Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computational Advances in Cancer Informatics (A)

Disease Biomarker Query from RNA-Seq Data

Henry Han[1,2] and Xiaoqian Jiang[3]

[1]Department of Computer and Information Science, Fordham University, New York, NY, USA. [2]Quantitative Proteomics Center, Columbia University, New York, NY, USA. [3]Division of Biomedical Informatics, University of California, San Diego, CA, USA.

ABSTRACT: As a revolutionary way to unveil transcription, RNA-Seq technologies are challenging bioinformatics for its large data volumes and complexities. A large number of computational models have been proposed for differential expression (DE) analysis and normalization from different standing points. However, there were no studies available yet to conduct disease biomarker discovery for this type of high-resolution digital gene expression data, which will actually be essential to explore its potential in clinical bioinformatics. Although there were many biomarker discovery algorithms available in traditional omics communities, they cannot be applied to RNA-Seq count data to seek biomarkers directly for its special characteristics. In this work, we have presented a biomarker discovery algorithm, SEQ-Marker for RNA-Seq data, which is built on a novel data-driven feature selection algorithm, nonnegative singular value approximation (NSVA), which contributes to the robustness and sensitivity of the following DE analysis by taking advantages of the built-in characteristics of RNA-Seq count data. As a biomarker discovery algorithm built on network marker topology, the proposed SEQ-Marker not only bridges transcriptomics and systems biology but also contributes to clinical diagnostics.

KEYWORDS: RNA-Seq, feature selection, biomarker discovery

SUPPLEMENT: Computational Advances in Cancer Informatics (A)

CITATION: Han and Jiang. Disease Biomarker Query from RNA-Seq Data. *Cancer Informatics* 2014:13(S1) 81–94 doi: 10.4137/CIN.S13876.

RECEIVED: March 25, 2014. **RESUBMITTED:** June 18, 2014. **ACCEPTED FOR PUBLICATION:** June 18, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: XJ is partially supported by iDASH (grant U54HL108460) and the NIH (grant R00LM011392). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: xhan9@fordham.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

RNA-Seq provides a revolutionary way to unveil transcriptome details by using ultra-high-throughput sequencing technologies to generate hundreds of million short reads from RNA molecules.[1] The short reads are a type of big data that usually need several or more gigabytes storage. They are usually aligned against a reference genome (eg, human genome) by using alignment tools such as *Bowtie*,[2] *SOAP3*,[3] or *Cufflinks*[4] to produce a genome-scale *transcription map* that consists of the expression level for all genes in transcription. The expression of each gene is represented by the number of short reads mapped to the gene in the alignment, which is believed to be linearly proportional to its abundance level in transcription.[1,5,6]

The transcription map can be represented by a nonnegative integer *read count matrix* $X \in \mathbb{Z}^{n \times p} \geq 0$, by collecting the number of short reads mapped to each gene across all samples. Each row and column in the matrix represents a gene and sample, respectively. The terminology gene actually refers to a more general biological feature such as a gene, exon, or transcript in context. Alternatively, each sample refers to a biological or technical replicate, where the biological replicates are all samples under a same biological condition and the technical replicates are alternative sequencing results of a same biological sample. Given an RNA-Seq count data (read count matrix), the number of variables (genes) is generally much greater than the number of samples (observations), ie, $n >> p$, where $n \sim 10^4$ and $P << 10^2$.

From a translational bioinformatics viewpoint, an essential issue in RNA-Seq data analysis is to answer the following two related queries. First, given a read count matrix, how to robustly determine whether the observed difference in read counts for a gene across two or more conditions is statistically significant? Second, how to retrieve disease biomarkers from RNA-Seq count data to provide a possible guide for disease diagnosis and prognosis? It is noted that we use terminologies "RNA-Seq data" and "RNA-Seq count data" interchangeably for the convenience of following description, unless there is a special notation.

Quite a few differential expression (DE) analysis methods have been proposed to answer the first query from different standing points.[7–12] They can be categorized as parametric and nonparametric approaches according to whether they rely on statistical parameter estimation modeling approaches. The parametric methods assume that read counts subject to a probability distribution and estimate corresponding parameters for the distribution before conducting a corresponding hypothesis test to rank genes.[7,8,10,12] For example, *DESeq* and *edgeR* methods both model read counts by a negative binomial (NB) distribution and employ a variation of Fisher's exact test to calculate *P*-values to rank each gene, whereas they estimate mean and variance parameters from different models.[7,12] Alternatively, the nonparametric methods, such as NOISeq, do not assume count data subject to any distribution.[13] Instead, their DE calls are based on an empirical distribution of some statistic derived from input data. For example, NOISeq determines differentially expressed (DE) genes by employing an odds ratio derived from two count statistics: log fold change and absolute expression difference.[13]

Unlike the first query, there was no previous work in the literature on RNA-Seq data biomarker discovery. However, compared with traditional microarray data, RNA-Seq data can provide a more reproducible, high-resolution digital expression for monitoring RNA transcription.[5] Especially, it makes each gene's expression in a single sample comparable with those of others.[7,14] On the other hand, it is almost impossible to compare the expression levels of genes within a sample for microarray data because of the strong background signals generated from the hybridization process of the microarray. Thus, it will be desirable to seek disease biomarker discovery from RNA-Seq count data for the sake of disease diagnosis and prognosis by taking advantage of these properties. However, RNA-Seq data biomarker discovery remains a challenging problem for the following major reasons.

First, the special characteristics of RNA-Seq count data present hurdles from reusing those biomarker discovery algorithms developed from traditional omics data, ie, microarray or proteomics data. For example, RNA-Seq count data have much fewer number of samples (eg, $P < 7$) compared with the traditional omics data, which challenges the effectiveness of the parameter estimation-oriented biomarker discovery methods (eg, Bayesian methods).[7]

Moreover, different from traditional omics data, which are usually normally distributed after normalization, RNA-Seq count data are usually modeled as the NB distribution or Poisson distribution.[7,14] The biomarker discovery methods developed under the normal distribution assumption usually cannot apply to RNA-Seq count data directly. In addition, most genes in RNA-Seq count data have quite good discriminative abilities under classification compared with microarray data. This is due to the built-in ultra-high resolution of RNA-Seq that leads to more accurate measurement and higher dynamic range for gene expression under different conditions.[1,15]

For example, we have found that the widely employed *filter–wrapper* biomarker discovery method in the microarray community[16] cannot work well for RNA-Seq count data. This is probably due to the fact that most genes have almost perfect performance under a classifier (eg, SVM),[17,18] with a leave one-out cross validation (LOOCV) in the wrapping process. In addition, there are no appropriate feature selection algorithms available for RNA-Seq count data in the filtering process, because most of widely used statistical testing-based feature selection methods in the filter-wrapper method assume that population data are normally distributed (eg, *t-test*), which cannot apply to RNA-Seq read count data.

Second, quite a lot biomarker discovery methods require an accurate *P*-value calculation to rank each gene in DE analysis. However, current DE analysis methods do not achieve it very well. Instead, quite a lot methods even suffer from high false positives in the *P*-value calculation with the increase of sequencing depth (SD).[13] A key reason is that all existing DE analysis methods usually invite almost all genes into DE calls without conducting a serious feature selection for high-dimensional RNA-Seq count data, although some of them simply filter genes with low count numbers before analysis. As such, the redundant or noise-contained genes will get involved and act as outliers in DE analysis, which inevitably leads to the increase of the false positive ratios in the hypothesis testing.

For example, some genes with low counts or same level high counts in two conditions are actually due to the artifacts of library preparation protocols, sequencing inaccuracy, or alignment imprecision, instead of reaction to treatment.[5,13] In fact, those genes, which do not have real contributions to data variations under a treatment, should not enter the DE call for the sake of the sensitivity of DE analysis, because they will distract the focus of DE analysis by increasing false positive ratios.

Moreover, almost all parametric DE analysis methods are SD-dependent methods[13] and they would falsely detect some non-DE gene as a DE gene, when the SD increases in a condition. That is, the count increases in the condition will be falsely diagnosed as a statistically significant DE change. Thus, a serious feature selection with an aim to remove the genes with no real contributions to data variations will contribute to the SD independence of these methods by decreasing false positive ratios in the DE call.

In this study, we presented a novel biomarker discovery algorithm: SEQ-Marker for RNA-Seq data from a network discovery standing point. The proposed SEQ-Marker algorithm is built on a proposed data-driven feature selection algorithm, nonnegative singular value approximation (NSVA), proposed in this study. As a feature selection algorithm to select genes according to its contribution to the whole data variance, it does not require any probability distribution assumption about count data. It also demonstrated a good consistency in identifying large variance genes (eg, long genes) as DE genes when integrated with classic DE analysis methods. Moreover, we compared our NSVA feature selection with two competing methods, count-based naive feature selection (NFS) and principal component analysis (PCA), to demonstrate its advantages in selecting meaningful genes.

Unlike the traditional biomarker discovery methods in the microarray community, the proposed SEQ-Marker employed a novel strategy to identify biomarkers from network markers. That is, it searched an inferred network marker at first for RNA-Seq count data and then identified meaningful biomarkers from the network marker by retrieving gene interaction and gene mutation information. The proposed biomarker discovery algorithm aimed at finding biomarkers by novelly viewing an inferred network marker as a small database. The database not only unveiled interaction information between genes but also made it possible to explore real disease biomarkers along an interaction "path". Such a search scheme is especially helpful to avoid the tissue-specific expression biomarkers and identify those real disease biomarkers, which may not express themselves in omics experiments.

It is worthwhile to point out that our biomarker search mechanism is specially designed according to the properties of the real disease biomarkers by answering the following queries: "which genes' mutations will affect most genes in the network marker?" and "which genes have the highest proximities with a most likely gene marker?". Obviously, such a biomarker search mechanism will have advantages to capture real disease markers by overcoming the weakness of traditional *P*-value-based tissue-specific expression biomarker discovery.

On the other hand, our proposed biomarker discovery model overcame the weakness of the traditional network marker and individual gene marker discovery. The former has been facing the difficulties in biomarker validation, because it could be prohibitive or impossible to conduct a biomarker validation for a network marker with more than hundreds of genes. The latter suffers from the poor reproducibility caused by adhocness of *P*-values, in addition to failing to provide information on significant molecular mechanism such as gene–gene interaction. In fact, the biomarkers from our model were convenient and less expensive for validation from a clinical viewpoint for its quantity. Alternatively, the proposed gene interaction-oriented search scheme in our model overcame the limitations of *P*-values and enabled the identification of biomarkers without strong *P*-values. To our knowledge, the proposed SEQ-Marker algorithm is the first work on RNA-Seq biomarker discovery that not only bridges transcriptomics and systems biology, but also contributes to clinical diagnosis.

Biomarker Discovery for RNA-Seq Count Data

As we pointed out before, quite a lot biomarker discovery methods on traditional omics data cannot be applied to RNA-Seq count data directly, because of the special characteristics of RNA-Seq count data. The question is "what kind of biomarker discovery algorithms of RNA-Seq count data should be?" We believe a desirable RNA-Seq biomarker discovery should satisfy at least the following criteria.

First, only a portion of meaningful genes instead of all genes should participate biomarker discovery, because a lot of genes are not informative enough to contribute to disease diagnosis. For example, some genes with very low variances in two conditions are actually not reaction to treatment but results of the artifacts of library preparations or relaxed alignment constraints. Alternatively, it is computationally expensive or even prohibitive to include all genes of a read count dataset in biomarker discovery. As such, we need a feature selection algorithm to filter the genes before starting biomarker discovery officially.

Second, a robust DE analysis model to accurately calculate *P*-values for each gene is needed in biomarker discovery. The DE analysis model should demonstrate capabilities to avoid high false positive rates in a conservative approach compared to its peers, no matter it is a parametric or nonparametric DE analysis model.

Third, the biomarker discovery algorithm should demonstrate some potential to overcome the weakness of traditional omics biomarker discovery methods by enhancing identified biomarkers' reproducibility and validation. For example, gene–gene (protein–protein) interaction information should be included in the biomarker discovery to improve the reproducibility of biomarkers. In particular, checking the interaction information of identified biomarkers would help to bridge the gap between biomedical research and clinical practice by identifying "real" or new markers.[19] This is because that some identified biomarkers with strong statistical support (eg, *P*-values) may not be found "useful" in real clinical diagnosis. In contrast, some well-known gene markers (eg, BRCA1 for breast cancer) widely employed in clinical practice may not be identified as biomarkers due to the complexity of disease, and limitations of existing technologies and mathematical modeling.[19] However, including gene–gene interaction information for biomarkers will contribute to fixing such a gap and improving the biomarkers' validation in clinical diagnostics. In this study, we presented a novel biomarker discovery algorithm, SEQ-Marker, for RNA-Seq count data to meet the three standards described previously. Our proposed SEQ-Marker has the following main components: a proposed data-driven feature selection algorithm: NSVA; a "new" DE analysis method: NSVA-DESeq by integrating our NSVA feature selection

with the parametric *DESeq* analysis; and a novel network-marker-oriented biomarker identification search strategy. In the following sections, we focus on NSVA feature selection before unveiling the SEQ-Marker algorithm.

Feature Selection for RNA-Seq Count Data

Although various feature selection algorithms are available in traditional omics data communities, most of these statistical testing-based methods may not be applied to RNA-Seq count data directly, because they usually assume that population data are normally distributed (eg, *t-test*).[8–10] On the other hand, traditional transform-based feature selection methods, such as PCA, ICA, or NMF,[16,20] also face difficulties in ranking each gene effectively. This is because they have to transform RNA-Seq count data to a subspace generated by principal components (PCs), independent components, or nonnegative bases to seek the meaningful linear combinations of features (genes). However, it is hard to distinguish an individual gene's contribution to the linear combination of all genes because of the nature of these transforms.[16,20]

As such, it is believed that a desirable feature selection for RNA-Seq count data should satisfy the following criteria. First, it should be a data-driven method that does not have any prior assumption about data distribution to avoid possible biases from the distribution itself. Second, it should avoid evaluating each gene's significance from the linear combinations of all genes in a subspace induced by a linear or nonlinear transform. Third, it should take consideration of the nonnegative characteristic of RNA-Seq count data instead of treating them as generic data. As such, we presented a novel data-driven feature selection method, NSVA, which did not have any prior data distribution assumption and enables the gene-significant ranking by fully taking advantages of nonnegativity of RNA-Seq count data. To some degree, it can be viewed as a special singular value decomposition (SVD) for nonnegative data. However, the characteristics related to SVD applied to nonnegative data are first unveiled and proposed in this work. We describe the classic SVD as follows before introducing NSVA.

Singular value decomposition. Given matrix $A \in \Re^{n\times p}$ with a rank $r = \min(n, p)$, it has the following SVD:

$$A = U\Sigma V^t = \sum_{j=1}^{r} s_j u_j v_j^T \quad (1)$$

where $U=[u_1, u_2, \ldots, u_n] \in \Re^{n\times n}$ and $V=[v_1, v_2, \ldots, v_p] \in R^{p\times p}$ are orthogonal matrices, and $\Sigma \in \Re^{n\times p}$ is a diagonal matrix mainly consisting of singular values, ie, $\Sigma = \text{diag}(s_1, s_2, \ldots s_r, \ldots, 0)$, $s_1 \geq s_2 \geq \ldots s_r > 0$.

Different from SVD that treats nonnegative read count data as genetic data, our proposed NSVA is a more data-driven algorithm that assumes the non-negativity of input data to match the key characteristic of RNA-Seq count data. Our NSVA is built upon the Perron–Frobenius theorem, which has been widely used in Google webpage ranking,[21] as follows.

Perron–Frobenius theorem. Given a nonnegative square matrix $A = (a_{ij}) \geq 0, A \in \Re^{n\times n}$, let λ_m be the largest eigenvalue of A and ν_m be its corresponding eigenvector (ie, $A\nu_m = \lambda_m \nu_m$), then it has the following properties:

1. $\lambda_m > 0$ and $\nu_m \in \Re^n$ has only nonnegative entries, ie, $\nu_m \geq 0$, or $\Pr(\nu_m < 0) = 0$.
2. Given $A\nu = \lambda\nu$ and $\lambda \neq \lambda_m$, then the eigenvector $\nu \in \Re^n$ must contain at least one negative entry.

Applying it to the SVD decomposition of a nonnegative matrix, we have our NSVA, whose proof details are skipped for the conciseness of description.

Nonnegative singular value approximation. Given a nonnegative matrix $A \in \Re^{n\times p}, A \geq 0$ with a rank $r = \min(n, p)$, and its SVD decomposition $A = \sum_{j=1}^{r} s_j u_j v_j^T$, we have the following results:

1. Both vectors $u_1 \in \Re^n$ and $v_1 \in \Re^p$ contain only nonnegative entries, ie, $u_1^{(j)} \geq 0, \nu_1^{(i)} \geq 0, j = 1, 2, \ldots, n, i = 1, 2, \ldots, p$.
2. The vectors $u_j \in \Re^n$ and $\nu_k \in \Re^p$ contain at least one negative entry for $2 \leqslant j \leqslant n$, and $2 \leqslant k \leqslant p$.
3. Matrix A has the following approximation along the first singular value direction:

$$A \sim s_1 u_1 \nu_1^T = \sum_{j=1}^{n} \sum_{k=1}^{p} s_1 u_1^{(j)} v_1^{(k)}.$$

It is noted that NSVA guarantees a purely additive decomposition of a nonnegative matrix A, which is an RNA-Seq read count matrix in our context, $A \sim s_1 u_1 v_1^T = \sum_{i=1}^{n} \sum_{k=1}^{p} s_1 u_1^{(i)} v_1^{(k)}$ ie, there are no negative components in the decomposition along the first singular value direction $\nu_1 \in \Re^p$. 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 ith gene of input data, in the "space" spanned by all entries of ν_1, ie, $S = \text{span}(\nu_1^{(1)}, \nu_1^{(2)}, \cdots, \nu_1^{(p)})$ with a weight s_1.

From a single gene viewpoint, NSVA means that each gene is approximated by the projection of its corresponding entry in vector u_1 on the singular value direction ν_1, ie, $A_i^T \sim s_1 u_1^{(i)} \nu_1^T, i = 1, 2, \ldots, n$. It is worthwhile to point out that such an approximation makes it possible to rank each gene by using its coefficient in the spanned space S, where each $\nu_1^{(k)}$ can be viewed as the meta-sample corresponding to the kth sample, and $u_1^{(i)}$ indicates the ith gene A_i^T's contribution to all the meta-samples. We describe the detailed NSVA feature selection in the following section.

NSVA feature selection for RNA-Seq count data. Our proposed NSVA makes it possible to represent each gene A_i^T by its contribution $u_1^{(i)}$ to all meta-samples $[\nu_1^{(1)}, \nu_1^{(2)}, \ldots, \nu_1^{(n)}]$. Since each meta-sample is the prototype of its original sample in the direction corresponding to the largest singular value s_1, it is natural to define a gene distribution score to quantify a gene's

contribution to all original samples of RNA-Seq count data by evaluating each gene's contribution to all meta-samples.

A *gene contribution score* measures a gene's contribution to all samples of an RNA-Seq count data $A \in \Re^{n \times p}$ by evaluating its contribution to all meta-samples in the low dimensional space S. Since it is positive or at least nonnegative for each gene according to our proposed NSVA algorithm, it guarantees the comparability for all genes. The gene contribution score of the ith gene to all samples is defined as $u_1^{(i)} = s_1^{-1} \sum_{j=1}^{p} a_{ij} v_1^{(j)} = s_1^{-1} A_i^T v_1$ by applying NSVA, that is

$$u_1^{(i)} = s_1^{-1} A_i^T \begin{pmatrix} v_1^{(1)} \\ v_1^{(2)} \\ \vdots \\ v_1^{(p)} \end{pmatrix} \quad (2)$$

As a measure to rank each gene's contribution to all samples of an RNA-Seq count dataset, the gene contribution score is independent of data distribution. In other words, such a property guarantees that it will integrate with any data analysis methods smoothly, no matter they are parameter estimation methods or not. Moreover, it avoids the technical difficulty faced by the traditional transform-based feature selection methods (eg, PCA) to evaluate each gene's significance from the linear combinations of all genes by taking advantage of the non-negativity of RNA-Seq count data. Thus, as a measure induced by NSVA, it appears as a desirable way to conduct feature selection for RNA-Seq count data. We call such a gene contribution score-based feature selection as NSVA feature selection and describe it as follows.

The NSVA feature selection has the following two steps. The first is to conduct NSVA for input RNA-Seq count data and compute the gene contribute score for each gene. The second is to employ the gene contribution scores to rank the importance of each gene and filter the genes with small gene contribution scores. For instance, sort all gene contribution scores and pick the top 2000 genes with largest scores. The genes with large gene contribution scores, which are potentially "good" genes, will enter the following data analysis (eg, DE analysis). It is noted that the gene contribution score is a weight to evaluate each gene's contribution to all samples instead of a percentage value, that is the summation of all gene contribution scores is not equal to 1.

It is worthwhile to point out that NSVA feature selection is actually a variance-based feature selection, where NSVA filters the genes according to their gene contribution scores, which is equivalent to filtering genes by the count variance. Each gene, $A_i^T, i = 1, 2, \cdots, n$, can be decomposed as $A_i^T \sim s_1 u_1^{(i)} v_1^T$ along the first singular value direction v_1, where the gene contribution score $u_1^{(i)}$ can be viewed as the variance term for the gene. It is obvious that the gene count variance is linearly proportional to its gene contribution score.

Data variation explanation ratios. It is noted that the first singular value s_1 is usually quite large for RNA-Seq count data compared with the other singular values. To evaluate the percentage of information that can be represented in NSVA, we define $\rho = s_1 / \sum_{i=1}^{r} s_i$ as the *data variation explanation ratio*, which is the ratio between the first singular value and the sum of singular values. Because each singular value is the square root of the corresponding eigenvalue of AA^t, the ratio actually represents the percentage of the data variances along the maximum eigenvector direction of AA^t among total data variances, ie, first singular value direction. Unlike other omics data, we found that the data variation explanation ratio of most RNA-Seq count data can reach 60% or higher (eg the ratio is 60.49% for the *Kidney-Liver* data), which means our NSVA is a reasonable approximation of the original count data. Similarly, the data variation explanation ratio is 85.60% for another *Prostate* data used in this study.

SEQ-Marker Algorithm

The proposed SEQ-Marker consists of the following three major procedures. First, it employed NSVA feature selection to obtain a gene set G, with the largest gene contribution scores from RNA-Seq count data $X \in \Re^{n \times p}$, $|G| < n$. It is noted that we employed *DESeq* analysis normalization to normalize input RNA-Seq data before applying NSVA feature selection to mitigate the biases from SD and gene length on gene counts. As a normalization method developed in *DESeq* analysis package,[7] the *DESeq* analysis normalization calculated a scaling factor (size factor) for each sample by using a pseudo-reference sample that consisted of geometric means of all genes, which was reported as one of the two best normalization methods for RNA-Seq data according to Dillies et al's work.[22]

Second, it applied *DESeq* analysis to the gene set G to calculate P-values for all genes. The reason we employed *DESeq* analysis was because it was the most robust parametric DE analysis method that demonstrated strong advantages over other parametric methods (eg, *edgeR*) to achieve low false-positive ratios.[7,12] Since such *DESeq* analysis was applied to the genes selected by NSVA feature selection instead of all original genes, we distinguished it with the original *DESeq* analysis applied to the whole data set by naming it as NSVA-DESeq for the convenience of description. We will demonstrate that the NSVA-DESeq would produce more meaningful P-values than applying *DESeq* analysis to the original data for benchmark RNA-Seq datasets (see Results section).

Third, it implemented a novel biomarker search strategy by searching biomarkers from an inferred network marker rather than from RNA-Seq count data directly. The proposed SEQ-Marker algorithm at first employed *jActiveModule* to seek several network marker candidates and merged them to a "final" network marker under a threshold. Then, it searched the inferred "final" network marker to identify "core genes"

with the largest interactions, and clustered the network marker to find densely connected gene regions. Finally, it further collected biomarkers by identifying those genes with the closest correlation distances with the core genes in the clusters.

The reason we identified the core genes in the network marker was to answer the query: "which genes will most likely act as key genes in the network marker and its mutations will affect other genes mostly?" Previous studies reported that the genes with the largest interactions will play an essential role in identifying disease molecular signatures and their mutations will have most impacts on those of other genes in the network marker.[19,23,24] Moreover, collecting genes with closest correlations with identified core genes will identify those genes with highest proximities with the key genes. That is, their mutations may trigger those of the identified core genes or vice versa. It is equivalent to answering the query: "which gene will most likely mutate with the key genes in the network marker?"

The reason we picked *jActiveModule* to infer network markers was mainly because it only required the expression data and *P*-values and did not have specific data distribution assumption, although it was developed for normally distributed gene expression array data, in addition to the fact that it has more robust support from *Cytoscape* and its related plugins than other peers.[25,26] It is noted that we employed *jActiveModule* 1.8 in *Cytoscape* 3.02 in our network marker inference. Figure 1 illustrated the flowchart of the proposed SEQ-Marker algorithm, which consisted of the following steps.

1. *NSVA feature selection:* conduct NSVA feature selection to calculates a gene set $G = \{g_1, g_2, \ldots, g_k\}$ from the normalized read count data $X \in \Re^{n \times p}$, such that $|G| < [n/2]$. For instance, $|G| = 2000$, which means the top 2000 genes with the largest gene contribution scores were used to seek the network marker. It is noted that each gene is assumed to have its Ensemble ID or obtain it from *BioMart* or other databases.
2. *NSVA-DESeq analysis*: conduct *DESeq* analysis for the gene set *G* to calculate the *P*-value for each gene, where each *P*-value is adjusted by using Benjamini–Hochbert procedure by choosing false discovery ratio (FDR) threshold 0.001.[27]
3. *Network marker inference*: input each gene in *G* with its Ensemble ID and its *P*-value to *jActiveModule*, along with corresponding read count data,[25] to seek *k* (eg, $k = 5$) in DE network markers M_k, $k = 1, 2, \ldots, 5$, by using the BioGrid human PPI network,[28] which has 17,580 proteins and 217,217 interactions, as the global network in our context.
4. Merge the network markers $M = \cup_{k=1}^{l} M_k$, provided their scores are greater than a threshold value of 5, which is set as 70th percentile of all network marker scores in our experiment, and drop the others with low scores.

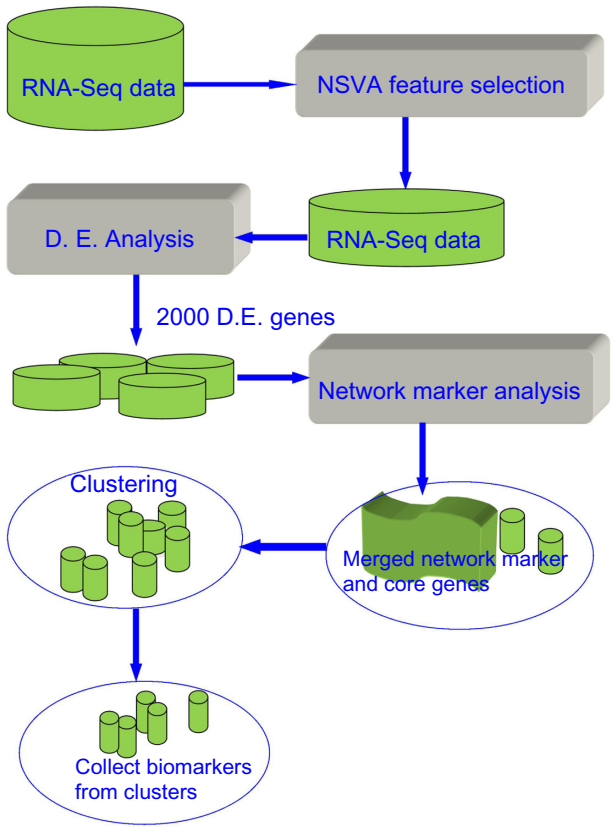

Figure 1. The flowchart of the proposed SEQ-Marker algorithm. The SEQ-Marker algorithm consists of the following main components: a data-driven feature selection algorithm: NSVA; a "new" DE analysis method: NSVA-DESeq, by integrating our NSVA feature selection with the parametric DESeq analysis; and a novel network marker-oriented biomarker identification search strategy.

5. Search the network marker to identify the first *l* core genes (eg $l = 5$) with largest interactions in the merged network marker *M*.
6. Cluster the network marker with a degree threshold (eg, 2) and seek an associative gene for each core gene that has the nearest correlation distances with the core gene with adjusted *P*-value < 0.001 in the clusters to collect the biomarkers left under a correlation threshold $\tau_c = 85\%$. If a cluster includes a core gene, then the associate gene will be searched in the cluster. Otherwise, the search will be done for the whole network marker. Similarly, if the nearest gene from the cluster has a correlation value less than the threshold, it will be dropped and a new search will be conducted for all the other genes in the network marker.

It is noted that we could have more than one associative gene for each core gene theoretically. However, we preferred to implement only one associative gene search in our implementation for the sake of biomarker validation. Moreover, the network marker acted as a small database that provides interaction information for biomarker identified. On the other

hand, the biomarkers identified will reveal the essential genes in the network marker, both of which contribute to unveiling the disease signature in a comprehensive approach.

Results

We presented our results on RNA-Seq count data biomarker discovery by applying our SEQ-Marker algorithm to two benchmark datasets: *Kidney–Liver* and *Prostate* in this study. The former is a dataset evolved from the original *Marioni* data by filtering genes with counts less than 5.[5] The latter is a dataset aligned and preprocessed by ourselves. We introduce the detailed information about these two datasets as follows before presenting our results.

Kidney–Liver data originally consist of 32,000 genes across 14 samples after Illuminasupplied alignment algorithm ELAND.[5] 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. We filtered the genes with counts ≤ 5 and obtained a dataset with 15,514 genes for the sake of meaningful DE analysis.

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). We employed *Bowtie and SAMtools*[2,29] to align the sequence data, which can be found in Li et al's work,[30] with respect to the human genome indexes (NCBI version 37), and collected read counts for each gene. Finally, we obtained a nonnegative integer matrix with four DHT-treated and three Mock-treated samples across 23,068 genes.

The success of the proposed biomarker discovery algorithm largely relies on *DESeq* analysis on the gene set obtained from the NSVA feature selection, ie, NSVA-DESeq. Thus, we need to evaluate its performance on the two datasets to answer the query: "what will happen in the first two steps of SEQ-Marker algorithm, where *DESeq* analysis is applied to the gene set selected by NSVA feature selection?"

Figure 2 answered the query by comparing NSVA-*DESeq* with *DESeq* on the two dataset, where NSVA selected 2000, 3000, 5000, and 8000 genes from each data and *DESeq* analysis was applied to these selected genes and their original datasets, respectively. The FDR cutoff was chosen as 0.001 in all DE analyses. Each horizontal and vertical axis in the sub-plots represents the $\log_2$ mean of each gene and the corresponding $\log_2$ fold changes under two different conditions. We had the following interesting findings from these results.

First, our NVSA feature selection demonstrated a good sensitivity to filter those non-DE genes, no matter that DE genes were the majority or not among all the input genes. For instance, non-DE genes were the majority among all genes in the *Prostate* data but a minority in the *Kidney–Liver* data. However, the proposed NSVA tended to selectively filter the non-DE genes by picking genes with large gene contribution scores. Such a mechanism made following DE analysis focused more on the potentially "good genes" and actually contributed to decreasing false positives.

We compared our NSVA feature selection with its two competing methods: count-based naive feature selection (NFS) and principal component analysis (PCA) feature selection to further demonstrate its advantage in picking potential DE genes. The count-based NFS selected genes according to its counts completely. It consisted of the following two steps. The first step selected all genes whose entries are more than or equal to the median count of the input data. The second step sorted all genes according to its coverage, ie, the sum of its counts, and selected the top-ranked genes (eg, 2000 genes).

On the other hand, PCA feature selection ranked each gene by using the 2-norm of its projection in the subspace spanned by the first three PCs. It represented the gene's contribution to all PCs and reflected its significance in the spanned subspace. PCA feature selection consisted of the following three steps. The first step conducted PCA for input data and projected it to the first three PCs. The second step calculated the 2-norm for the projection data of each gene in the subspace spanned by the three PCs. The third step sorted the genes according to the calculated 2-norm and selected the top-ranked genes. It was interesting to point out that our PCA feature selection had very high explanation ratio (>99%) for both datasets, compared with NSVA feature selection.

The northeast and northwest plots in Figure 3 compared the DE ratios from the three feature selection methods: NSVA, PCA, and NFS under DESeq analysis on corresponding 2000, 3000, 5000, and 8000 selected genes from the two original RNA-Seq datasets. The DE ratio was defined as the ratio of DE genes among all the genes of input data. It was interesting to see that the DE ratios from NSVA feature selection were much higher than those of NFS and PCA feature selection for all selection cases of two datasets. Since we only employed DESeq for DE analysis for all datasets, it was clear that the proposed NSVA feature selection demonstrated its advantage in selecting potential DE genes than the NFS and PCA feature selection.

All DE ratios showed stable increasing patterns with the increase in the number of genes filtered on the *Prostate* data, in which most genes were non-DE. However, the DE ratios of PCA feature selection reached only 17.25% on the 8000 selected genes, which was much lower than 24.30%, the DE ratio achieved by the original *Prostate* data under *DESeq* analysis without any feature selection. Moreover, it was obvious that the DE ratios from PCA were the lowest among the three feature selection methods. It indicated that the transform-based feature selection may not be a good choice for RNA-Seq count data, even if it had high data variation explanation ratios.

On the other hand, the largest DE ratios from PCA and NFS only reached 83.05% and 87.45% on the selected

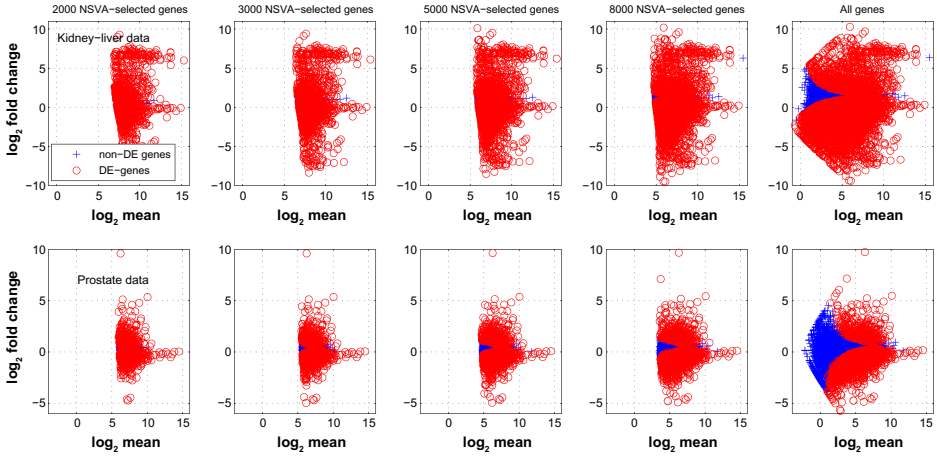

Figure 2. The scatter plots of $\log_2$ mean versus $\log_2$ fold changes by comparing *DESeq* and *NSVA-DESeq* on *Kidney–Liver* and *Prostate* data, where 2000, 3000, 5000, and 8000 genes are selected by NSVA from each data, and DESeq analysis is then applied to these selected genes and their original datasets, respectively. It is interesting to see that non-DE genes dropped remarkably when NSVA feature selection is applied to each dataset.

2000 genes for the *Kidney–Liver* data. However, the DE ratio reached 88.16% for this dataset without any feature selection, and the DE ratios from NVSA reached 90.05%, 89.80%, 91.40%, and 91.90%, respectively, on the selected 2000, 3000, 5000, and 8000 genes. In other words, NFS and PCA feature selection did not demonstrate any advantage in enhancing DE analysis on the *Kidney–Liver* data, where most genes are DE. Alternatively, the DE ratios of NSVA kept a slightly increased pattern compared with the original DE ratio 88.16%, with the increase in the number of genes filtered. It indicated that the false positive ratios were forcibly dropped in DE analysis under such a "conservative" feature selection.

Second, the proposed NSVA tended to select the DE genes with high counts with increase in the number of genes filtered. The southeast and southwest plots in Figure 3 compared the DE gene median counts from three feature selection methods. It was interesting to see that the DE gene median counts from NSVA was generally lower than those from NFS and PCA on two datasets, except those that were greater than the DE gene median counts from PCA at the 5000 and 8000 gene selection cases. In other words, NSVA feature selection had the highest DE ratios but shortest median counts for DE genes among all the three methods. It suggested that DE genes in NSVA-DESeq were mostly high-count genes instead of low-count ones, which was consistent with the previous results.[5,14]

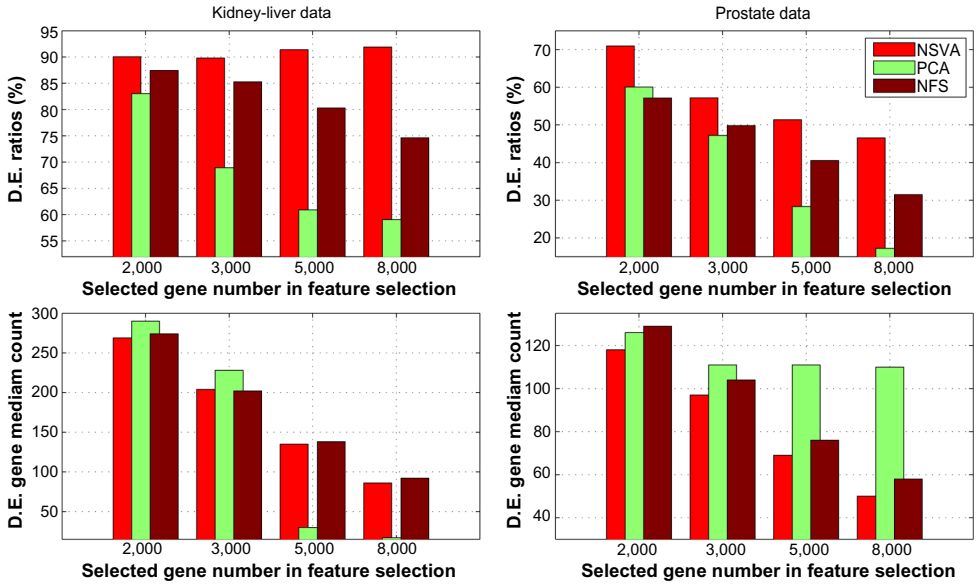

Figure 3. The comparisons of DE ratios and DE gene median counts for NSVA, PCA, and NFS feature selection methods under *DESeq* analysis on the *Kidney–Liver* and *Prostate* data. The proposed NSVA feature selection demonstrated strong advantages in selecting potential DE genes than the two competing methods. The DE gene median counts from NSVA are generally lower than those of PCA and NFS.

However, it does not mean that high-count genes will be DE genes because our result demonstrated the DE ratios from NFS were much lower than those of NSVA under *DESeq* analysis, especially when more genes were selected in feature selection. On the other hand, because the DE gene median counts were only 39 and 35 for the original *Kidney–Liver* and *Prostate* data, there were quite a few false positives removed in DE analysis due to NSVA feature selection, because most false positive genes were reported as low-count genes.[9,13]

Third, we found that the DE genes among NSVA-selected genes tended to be relatively long genes, which was also true for NFS- and PCA-selected genes. Figure 4 compared the gene length medians of NSVA-, NFS-, and PCA-selected genes and DE genes among these selected genes. It was interesting to see that PCA selected the shortest genes among three of them. For example, the gene length medians for its DE genes were quite low in the 3000, 5000, and 8000 gene selection cases. Considering the low DE ratios and low counts for PCA-selected methods, it is reasonable to say that PCA tends to select those genes with low counts or short lengths, most of which are obviously not DE genes.

Alternatively, NFS-selected genes and the DE genes among them had the longest gene lengths, but the DE ratios from NFS were quite low compared with those of NSVA, especially under the 5000 and 8000 gene selection cases. It strongly suggested that only picking high-count genes would not contribute to enhance DE analysis, because a large number of pseudo high-count genes could be generated from those lanes with high SD in RNA-Seq.[7,13]

On the other hand, the median gene lengths from NSVA-selected genes were shorter than those from NFS but higher than the DE gene median length (26,445 bp) of all genes for the *Kidney–Liver* data. For example, its DE gene median length reached 27,659 bp on the 2000 gene selection case, which was much lower than that of NFS (29,328 bp). Compared with the results from NFS and PCA, the proposed NSVA demonstrated the consistency in identifying meaningful genes in DE analysis that consisted of relatively long genes with reasonable counts instead of all high-count or long genes. It strongly suggested that our NSVA-DESeq would have an SD independence property in DE analysis, although the original *DESeq* analysis was dependent on the SD.[9,13]

In summary, NSVA feature selection seems to make the following *DESeq* analysis more targeted on the genes with large gene contribution scores, most of which were proved to be DE genes because of their large contributions to the whole data variations. Obviously, NSVA will contribute more to the decrease in false positives than NFS and PCA, because quite a lot of false positive candidates were filtered before the DE analysis. In fact, all these results demonstrated that NSVA and following *DESeq* analysis, the first two steps in our biomarker discovery algorithm, will prepare meaningful genes for next steps of the SEQ-Marker algorithm.

Biomarker Discovery for Kidney–Liver and Prostate Data

We applied our SEQ-Marker algorithm to seek biomarkers for the *Kidney–Liver* data. At first, we applied *DESeq* analysis to 2000 genes selected by NSVA, which consisted of 1801 DE genes and 199 non-DE genes. Figure 5 illustrated those non-DE genes, DE genes, and all 2000 genes in the left, middle, and right plots, respectively. It was interesting to see that the non-DE genes had fold changes in a much smaller range than the DE genes, which guaranteed that 90% genes in network marker inference were DE genes and leaded to more meaningful network markers.

In addition, we obtained five network markers with scores 8.219, 7.922, 7.754, 7.735, and 7.637, respectively from *jActiveModule*, which were further merged to a "single" network marker with 102 genes and 194 interactions. We then identified

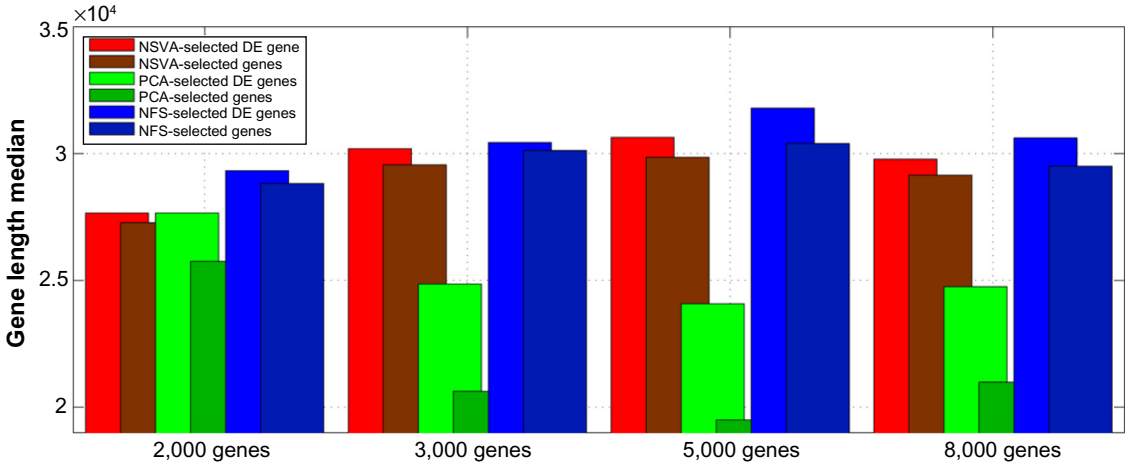

Figure 4. Comparisons of the gene length medians of the genes selected by NSVA, PCA, and NFS methods and DE genes among the selected genes for the *Kidney–Liver* data. The DE genes have longer gene length than those selected genes from each feature selection method. The DE genes from NSVA-selected genes seem to be shorter than NFS-selected genes but longer than the PCA-selected genes.

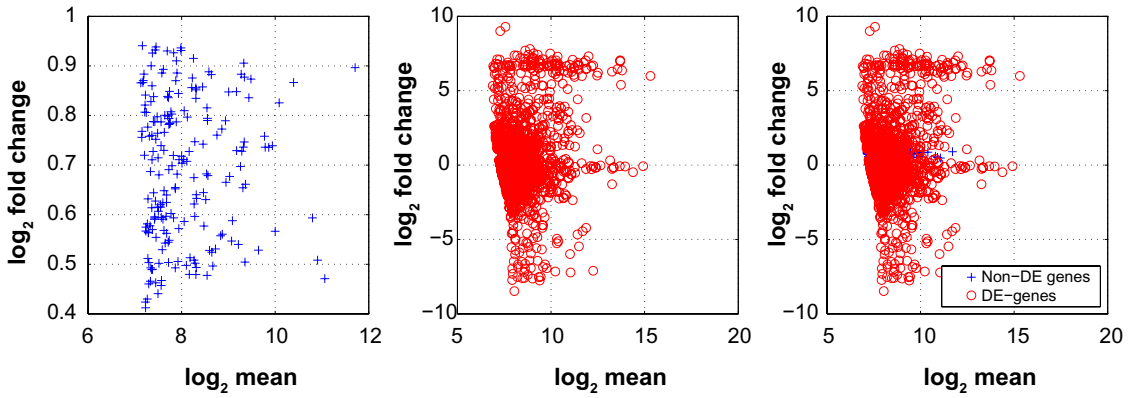

Figure 5. The plots of 1801 DE genes and 199 non-DE genes of 2000 genes selected by NSVA for the Kidney–Liver data. Unlike the DE genes, the non-DE genes have fold changes in a quite small range.

that there are $k = 5$ core genes FN1, ESR1, HSB90 A, HSPB1, and CTNNB1 related to liver and kidney diseases by examining the genes with the largest interactions in the network marker. Figure 6 illustrated the network marker where the core genes with the largest interactions (degrees) were emphasized in the network topology.

It was interesting to find that these core genes were actually gene markers closely related to kidney and liver diseases.

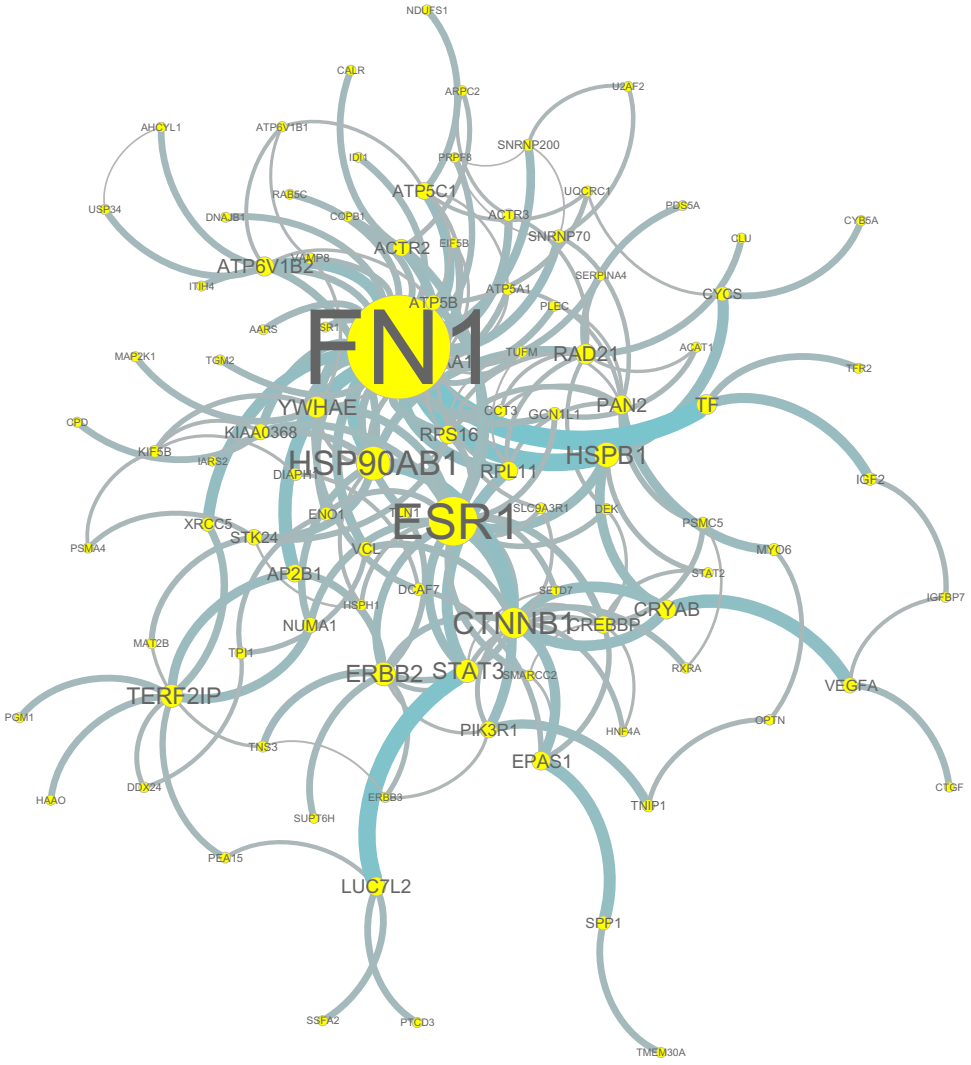

Figure 6. The network marker with 102 genes and 194 interactions identified by the SEQ-Marker algorithm for *Kidney–Liver* data. The five core genes with the largest interactions (degrees) were emphasized in the network topology.

Table 1. The five core genes identified for *Kidney–Liver* data.

GENE	*P*-VALUE	GENE LENGTH (bp)	AVG. GENE COUNT	FOLD CHANGE
FN1	7.0315e-236	75,611	1,185	8.8564
ESR1	7.7471e-175	297,588	144	13.0222
HSP90AB1	1.0320e-105	6,797	965	0.6216
CTNNB1	1.4246e-032	40,939	435	0.9591
HSPB1	1.9217e-141	1,690	223	0.4379

For example, the gene FN1 with the largest interactions was reported as a gene associated with the development of renal cell cancer (RCC), that is a cancer related to kidney.[31] The gene ESR1 with the second largest interactions in the network marker was usually differentially expressed in liver, kidney, and other human organism parts and its mutation was reported to relate to kidney- and liver-related diseases (eg, osteosarcoma of the kidney).[32] The gene HSB90AB1 with the third largest interactions has usually seen its DE in liver, kidney, and other organisms, and its mutations were reported to be associative with kidney diseases in the previous studies.[33] Alternatively, it was easy to enumerate the biological relevance of the other two core genes, HSPB1 and CTNNB1, in the network marker with respect to kidney or liver diseases in the literature. The HSPB1 interacting with p53 directly was reported to inhibit the lung and liver tumor progression,[34] and the somatic mutation of the CTNNB1 gene was associated with the liver, skin, and other cancers.[35]

Table 1 showed more detailed information about the five core genes. Interestingly, the core genes were DE genes with strong *P*-value support, which included both long and short genes. In fact, FN1 and ESR1 were up-regulated genes and others were down-regulated genes. In addition, only ESR1 had a relatively low average count number (144) compared with other genes, whose average gene numbers were much higher than 269 bp, the median count of DE genes among 2000 NSVA-selected genes.

We conducted clustering for the network marker with a degree threshold 2 by employing MCODE[26,36] and obtained four clusters: {ATP5B, FN1, ATP6V1B2, ATP5C1}, {PIK3R1, ERBB2, CTNNB1}, {HSPH1, HSP90 AB1, STK24}, and {SNRNP70, PAN2, PRPF8, RPL11, SNRNP200, PRKAA1, TUFM}. The first cluster had a cluster score of 3.33 and the other three had 3. We further found five corresponding associative genes for the core genes, where ATP5B and STK24 were the associative genes found in a same cluster as their core genes FN1 and HSP90ABf, respectively. Moreover, RPS9, ADH4, and RHOC were the associative genes of the core genes CTNNB1, ESR1, and HSPB1, respectively, Interestingly, almost all associative genes had high correlation values with their corresponding core genes, except RPS9 (core gene: CTNNB1, correlation value: 62.16%).

Table 2 showed the details about these associative genes. They shared almost the same characteristics as the core genes: all genes are DE genes with strong *P*-value support, where RPS9, STK24, and RHOC ADH4 could be viewed as "short-count" genes compared with the median DE gene counts. We further grouped all the 10 biomarkers and conducted diagnosis by using a linear support vector machine under Leave-one-out cross validation (LOOCV) and achieved 100% accuracy with 100% sensitivity and specificity, we are not surprised by such a result because we found that RNA-Seq count data have quite good discrimination ability than traditional omics data.

Similarly, we applied the SEQ-Marker algorithm to the *Prostate* data and obtained the following network marker with 203 genes and 730 edges as shown in Figure 7. We identified five core genes such as APP, HSP90AA1, NEDD8, HNRNPA1, and NPM1 from the inferred network marker. It was interesting to see that almost all core genes had strong *P*-value support except HSP90AAI. Although it was actually not a DE gene because of its *P*-value, 0.2051 statistically, our SEQ-Marker algorithm indicated it as a biomarker for prostate cancer, which was proved as a real prostate cancer marker by the previous studies.[33,37] In addition, all the five core genes were high-count genes whose average gene counts were much higher than the median DE gene count: 118 bp. For example, the average gene count of APP and HSP90AAI reached 630 bp and 397 bp, respectively. Interestingly, we found that almost all these genes were associated or closely related to prostate cancer from previous studies,[37,38] for instances, APP was identified as a well-known gene marker to promote prostate cancer growth according to Takayama et al's work,[39] and NEDD8 conjugation pathway is essential for understanding prostate cancer or other complex cancer diseases.[38,40] We identified the corresponding associative genes for the core genes and included their corresponding correlation values: FKBP5

Table 2. The five associative genes identified for *Kidney–Liver* data.

GENE	*P*-VALUE	GENE LENGTH (BP)	AVG. GENE COUNT	FOLD CHANGE	CORRELATION
ATP5B	3.1498e-224	7,890	1,793	0.2192	99.92%
RPS9	5.0690e-012	6,790	203	1.1434	87.76%
STK24	8.0021e-144	12,6942	204	0.4046	96.88%
ADH4	0.000000000	20,618	1,164	108.0622	99.95%
RHOC	4.1735e-140	6,277	215	0.4379	98.91%

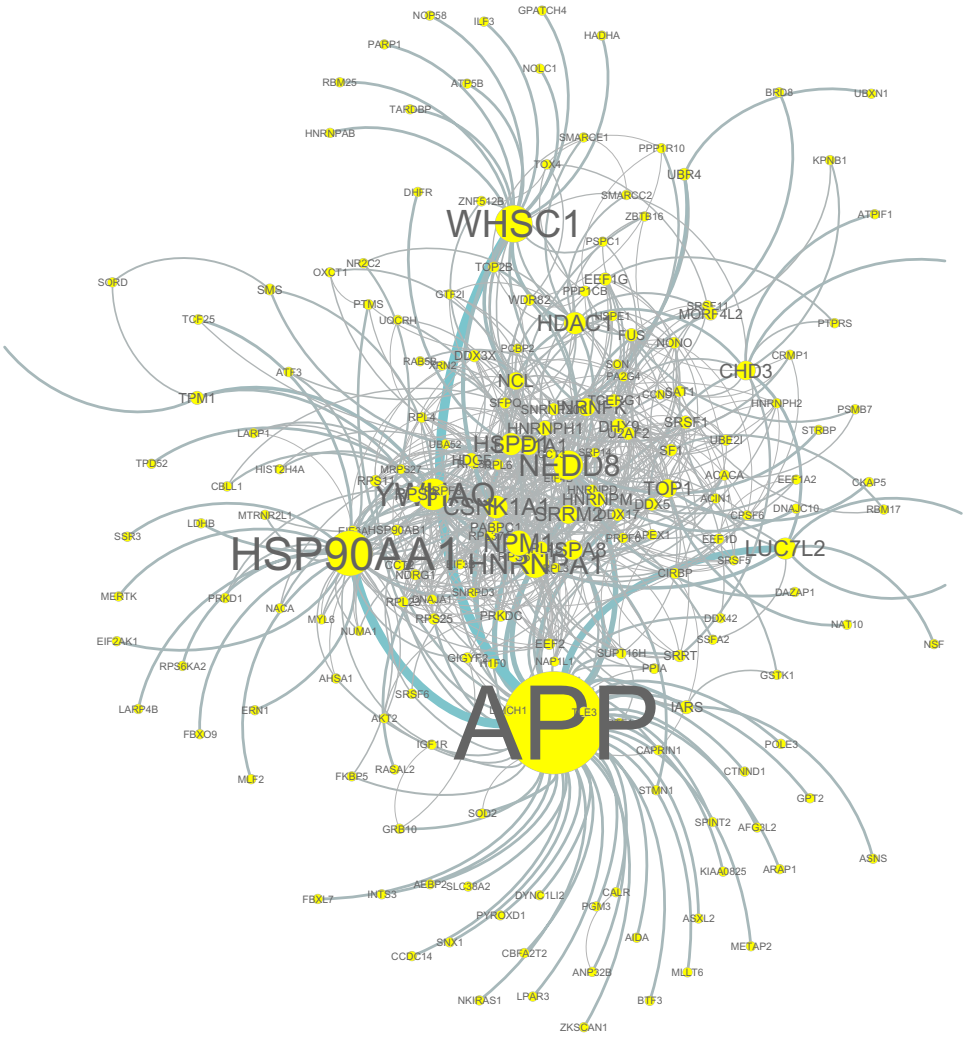

Figure 7. The network marker with 203 genes and 730 interactions identified by SEQ-Marker algorithm for *Prostate* data. The core genes with the largest interactions (degrees) were emphasized in the network topology.

(99.91%), SPTLC1 (98.57%), NEDD8-MDP1 (99.60%), DARS (99.53%), and MY06 (99.54%). It was interesting to find that FKBP5, DARS, and MY06 were well-known prostate cancer marker according to previous studies.[41–43] Similar to the *Kidney–Liver* data, we achieved 100% accuracy with 100% sensitivity and specificity by using the 10 biomarkers to conduct diagnosis under a linear support vector machine with LOOCV.

Discussion

In this study, we proposed a novel biomarker discovery algorithm SEQ-Marker for RNA-Seq count data. Our biomarker discovery algorithm is based on NSVA algorithm proposed in this work. As a data-driven feature selection algorithm, our NSVA algorithm demonstrated the advantages in selecting meaningful genes before DE analysis by contributing to lowering false positive rates and improving the sequence depth independent of DE analysis when compared with its peers: PCA and NFS feature selection.

Moreover, as a first algorithm to address biomarker discovery for RNA-Seq count data, SEQ-Marker identified biomarkers by searching an inferred network marker, which acted as a database to provide gene interaction for biomarkers. The database unveiled essential gene interaction information and provided the opportunity to identify real disease biomarkers along an interaction "path". As such, the biomarkers identified will reveal more network dynamics and contribute to unveiling disease signatures in a comprehensive approach.

It is worthwhile to point out that our proposed biomarker discovery model, SEQ-Marker for RNA-Seq data, overcomes the limitations of traditional omics biomarker discovery by finding the biomarkers without strong *P*-value support, in addition to contributing to convenient biomarker validation from a clinical viewpoint. Just as we pointed out before, our work not only bridges transcriptomics and systems biology, but also contributes to clinical diagnostics.

An issue of particular importance is that our studies have quite different focuses from the previous studies,[5,30] although almost same data were employed in these studies. For

examples, the original *Maroini* dataset, which is the source data of our *Kidney–Liver* data, was mainly used to demonstrate the advantages of reproducibility of RNA-Seq data with respect to microarrays.[5] Furthermore, the prostate dataset was employed to analyze alternative splicing and estimate the number of short reads (tags) required to detect specific genomic features under an androgen-sensitive prostate cancer model.[30] Thus, our work is the first to address RNA-Seq data biomarker discovery.

The proposed NSVA demonstrated a good strategy in enhancing the robustness of *DESeq* analysis by filtering less likely DE genes using each gene's contribution to the total data variances. Compared with PCA and NFS feature selection, it not only achieved the highest DE ratios for two benchmark datasets, but also avoided weakness of selecting high-count or long-length-based gene selection. However, how to achieve an optimal feature selection to reach a robust DE analysis is still a challenge theoretically and practically. This is because NSVA may remove some real DE genes before DE analysis and lead to the increase of false-negative ratios potentially. As such, we plan to employ information measures such as entropy to employ its potential in optimal NSVA feature selection.[44] On the other hand, our current NSVA algorithm only conducts feature selection along the first singular value direction, because RNA-Seq count data usually reach a high data variation explanation ratio (eg, >60%) on it. How to extend the proposed NSVA algorithm to include any specified number of singular value directions in feature selection, while maintaining its purely additive property for nonnegative RNA-Seq count data, is an important problem that deserves more investigation. We are developing a multi-resolution data model to decompose RNA-Seq count data and recursively conduct NSVA to achieve it.

Although we only integrated NSVA with the parametric method DESeq in our biomarker discovery in this study, we are integrating NSVA with nonparametric DE analysis algorithms such as NOISeq to investigate its performance in biomarker discovery.[13] In addition, because our proposed SEQ-Marker faces quite a high computing demand due to the complexities of identifying DE network markers in jActiveModule,[25] we plan to use Graphics processing unit (GPU) computing way to tackle the computing burden in the network marker discovery, in addition to applying it to RNA-Seq data-based clinical diagnosis.[45]

Author Contributions

Conceived and designed the experiments: HH. Analyzed the data: HH. Contributed to the writing of the manuscript: HH. Agree with manuscript results and conclusions: HH, XJ. Jointly developed the structure and arguments for the paper: HH, XJ. Made critical revisions and approved final version: HH, XJ. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57–63.
2. Bowtie. 2014. Available at http://bowtie-bio.sourceforge.net/index.
3. Luo R, Wong T, Zhu J, Liu C-M, Zhu X, et al. (2013) SOAP3-dp: Fast, Accurate and Sensitive GPU-Based Short Read Aligner. PLoS ONE 8(5): e65632. doi:10.1371/journal.pone.0065632Dillies.
4. Trapnell C, Hendrickson DG, Sauvageau M, Go L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 2013;31:46.
5. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509–17.
6. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464(7289):768–72.
7. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106.
8. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*. 2010;11:422.
9. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*. 2013;14:R95.
10. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*. 2012;13:523–38.
11. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing. *BMC Genomics*. 2012;13:484.
12. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40.
13. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res*. 2011;21:2213–23.
14. Anders S, McCarthy DJ, Chen Y, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc*. 2013;8:1765–86.
15. Oshlack A, Robinson M, Young M. From RNA-seq reads to differential expression results. *Genome Biol*. 2010;11:220.
16. Han X. Nonnegative principal component analysis for cancer molecular pattern discovery. *IEEE/ACM Trans Comput Biol Bioinform*. 2010;7(3):537–49.
17. Hus C, Lin C. A comparison of methods for multi-class support vector machines. *IEEE Trans Neural Netw*. 2012;13(2):415–25.
18. Vapnik V. *Statistical Learning Theory*. New York: John Wiley; 1998.
19. Ideker T, Sharan R. Protein networks in disease. *Genome Res*. 2008;18(4):644–52.
20. Jolliffe I. *Principal Component Analysis*. New York: Springer; 2002.
21. Langville AN, Meyer CD. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton, NJ: Princeton University Press; 2002.
22. MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*. 2013 Nov;14(6):671–83. doi: 10.1093/bib/bbs046. Epub 2012 Sep 17.
23. Kreeger KP, Lauffenburger DA. Cancer systems biology: a network modeling perspective. *Carcinogenesis*. 2010;31(1):2–8.
24. Han H, Li XL, Ng SK, Ji Z. Multi-resolution-test for consistent phenotype discrimination and biomarker discovery in translational bioinformatics. *J Bioinform Comput Biol*. 2013;11(6):1343010.
25. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*. 2002;18(suppl 1):S233–40.
26. Cytoscape. Available at http://www.cytoscape.org/. 2013.
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57:289–300.
28. BioGrid. Available at http://thebiogrid.org/. 2013.
29. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
30. Li H, Lovci MT, Kwon Y-S, Rosenfeld MG, Fu X-D, Yeo GW. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci U S A*. 2008;105(51): 20179–84.
31. Waalkes S, Atschekzei F, Kramer MW, et al. Fibronectin 1 mRNA expression correlates with advanced disease in renal cancer. *BMC Cancer*. 2010;10:503.
32. Cioppa T. Primary osteosarcoma of the kidney with retroperitoneal hemorrhage. Case report and review of the literature. *Tumori*. 2007;93(2):213–6.
33. Koshimizu TA. Inhibition of heat shock protein 90 attenuates adenylate cyclase sensitization after chronic morphine treatment. *Biochem Biophys Res Commun*. 2007;392(4):603–7.
34. Choi SH, Lee HJ, Jin YB, et al. MMP9 processing of HSPB1 regulates tumor progression. *PLoS One*. 2014;9(1):e85509.
35. Kimelman D, Xu W. beta-catenin destruction complex: insights and questions from a structural perspective. *Oncogene*. 2006;25(57):7482–91.
36. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4:2.

37. Centenera MM, Fitzpatrick AK, Tilley WD, Butler LM. Hsp90: still a viable target in prostate cancer. *Biochim Biophys Acta*. 2013;1835(2):211–8.
38. Hori T, Osaka F, Chiba T, et al. Covalent modification of all members of human cullin family proteins by NEDD8. *Oncogene*. 1999;18(48):6829–34.
39. Takayama K, Tsutsumi S, Suzuki T, et al. Amyloid precursor protein is a primary androgen target gene that promotes prostate cancer growth. *Cancer Res*. 2009;69(l):137–42.
40. Soucy TA, Smith PG, Milhollen MA, et al. An inhibitor of NEDD8-activating enzyme as a new approach to treat cancer. *Nature*. 2009;458(7239):732–6.
41. Nelson PS, Clegg N, Arnold H, et al. The program of androgen-responsive genes in neoplastic prostate epithelium. *Proc Natl Acad Sci U S A*. 2002;99(18):11890–5.
42. Wei S, Dunn TA, Isaacs WB, et al. GOLPH2 and MY06: putative prostate cancer markers localized to the Golgi apparatus. *Prostate*. 2008;68(13):1387–95.
43. Tu LC, Yan X, Hood L, Lin B. Proteomics analysis of the interactome of N-myc downstream regulated gene 1 and its interactions with the androgen response program in prostate cancer cells. *Mol Cell Proteomics*. 2010;6(4):575–88.
44. Kapur JN, Kesevan HK. *Entropy Optimization Principles with Applications*. Boston: Academic Press; 2002.
45. Dematté L, Prandi D. GPU computing for systems biology. *Brief Bioinform*. 2010;11(3):323–33.