compares the DE ratios of NSVA-selected datasets and original data under NOISeq. It is clear to see that DE ratios increase for all three datasets when more genes are filtered in NSVA-feature selection. For example, when the 2000 most significant genes are selected from the *Marioni* dataset, 81.9% of those genes are determined to be differentially expressed. But the DE ratios of the original dataset without feature selection has only 17.52%. On the other hand, the DE ratio of the original *Fly* dataset is only 0.32%, but such a ratio reaches 2.4% when the 2000 most significant genes are selected in feature selection.

We also conduct naive feature selection (NFS) for NOISeq by removing all genes with count < 10. However, it can't achieve good DE ratios as NSVA. For example, the DE ratio is only 42.92% for 5606 NFS-selected genes for the *Marioni* dataset, but the DE ratio under the 5000 NSVA-selected genes is 72.56%. In addition, the DE ratio is only 30.93% for 1809 NFS-selected genes for the *Prostate* dataset, but the DE ratio under the 2000 NSVA-selected genes is 54.45%. Such results again indicate

the proposed feature selection performs better than naive feature selection (NFS) in  
 485 selecting meaningful genes.

### 3.5. Compare NSVA-seq with peer D.E. analysis models

To demonstrate the effectiveness of proposed NSVA-seq, we apply it to the gene  
 set consisting of top 10% genes ranked by NSVA from each dataset normalized by  
 DESeq normalization [8][10]. Then, we compare its adjusted p-value distributions  
 490 with those of four peer methods: NSVA-DESeq, NSVA-edgeR and NSVA-NOISeq and  
 mFET, where mFET is applied to the original normalized data. It should be noted that  
 the notations NSVA-DESeq/edgeR/NOISeq refers to applying DESeq/edgeR/NOISeq  
 analysis to the NSVA-selected genes respectively. We employ Benjamini-Hochbert  
 (BH) procedure to adjust all p-values under a FDR 0.01 in such a comparison.

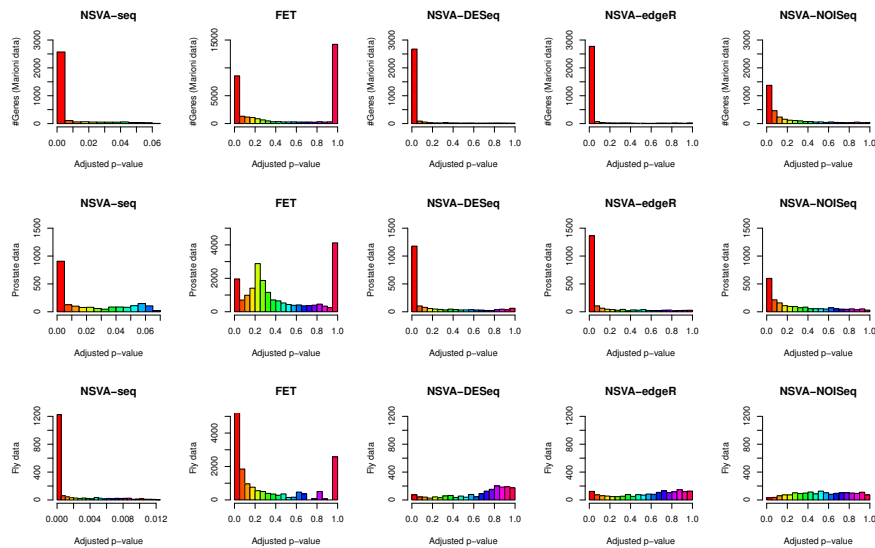


Figure 6: The comparisons of the adjusted p-value distributions of NSVA-Seq, mFET, NSVA-DESeq, NSVA-edgeR, and NSVA-NOISeq. NSVA-seq demonstrates a conservative D.E. analysis for the *Marioni* and *Prostate* datasets. But it overcomes the weakness of its peers in D.E. analysis of the *Fly embryos* dataset with few samples.

495

Figure 6 illustrates the scale of the adjusted p-values from NSVA-seq is in a quite  
 small range compared with those of the others. Such a result strongly suggests that

NSVA-seq is more sensitive to identify those genes with quite small adjusted p-values in differential expression analysis than the others. For example, almost all adjusted  
500 p-values are less than 0.03 that indicates these identified DE genes have a strong p-value support. In contrast, mFET generates a large amount of non-DE genes without NSVA feature selection for each dataset. It implies that NSVA tends to pick DE genes in feature selection, which directly contributes to the high DE ratios of NSVA-seq. As such, NSVA-seq has a much lower false positive ratio than mFET.

505 On the other hand, NSVA-DESeq and NSVA-edgeR have similar distribution patterns due to the same underlying assumption on the count data distribution of the DESeq and edgeR models. It is even hard to claim the advantage of DESeq than edgeR in D.E. analysis under NSVA-feature selection [8, 13, 15]. Such a result further implies that power of NSVA in selecting potential DE genes.

510 Interestingly, NSVA-seq seems to be more conservative in D.E. analysis than NSVA-DESeq and NSVA-edgeR for the *Prostate* and *Marioni* datasets. However, it actually identifies more DE genes for the *Fly embryos* dataset that has only 4 samples than the other methods, which seem to identify almost all genes as non-DE genes. This is because the datasets with few samples have some disadvantage in estimating accurate mean and variance parameters parameters for parametric D.E. analysis models like  
515 DESeq and edgeR [8, 13]. For example, a dataset with few samples may cause difficulties for the local fit procedure in the DESeq model [8, 15]. On the other hand, too small sample size can lead to low odd-ratios of  $M$  and  $D$  in NOISeq due to the lack of replicates and the likelihood to miscount meaningful expression signals as noise [11].  
520 However, our NSVA-seq avoids parameter estimation or  $M$ - $D$  odd ratio comparisons by comparing each gene's average expression with a set of selected genes' average expressions. It mitigates the side-effect from the small number of samples and provides a fair D.E. analysis environment for RNA-seq datasets. It is worthwhile to point out that similar results can be obtained when TMM is employed as the procedure in data  
525 normalization rather than DESeq normalization [8, 9].

Furthermore, NSVA-seq has a more transparent D.E. analysis mechanism than NOISeq though both are nonparametric D.E. analysis models. Its modified FET based D.E. analysis under NSVA-feature selection is more direct than NOIseq that relies on

M-D odd ratio comparisons [9]. However, NSVA-seq demonstrates advantages in over-  
530 coming the weakness of the existing D.E. models in handling datasets with few sam-  
ples, besides more conservative D.E. analysis results for other datasets. Such a charac-  
teristic can be essentially useful for clinical D.E. analysis, in which no enough samples  
are generally available [9, 13, 15].

### 3.5.1. *Nonnegative singular value approximation based biomarker discovery, a case 535 study*

We further demonstrate the effectiveness of NSVA in biomarker discovery from  
by using jActiveModule to search active subnetwork modules [37]. We use the top-  
ranked 2000 genes with smallest probability values under NSVA-DESeq to find possi-  
ble biomarkers for the *Prostate* data, in which input dataset consisting of 5000 genes  
540 selected by NSVA from the original *Prostate* data. We have found that there are several  
networks with varying active path scores of: 4.97, 5.22, 5.24, 5.29, and 5.98. We use  
the module with the highest score: 5.98, as our network marker that has 179 nodes and  
630 edges.

Figure 7 illustrates the network marker by high-lightening those genes with most  
545 protein-protein interactions. Although detailed analysis of such a network marker is  
beyond the scope this study, we would like to analyze the genes with the largest inter-  
actions in the network marker. YWHAE, TARDBP, and CALM1 are the three genes  
with the most interactions among the network marker. It is interesting to see that all  
of them have much closer relationships with prostate cancer. For example, YWHAE  
550 belongs to the 14-3-3 family of proteins which mediate signal transduction by binding  
to phosphoserine-containing proteins and has been reported to have express in prostate  
cancer [38]. Furthermore, it was reported to interact with YWHAZ, a widely known  
biomarker of prostate cancer [39, 40]. In addition, TARDBP has been found to have  
multiple functions in transcriptional repression, pre-mRNA splicing and translational  
555 regulation. It was reported as one of biomarkers to distinguish prostate cancer from  
benign prostatic hyperplasia in patients[41]. Moreover, CALM1's mutation was re-  
ported to connect with prostate cancer and was one of verifiable biomarkers of prostate  
analysis using urinary shotgun proteomics [42, 43]. Such a meaningful biomarker cap-

turing indicates the usefulness of our network marker though more detailed analysis  
560 can be done for this network marker to retrieve more comprehensive information. Al-  
ternatively, it demonstrate the effectiveness of NSVA in D.E. analysis and biomarker  
discovery.

## Conclusion

In this study, we propose a novel NSVA feature selection and NSVA-seq differ-  
565 ential expression analysis method for RNA-seq data. The NSVA feature selection is  
rooted in a rigorous mathematical result from singular value decomposition for non-  
negative RNA-seq read count data. The proposed NSVA-based feature selection al-  
gorithm demonstrates robustness to depth and gene length robustness by overcoming  
the weakness of widely used naive count feature selection (NFS). It demonstrates ad-  
570 vantages in picking meaningful potential DE genes for different D.E. analysis models  
by enhancing the efficiency of D.E. analysis by comparing with its five peer feature  
selection methods.

As a data-driven D.E. analysis, NSVA-seq provides more freedom in D.E. analy-  
sis by allowing both original count data and normalization data in D.E. analysis. It  
575 not only avoids the parameter estimation process, but also provides a more direct and  
transparent nonparametric D.E. analysis, which contributes to easy understanding and  
implementation. More importantly, it overcomes the limitations of the existing D.E.  
analysis models by providing a fair D.E. analysis for those datasets with few samples  
besides producing a relatively conservative D.E. analysis for the other datasets. Fur-  
580 thermore, the biomarker discovery results demonstrate the effectiveness of NSVA in  
capturing meaningful genes, and its positive impacts on D.E. analysis and meaningful  
gene marker capturing for complex diseases.

However, how to achieve an optimal feature selection for the sake of robust dif-  
ferential expression analysis remains a challenge for this method. We are employing  
585 information measures such as mutual information or entropy to explore its potential in  
NSVA feature selection [44]. Moreover, we are interested in conducting novel path-  
way analysis for the network marker obtained in this study to dig more knowledge and

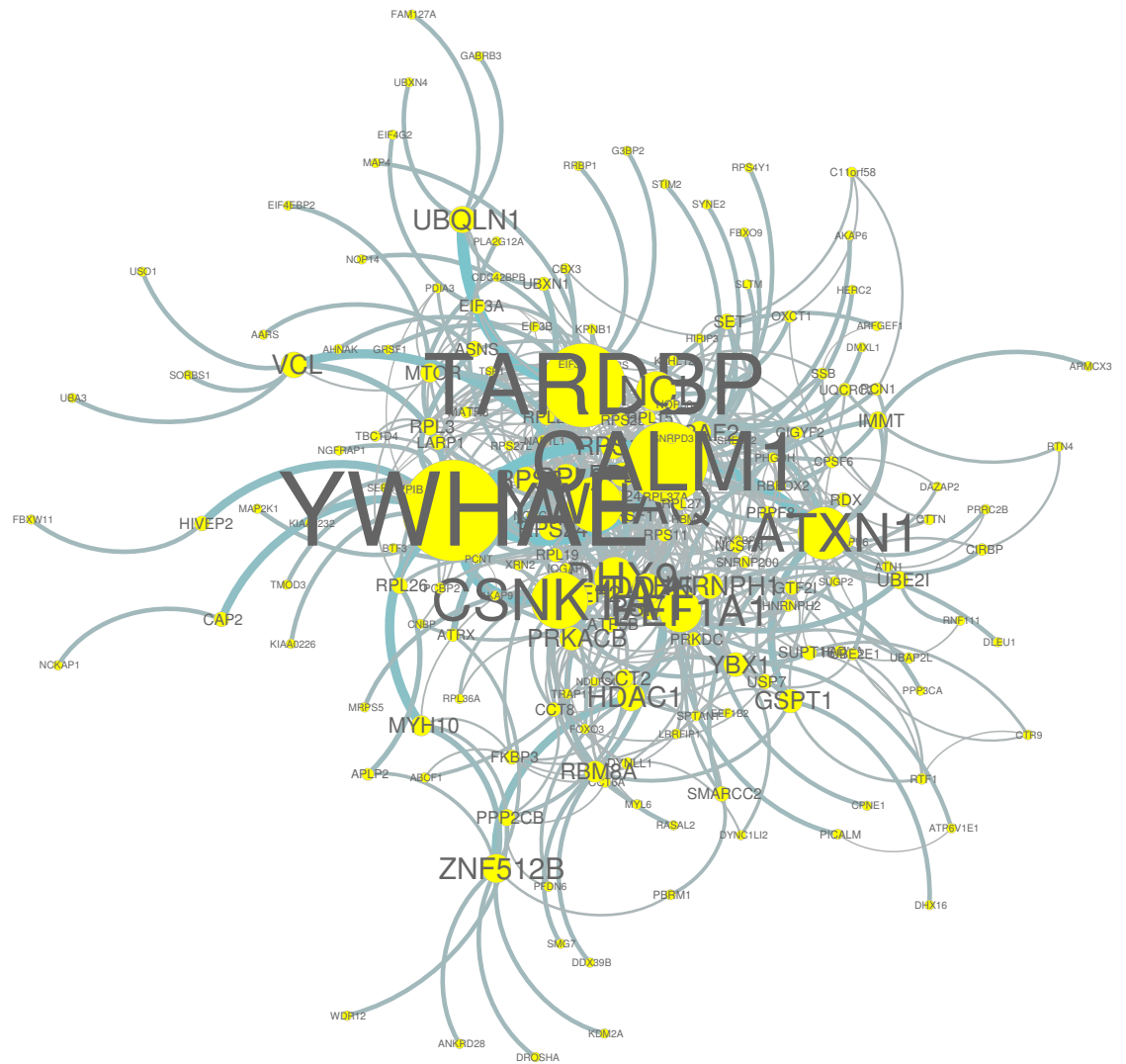


Figure 7: The network marker of the *Prostate* dataset based on the 2000 most significant genes from NSVA-DESeq. The top three gene with most interaction are YWHAE, TARDBP, an CALM1.

further enhance its repeatability and validity [45].

In addition, we are applying NSVA and NSVA-seq to RNA-seq datasets retrieved  
590 from TCGA portal, which are a type of structured big data, to further investigate the  
effectiveness of our methods [46, 47]. Those datasets can be high-dimensional imbal-  
anced data (HDI): high-dimensional data with skewed label distributions. They usually  
bring hard time in disease diagnosis when there is no feature selection done [48]. We  
are interested in investigating the impacts of NSVA feature selection on such data to  
595 further explore its potential in disease diagnosis.

### Acknowledgement

Author sincerely thanks Dr. Wentian Li for his invaluable suggestions and com-  
ments for this work.

### References

- 600 [1] Sara A. Byron, Kendall R. Van Keuren-Jensen, David M. Engelthaler, John D.  
Carpenter & David W. Craig. 2016. Translating RNA sequencing into clinical diag-  
nostics: opportunities and challenges, *Nat Rev Genet.* 17(5):257-71
- [2] Conesa et al. 2016. A survey of best practices for RNA-seq data analysis, *Genome  
Biology* 17:13
- 605 [3] Wang et al. 2009. RNA-Seq: a revolutionary tool for transcriptomics, *Nat Rev  
Genet.* 10(1): 5763
- [4] Xie *et al.* 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short  
RNA-seq reads. *Bioinformatics.* 30(12):1660-6
- [5] Langmead *et al.* 2009. Ultrafast and memory-efficient alignment of short DNA  
610 sequences to the human genome, *Genome Biology* 2009, 10:R25
- [6] Luo, R *et al.* 2013. SOAP3-dp: Fast, Accurate and Sensitive GPU-Based Short  
Read Aligner, *PLoS One.*



- [7] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18(9):1509–1517
- 615
- [8] Anders S, Huber W., 2010. Differential expression analysis for sequence count data. *Genome Biol* 2010, 11:R106.
- [9] Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11:R25
- [10] Dillies et al: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, *Briefings in Bioinformatics* 2012
- 620
- [11] Tarazona et al.: Differential expression in RNA-seq: a matter of depth. *Genome Res* 2011, 21:2213–2223.
- [12] Li J, Tibshirani R.: Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data, *Stat Methods Med Res.* 2013, 22(5):519-36
- 625
- [13] Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010 26: 139-140.
- [14] Hardcastle TJ, Kelly KA: baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 2010, 11:422.
- 630
- [15] Rapaport et al: Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data, *Genome Biology* 2013, 14:R95
- [16] Li, W., Yang, Y. 2002. How Many Genes Are Needed for a Discriminant Microarray Data Analysis? *Methods of Microarray Data Analysis*, Kluwer Academic, 137-150
- 635
- [17] Han, X. 2006. Inferring Species Phylogenies: A Microarray Approach, *Computational Intelligence and Bioinformatics, Lecture Notes in Bioinformatics (LNBI)* 4115, 485493, Springer-Verlag Berlin Heidelberg

- 640 [18] Fan, *et al.* 2010. Nonparametric estimation of genewise variance for microarray data, *The Annals of Statistics*. 38(5): 2723-2750.
- [19] Jolliffe, I.: Principal component analysis, Springer, New York 2002
- [20] Lee, D. and Seung, H.: Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999 401: 788-791
- 645 [21] Han, X. 2007. Cancer Molecular Pattern Discovery by Subspace Kernel Classification. Proceeding of Computational System Bioinformatics (CSB), V6, 55-66
- [22] Han, X. 2010. Improving Gene Expression Cancer Molecular Pattern Discovery using Nonnegative Principal Component Analysis, *Genome Informatics*, 21, 200-211
- 650 [23] Nguyen and Rocke. 2002. Tumor classification by partial least squares using microarray gene expression data, , *Bioinformatics*, 18:39-50
- [24] Han, H: Derivative Component Analysis for Mass Spectral Serum Proteomic Profiles, *BMC Medical Genomics*, Vol 7:S1 2014
- [25] Lax, P. 1997. Linear Algebra, John Wiley & Sons, 1997
- 655 [26] Brin, S., Page, L. 2000. The Anatomy of a Large-Scale Hypertextual Web Search Engine., <http://infolab.stanford.edu/backrub/google.html>
- [27] Golub and Van Loan 1996. Matrix Computations, The Johns Hopkins University Press
- [28] Choi L, Blume JD, Dupont WD. 2015. Elucidating the Foundations of Statistical  
660 Inference with 2 x 2 Tables. *PLoS ONE*. 10 (4): e0121263
- [29] Agresti, A. 1992. A Survey of Exact Inference for Contingency Tables. *Statistical Science*. 7 (1): 131-153.
- [30] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup.

- 665 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-9.
- [31] Venet D, Detours V, Bersini H. 2012. A Measure of the Signal-to-Noise Ratio of Microarray Samples and Studies Using Gene Correlations. *PLoS ONE* 7(12): e51013. <https://doi.org/10.1371/journal.pone.0051013>
- 670 [32] Han, X. 2010. Nonnegative Principal component Analysis for Cancer Molecular Pattern Discovery, *IEEE/ACM Transaction of Computational Biology and Bioinformatics* Vol 7 (3), 537-549
- [33] Gonzalez JG, Paredes JL, Arce GR. 2006. Zero-order statistics: a mathematical framework for the processing and characterization of very impulsive signals, *IEEE*  
675 *Trans. on Signal Processing*, 54:38393851.
- [34] Li, H., Lovci, et al. 2010. Determination of Tag Density Required for Digital Transcriptome Analysis: Application to an Androgen-Sensitive Prostate Cancer Model. *PNAS*, 105(51), 20179-20184.
- [35] Li, W. 2012. Volcano plots in analyzing differential expressions with mRNA microarrays, *J. Bioinform. Comput. Biol.*, 10, 1231003  
680
- [36] Benjamini, Y. and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.
- [37] Ideker, T., et al. (2002) Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics*, 18 Suppl 1, S233-240.
- 685 [38] YWHAE: <https://www.proteinatlas.org/ENSG00000108953-YWHAE/>, retrieved 2017
- [39] Murata T, Takayama K, Urano T, Fujimura T, Ashikari D, et al. (2012) 14-3-3zeta, a novel androgen-responsive gene, is upregulated in prostate cancer and promotes prostate cancer cell proliferation and survival. *Clin Cancer Res* 18: 5617–5627.
- 690 [40] Kasinski et al. (2014) Transcriptional Regulation of YWHAZ, the Gene Encoding 14-3-3. *PLoS ONE*9(4): e93480. <https://doi.org/10.1371/journal.pone.0093480>

- [41] O'Rourke *et al.* 2011 Autoantibody signatures as biomarkers to distinguish prostate cancer from benign prostatic hyperplasia in patients with increased serum prostate specific antigen, *Clinica Chimica Acta*, Volume 413, 5(6), 22, 561-567
- 695 [42] Marsman, *et al.* 2014 A Mutation in CALM1 Encoding Calmodulin in Familial Idiopathic Ventricular Fibrillation in Childhood and Adolescence, *Journal of the American College of Cardiology*, Volume 63(3): 28 259-266
- [43] Adeola et al. 2016 Novel potential serological prostate cancer biomarkers using CT100+ cancer antigen microarray platform in a multi-cultural South African cohort, *Oncotarget*. 2016 Mar 22; 7(12): 13945–13964.
- 700 [44] Cilibrasi, et al. 2005. Clustering by compression. *IEEE Transactions on Information Theory*. 51 (4): 1523–1545
- [45] Garca-Campos, et al. 2015. Pathway analysis: State of the art. *Frontiers in Physiology*. 6:383
- 705 [46] Zhang Q et al. 2014. Integrative network analysis of TCGA data for ovarian cancer, *BMC Systems Biology* 8:1338
- [47] Han H, Liu Y. 2016. Transcriptome marker diagnostics using big data. *IET Systems Biology*. 10: 41-48
- [48] Han, H. 2015. Diagnostic biases in translational bioinformatics, *BMC Medical Genomics*, 8:46
- 710