Characteristic Gene Selection via Weighting Principal Components by Singular Values

Jin-Xing Liu[1,5], Yong Xu[1,6]*, Chun-Hou Zheng[2], Yi Wang[3], Jing-Yu Yang[4]

1 Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong, China, **2** College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui, China, **3** School of Mechanical Engineering and Automation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong, China, **4** School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, Jiangsu, China, **5** College of Information and Communication Technology, Qufu Normal University, Rizhao, Shandong, China, **6** Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong, China

Abstract

Conventional gene selection methods based on principal component analysis (PCA) use only the first principal component (PC) of PCA or sparse PCA to select characteristic genes. These methods indeed assume that the first PC plays a dominant role in gene selection. However, in a number of cases this assumption is not satisfied, so the conventional PCA-based methods usually provide poor selection results. In order to improve the performance of the PCA-based gene selection method, we put forward the gene selection method via weighting PCs by singular values (WPCS). Because different PCs have different importance, the singular values are exploited as the weights to represent the influence on gene selection of different PCs. The ROC curves and AUC statistics on artificial data show that our method outperforms the state-of-the-art methods. Moreover, experimental results on real gene expression data sets show that our method can extract more characteristic genes in response to abiotic stresses than conventional gene selection methods.

Citation: Liu J-X, Xu Y, Zheng C-H, Wang Y, Yang J-Y (2012) Characteristic Gene Selection via Weighting Principal Components by Singular Values. PLoS ONE 7(7): e38873. doi:10.1371/journal.pone.0038873

Editor: Yong Fan, Institution of Automation, CAS, China

Received December 19, 2011; **Accepted** May 13, 2012; **Published** July 10, 2012

Funding: This work was supported by fund for Key Research Project of National Natural Science Foundation of China under grant No. 90820306; Program for New Century Excellent Talents in University (No.NCET-08-0156), NSFC under grant No. 61071179; Provincial natural science research program of higher education institutions of Anhui province, No. KJ2012A005; China Postdoctoral Science Foundation Funded Project, No. 2012M510148; The Foundation of Qufu Normal University under grant, No. XJ200947.The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: laterfall2@yahoo.com.cn

Introduction

The growth of plants is greatly affected by a variety of abiotic stresses, such as cold, drought, salt, heat, UV-B light, osmotic press, and so on. In response to these abiotic stresses, plants have evolved a number of defense mechanisms that can increase tolerance to the adverse conditions. The underlying concept is that there exists a specific set of interacting genes responding to the abiotic stresses. Therefore, understanding abiotic stress responses is now thought to be one of the most important topics in plant science [1].

In order to obtain the characteristic genes responding to these stresses, many conventional experimental methods were proposed, such as RT-PCR [2,3] and Northern blotting [4,5], etc. RT-PCR can accurately position the genes in tissue or cell and Northern blotting can display the information of detected genes. However, these two methods have the following fatal flaw: only a limited amount of genes can be simultaneously studied. To overcome this flaw, people have developed the gene microarray technology [6–8], which makes it possible to monitor gene expression levels on a genomic scale [9,10].

With the rapidly development of gene microarray technology, how to efficiently analyze gene expression data becomes a matter of great urgency. During the last decade, feature selection from gene expression data has been extensively studied. The most commonly used methods of feature selection first calculate a score for each gene, respectively, then select the genes with high scores [11]. These methods are often denoted as univariate feature selections (UFS). The main virtues of UFS are: (a) intuitive and easy to understand; (b) computationally simple; and (c) fast [12]. A common disadvantage of UFS based methods is that each feature is separately considered, thereby ignoring feature dependencies. In order to handle the problem, the method of multivariate feature selection (MFS), also denoted as dimension reduction, was introduced [13]. MFS uses all the gene expression data simultaneously to select the genes. Until now, many mathematical methods for MFS have been used for gene expression data analysis. For example, Park *et al.* gave the theoretical analysis on feature extraction capability of class-augmented PCA [14]. Ma *et al.* used PCA to identify differential gene pathways in [15]. De Haan *et al.* used PCA to analyze microarray data [16]. Musumarra *et al.* used PLS to identify genes for new diagnostics [17]. Boulesteix *et al.* provided a systematic comparison of the PLS methods for the analysis of gene data [18].

However, the classical methods, such as PCA and PLS, still have some drawbacks, e.g. the principal components (PCs) of PCA or the latent components (LCs) of PLS are usually dense, which makes it difficult to interpret PCs or LCs without subjective judgment. To overcome these drawbacks, many mathematical tools have been devised to reduce the complexity of the data. Among them, sparse methods have significant advantage, while giving up little statistical efficiency. For example, Zou *et al.*

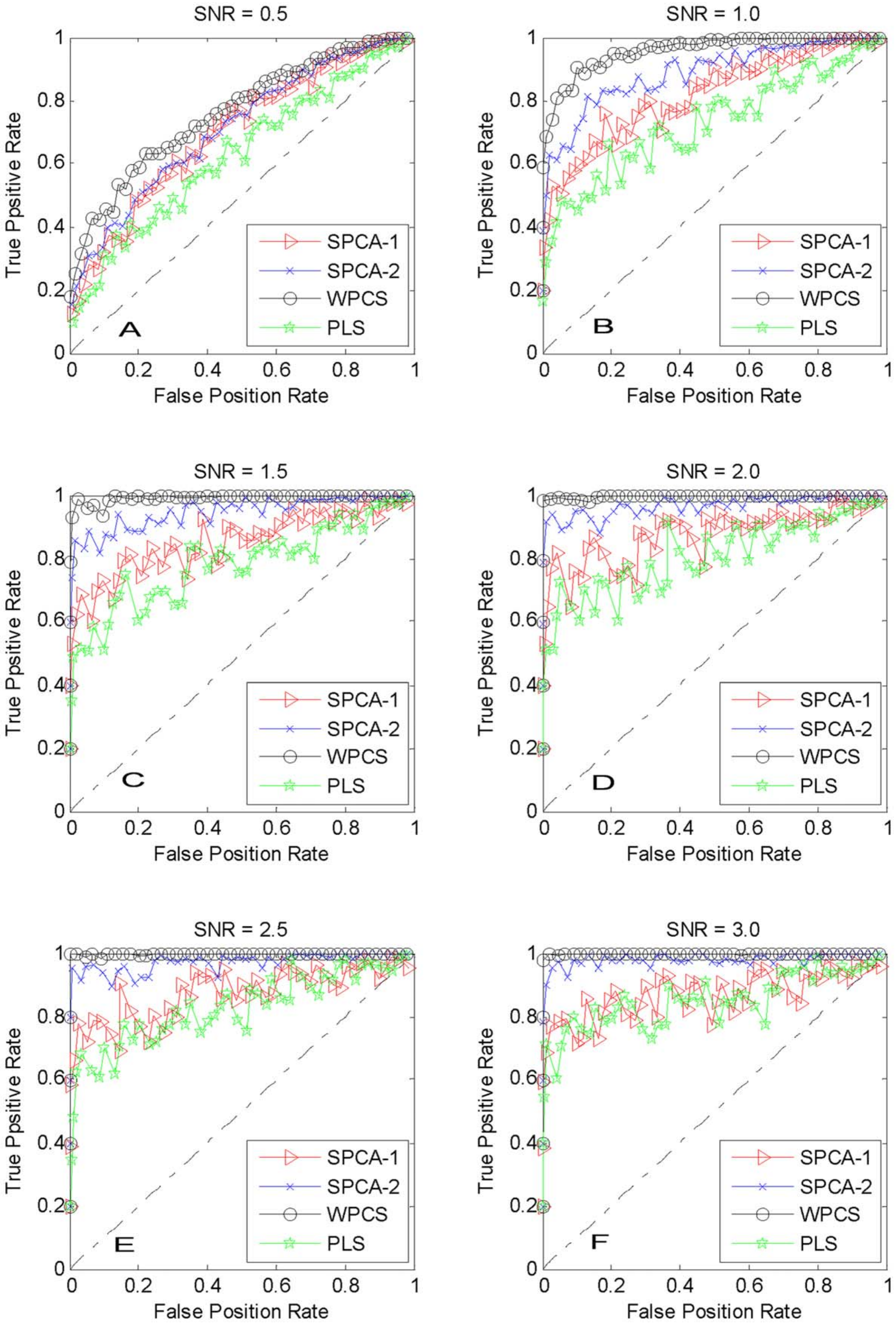

Figure 1. ROC curves for artificial data. (SNR denotes the signal-to-noise ratio).
doi:10.1371/journal.pone.0038873.g001

proposed sparse PCA using the lasso [19]. In [20], Journée *et al.* described a general power method for sparse PCA. Lai *et al.* used sparse local discriminant projections for feature extraction [21]. Moreover, many sparse methods have been widely used for gene expression data analysis. Lass *et al.* used the SPCA for clustering and feature selection [22]. In [23], Witten *et al.* proposed a penalized matrix decomposition, which was used to analyze plant gene expression data by Liu *et al.* [24]. Cao *et al.* used sparse PLS discriminant analysis for biologically relevant feature selection [25].

Though sparse methods are useful, yet these methods used the first LC of PLS or PC of SPCA to select feature. For example, Boulesteix *et al.* used the first LC of PLS to select the important genes [18]. Liu *et al.* used the first PC of SPCA for characteristic genes selection [26]. These methods indeed assumed that the first component of PLS or SPCA plays a dominant role in gene selection. However, we identify that in a number of cases this assumption is not satisfied and conventional PCA-based gene selection methods usually provide poor results. Actually, not only the first LC or PC but also the remaining LCs or PCs include the important information for gene selection [27]. So if only the first LC or PC is used in selecting genes, the poor results may have been obtained with loss of some important information.

In this paper, in order to select characteristic genes, a novel method is proposed that is based on the weighted PCs of SPCA by singular values (WPCS). First, the PCs of SPCA and singular values are calculated. Second, using the singular values as the weights of PCs, the weighted PCs (WPC) are gotten. Then, as the absolute value of the *i*-th entry of every WPC somewhat denotes the importance of the *i*-th gene, we take the sum of all the *i*-th entries of all the WPC as the extent of importance of the *i*-th gene and use it to select features. The genes corresponding to the first m largest sums are selected as characteristic genes. The experimental results show that our method is efficient and powerful for gene selection. Our work has the following contributions: first, it proposes, for the first time, the method of weighting PCs by singular values for gene selection. Second, from the viewpoint of minimizing reconstitution error, it clearly presents the idea of the proposed method. Third, it conducts a large number of gene selection experiments.

Results and Discussion

In this section, our proposed WPCS method is compared with the following existing methods. (a) SPCA-1 method uses only the first one PC of SPCA (proposed by Journée *et al.* [20]) to identify the characteristic genes; (b) SPCA-2 method uses the first two PCs of SPCA to identify the characteristic genes; (c) PLS method uses partial least squares regression (PLS) (proposed by Boulesteix *et al.* [18]) to identify the characteristic genes.

Table 1. AUC statistics for artificial data.

	SNR = 0.5	SNR = 1	SNR = 1.5	SNR = 2	SNR = 2.5	SNR = 3
SPCA-1	0.6985000	0.8186722	0.8517611	0.8738500	0.8733944	0.8665222
SPCA-2	0.7131000	0.8888778	0.9469222	0.9668333	0.9730667	0.9800500
WPCS	0.7615000	0.9579889	0.9918333	0.9971667	0.9988222	0.9989444
PLS	0.6282000	0.7269556	0.7812667	0.8061889	0.8261889	0.8506667

doi:10.1371/journal.pone.0038873.t001

First, these methods are carried on the artificial data. Then, these methods are used to extract the characteristic genes responding to abiotic stresses from real gene expression data.

Simulation on Artificial Data

To investigate the performance of the methods, the average receiver operator characteristic (ROC) curves are shown in Figure 1 with six different SNR.

Figure 1(A) and 1(B) show that our WPCS and the competitive methods can identify the patterns even with very low SNR. As shown in Figure 1, with different SNR, WPCS outperforms other methods. For example, with SNR = 1.0, PLS method are dominated by SPCA-2 and SPCA-1; and our WPCS can achieve the best results.

The area under curve (AUC) statistics are listed in Table 1, from which we can conclude that under the same SNR, the ascending order of accuracy given by these methods is: PLS, SPCA-1, SPCA-2 and WPCS.

From the experiments on artificial data, a conclusion can be drawn that WPCS method outperforms other methods for feature selection.

Gene Ontology (GO) Analysis

The Gene Ontology (GO) Term Enrichment tool can be used to help discover what those genes may have in common [28]. GOTermFinder is a web-based tool that finds the significant GO terms shared among a list of genes. The analysis of GOTermFinder provides significant information for the biological interpretation of high-throughput experiments. In this paper, our proposed method will be evaluated by GOTermFinder [29], which is publicly available at http://go.princeton.edu/cgi-bin/GOTermFinder. Its threshold parameters are set as following: maximum p-value = 0.01 and minimum number of gene products = 2.

Here, only the main results of GO are given. Figure 2 shows the sample frequency of response to stimulus (GO: 0050896) given by the four methods. From Figure 2(A), WPCS method outperforms the others in all the data sets of shoot samples with six different stresses. Figure 2(B) shows that only in drought-stress data set of root samples, our method is dominated by SPCA-1 and SPCA-2 methods. In other data sets, our method is superior to the others.

Figure 3(A) shows the sample frequency of response to stress (GO: 0006950) in shoot samples. From Figure 3(A), it can be seen that only in drought-stress data set, the PLS method is slightly superior to our method. In other data sets, our method is superior to the other methods. Figure 3(B) shows that only in drought-stress data set of root samples, our WPCS gives a similar result to that of SPCA-1 and SPCA-2 methods, and exceeds that of PLS method. In other data sets, our WPCS method surpasses the others.

The remarkable results are listed in Tables 2–5. The number of genes responding to stimulus (GO: 0050896) selected by the four methods in shoot and root samples are listed in Table 2 and Table 3, respectively.

As Table 2 listed, in shoot samples, WPCS method outperforms the others in all the data sets with six different stresses. As Table 3 listed, in root samples, only in drought-stress data set, WPCS method is dominated by SPCA-2. For other stresses data sets, WPCS outperforms our competitive methods.

Table 4 and Table 5 give the gene numbers and P-value of response to stress (GO: 0006950) selected by the four methods in shoot and root samples, respectively.

To sum up, for all the data sets except drought-stress data set, our method is superior to other methods. For the drought-stress

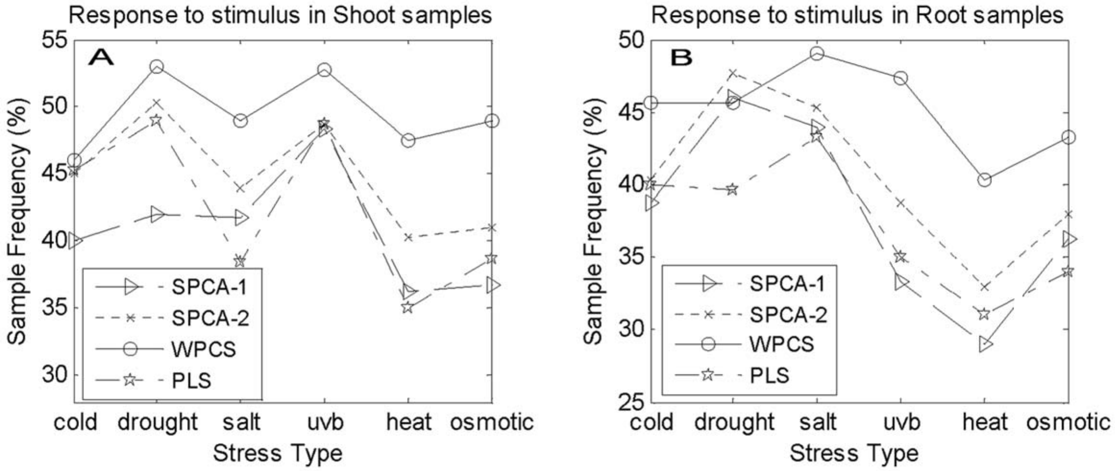

Figure 2. The Response to stimulus (GO: 0050896).
doi:10.1371/journal.pone.0038873.g002

data set in shoot samples, only the PLS method slightly suppresses our WPCS method.

To further study the characteristic genes closely related to the stresses, the cold stress in shoot samples and UV-B stress in root samples are analyzed. Table 6 lists the numbers of response to cold (GO: 0009410) in shoot samples selected by these methods. The background sample frequency of response to cold (GO: 0009410) is 0.9% (276/29887). As Table 6 listed, our method can select more genes than others.

In detail, we compare the genes selected by WPCS with the genes selected using others. Different genes selected using WPCS and neglected by other methods are listed in Table 7. As Table 7 listed, the functions of genes selected using WPCS are closely related with cold stress.

Table 8 gives the numbers of response to light stimulus (GO: 0009416) in root samples selected using these methods. The background sample frequency of response to light stimulus (GO: 0009416) is 1.8% (547/29887).

As Table 8 listed, WPCS can select more genes than others. Moreover, we compare the genes selected by WPCS with the ones by other methods. The genes selected using WPCS and neglected by others are listed in Table 9. As Table 9 listed, the functions of genes selected using WPCS are closely related with UV-B stress.

From the experiments and analyses on gene expression data, a conclusion can be drawn that WPCS method is very efficient and powerful for gene selection.

Conclusion

In this paper, a novel method of gene selection, WPCS, is proposed, that uses the weighted PCs by SVs as the basis of selection. The idea of WPCS is clearly shown. WPCS works as follows. First, it obtains the PCs of SPCA and SVs. Second, using the SVs as the weights of PCs, it obtains the WPC. Then, it sums the absolute value of the WPC in row, and sorts the sum in descending order. Finally, it selects the genes corresponding to the top part of the sum as the characteristic genes. A large number of experiments on artificial data and gene expression data demonstrate that the proposed WPCS method outperforms the state-of-the-art gene selection methods. For gene expression data, WPCS can extract more characteristic genes in response to abiotic stresses than the other methods.

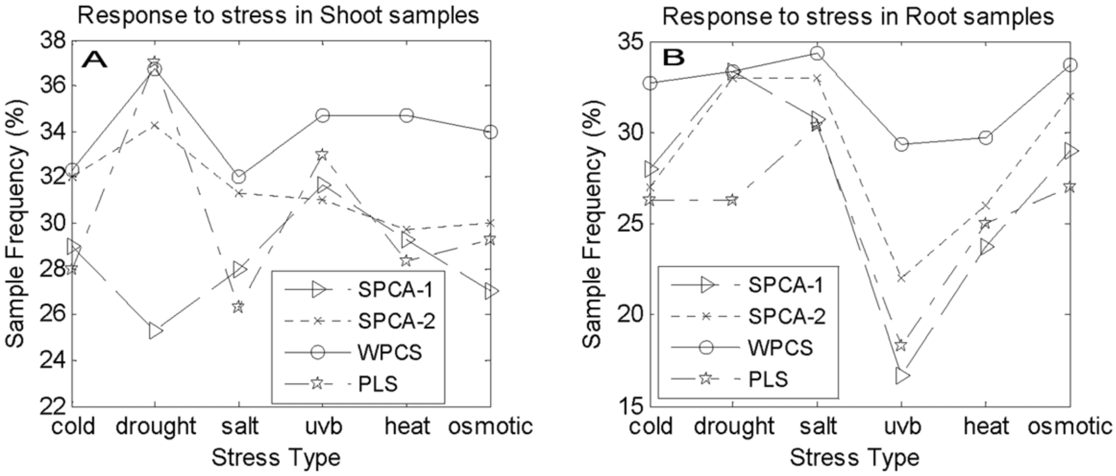

Figure 3. The Response to stress (GO: 0006950).
doi:10.1371/journal.pone.0038873.g003

Table 2. Response to stimulus (GO: 0050896) in shoot samples.

Stress	SPCA-1		SPCA-2		WPCS		PLS	
Type	SF[♀]	PV	SF	PV	SF	PV	SF	PV
Cold	120/300* 40.0%	2.39E-21	135/300 45.0%	2.19E-30	**138/300 46.0%**	2.33E-32	136/300 45.3%	5.19E-31
Drought	126/300 42.0%	8.50E-25	151/300 50.3%	1.62E-41	**159/300 53.0%**	1.17E-47	147/300 49.0%	2.36E-38
Salt	125/300 41.7%	3.27E-24	132/300 44.0%	1.19E-28	**147/300 49.0%**	8.34E-39	115/300 38.5%	8.57E-19
UV-B	145/300 48.3%	6.44E-37	146/300 48.7%	1.31E-37	**158/300 52.7%**	7.51E-47	146/300 48.7%	1.20E-37
Heat	109/300 36.3%	2.05E-15	121/300 40.3%	7.12E-22	**142/300 47.5%**	2.89E-35	105/300 35.0%	1.24E-13
Osmotic	110/300 36.7%	6.18E-16	123/300 41.0%	7.06E-23	**147/300 49.0%**	2.19E-38	116/300 38.7%	4.96E-19

Note: The response to stimulus on characteristic genes are shown, whose background frequency in TAIR set is 4570/29887 (15.3%), where 4570/29887 denotes having 4570 genes to respond to stimulus in whole 29887 genes set.

[♀]SF: sample frequency, PV: P-value.

*In the table, the sample frequency, e.g. 120/300, denotes the method select 300 genes, in which there are 120 genes responding to stimulus.

doi:10.1371/journal.pone.0038873.t002

Materials and Methods

Artificial Data

The artificial data are in R^p with $p=2000$ and generated as $\mathbf{A}\sim(0,\Sigma_4)$. Let $\tilde{\mathbf{v}}_1\sim\tilde{\mathbf{v}}_4$ be four 2000-dimensional vectors, such that $\tilde{\mathbf{v}}_{1k}=1,\ \ k=1,\cdots,50$, and $\tilde{\mathbf{v}}_{1k}=0,\ \ k=51,\cdots,2000$; $\tilde{\mathbf{v}}_{2k}=1,\ \ k=51,\cdots,100$, and $\tilde{\mathbf{v}}_{2k}=0,\ \ k\neq 51,\cdots,100$; $\tilde{\mathbf{v}}_{3k}=1,\ \ k=101,\cdots,150$, and $\tilde{\mathbf{v}}_{3k}=0,\ \ k\neq 101,\cdots,150$; and $\tilde{\mathbf{v}}_{4k}=1,\ \ k=151,\cdots,200$, and $\tilde{\mathbf{v}}_{4k}=0,\ \ k\neq 151,\cdots,200$. Let $\mathbf{E}$ be 2000-dimensional noise matrix, $E\sim N(0,1)$. Then the noise matrix is added to $\tilde{\mathbf{v}}$ with different Signal-to-Noise Ratios (SNR). The first four eigenvectors of $\mathbf{\Sigma}_4$ are chosen to be$\mathbf{v}_k=\tilde{\mathbf{v}}_k/\|\tilde{\mathbf{v}}_k\|,\ \ k=1,2,3,4$. To make these four eigenvectors dominate, we let the eigenvalues be$c_1=400,c_2=300,\ c_3=200,\ c_4=100$ and $c_k=1$ for $k=5,\cdots,2000$. Then the simulation scheme in [30] is used to generate the artificial data, which include ten samples in each test.

Gene Expression Data

The raw data include two classes: roots and shoots in each stress, which were downloaded from NASCArrays [http://affy.arabidopsis.info/] [31], reference numbers are: control, NASCArrays-137; cold stress, NASCArrays-138; osmotic stress, NASCArrays-139; salt stress, NASCArrays-140; drought stress, NASCArrays-141; UV-B light stress, NASCArrays-144; and heat stress, NASCArrays-146. The sample numbers of each stress type are listed in Table 10. There are 22810 genes in each sample. The data are adjusted for background of optical noise using the GC-RMA software by Wu *et al.* [32] and normalized using quantile normalization. The results of GC-RMA are gathered in a matrix for further processing.

Selection of the Parameters

In SPCA, we take l_0-norm penalty and set $\gamma=0.1$. In PLS, only the first component is used. For the sake of comparison, on gene expression data, 300 genes are roughly selected by all the methods.

Singular Value Decomposition (SVD)

In this subsection, the details of the WPCS method are presented. Let $\mathbf{A}$ denote an $n\times p$ matrix of real-valued gene expression data, which consists of p genes in n samples. In the case of gene expression data, a_{ij} is the expression level of the j-th gene in the i-th sample. The elements of the j-th column in $\mathbf{A}$ form the n-dimensional vector $\mathbf{r}_j$, which is referred to as the transcriptional response of the j-th gene. Correspondingly, the elements of the i-th row in $\mathbf{A}$ form the p-dimensional vector$\mathbf{s}_i$, referred to as the expression profile of the i-th sample. Usually$p>>n$, so it is a classical small-sample-size problem. To integrate SPCA and SV, the singular value decomposition (SVD) and the SPCA via cardinality penalty (l_0-penalty) are introduced as follows.

If the variables contained in the columns of $\mathbf{A}$ with rank $r\leq\min(n,p)$ are centered, the equation for SVD of $\mathbf{A}$ is as follows:

$$\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{1}$$

Table 3. Response to stimulus (GO: 0050896) in root samples.

Stress	SPCA-1		SPCA-2		WPCS		PLS	
Type	SF	PV	SF	PV	SF	PV	SF	PV
Cold	116/300 38.7%	4.79E-19	121/300 40.3%	1.12E-21	**137/300 45.7%**	1.78E-31	120/300 40.0%	3.46E-21
Drought	138/300 46.0%	3.62E-32	**143/300 47.7%**	1.48E-35	137/300 45.7%	1.55E-31	119/300 39.7%	1.12E-20
Salt	132/300 44.0%	2.52E-28	136/300 45.3%	6.80E-31	**147/300 49.0%**	2.06E-38	130/300 43.3%	4.23E-27
UV-B	100/300 33.3%	2.60E-11	116/300 38.7%	4.79E-19	**142/300 47.3%**	6.67E-35	105/300 35.0%	1.34E-13
Heat	87/300 29.0%	3.30E-06	99/300 33.0%	4.75E-11	**121/300 40.3%**	6.93E-22	93/300 31.0%	1.41E-08
Osmotic	109/300 36.3%	1.92E-15	114/300 38.0%	5.57E-18	**130/300 43.3%**	4.43E-27	102/300 34.0%	3.29E-12

doi:10.1371/journal.pone.0038873.t003

Table 4. Response to stress (GO: 0006950) in shoot samples.

Stress	SPCA-1		SPCA-2		WPCS		PLS	
Type	SF	PV	SF	PV	SF	PV	SF	PV
Cold	87/300 29.0%	5.19E-25	96/300 32.0%	1.52E-31	**97/300 32.3%**	2.61E-32	84/300 28.0%	6.72E-23
Drought	76/300 25.3%	9.02E-18	103/300 34.3%	5.07E-37	110/300 36.7%	7.81E-43	**111/300 37.0%**	1.68E-43
Salt	84/300 28.0%	6.06E-23	94/300 31.3%	3.57E-30	**96/300 32.0%**	1.15E-31	79/300 26.3%	8.58E-20
UV-B	95/300 31.7%	1.28E-30	93/300 31.0%	4.13E-29	**104/300 34.7%**	8.09E-38	99/300 33.0%	1.07E-33
Heat	88/300 29.3%	1.46E-25	89/300 29.7%	2.20E-26	**104/300 34.7%**	5.61E-38	85/300 28.3%	1.31E-23
Osmotic	81/300 27.0%	7.48E-21	90/300 30.0%	5.38E-27	**102/300 34.0%**	4.37E-36	88/300 29.3%	1.32E-25

Note: The response to stress (GO: 0006950) obtained by GO Term Enrichment Analysis are shown, whose background frequency in TAIR set is 2351/29887 (7.9%), where 2351/29887 denotes having 2351 genes to respond to stress in whole 29887 genes set.

doi:10.1371/journal.pone.0038873.t004

where $\mathbf{U}$ is an $n\times r$ matrix of left singular vectors with $\mathbf{U}^T\mathbf{U}=\mathbf{I}$, $\mathbf{V}$ is a $p\times r$ matrix of right singular vectors with $\mathbf{V}^T\mathbf{V}=\mathbf{I}$, and $\mathbf{D}$ is a diagonal matrix of singular values. Let $\mathbf{u}_k$ denote column k-th of $\mathbf{U}$, let $\mathbf{v}_k$ denote column k-th of $\mathbf{V}$, and note that d_k denotes the k-th diagonal element of the matrix $\mathbf{D}$. According to Eckart *et al.* [33],

$$\mathbf{A}^{(l)}=\sum_{k=1}^{l}\mathbf{u}_k d_k \mathbf{v}_k^T, \tag{2}$$

which is the closest rank-l matrix to $\mathbf{A}$. The term "closest" means that $\mathbf{A}^{(l)}$ minimizes the square error sum between the elements of $\mathbf{A}$ and $\mathbf{A}^{(l)}$.

Sparse Principal Component Analysis (SPCA)

The results given with l_0- and l_1- norm penalty in SPCA are similar, which is also shown in [20]. Since the l_0-norm is faster than l_1-norm, l_0-norm penalty is taken on SPCA. Let $\mathbf{F}=\mathbf{UD}$ and $\mathbf{Z}=\mathbf{V}$, eq.(1) can be written as follows:

$$\mathbf{A}=\mathbf{F}\mathbf{Z}^T, \tag{3}$$

where $\mathbf{Z}^T\mathbf{Z}=\mathbf{I}$ and $\mathbf{F}^T\mathbf{F}=\boldsymbol{\Sigma}=\mathbf{A}^T\mathbf{A}$. It is the classical PCA formation.

Extracting one principal component (PCs) amounts to computing the dominant eigenvector of $\boldsymbol{\Sigma}$ (or, equivalently, dominant right singular vector of $\mathbf{A}$). That is, PCA seeks to project the data onto the linear combination of variables that maximizes the sample variance. It is well-known that the solution to this problem is given by the right singular vector of $\mathbf{A}$. In general, PCs is not expected to have many zero coefficients. So, to makes it easy to interpret PCs without subjective judgment, Sparse PCA proposed by Journée *et al.* in [20] is used to generate the sparse PCs.

Let us consider the optimization problem.

$$\vartheta_{l_0}(\gamma)\overset{def}{=}\max_{z\in B^p} z^T\Sigma z-\gamma\|z\|_0, \tag{4}$$

with sparsity-controlling parameter $\gamma\geq 0$, $\|\cdot\|_0$ denotes the l_0-norm, that is, the number of non-zero components (cardinality).

According to [20], eq.(4) can be rewritten as follows:

$$\vartheta_{l_0}(\gamma)=\max_{v\in B^n}\max_{z\in B^p}\left(v^T\mathbf{A}z\right)^2-\gamma\|z\|_0, \tag{5}$$

where the maximization with respect to $z\in B^p$ for a fixed $v\in B^n$ has the closed form solution

$$z_i^*=z_i^*(\gamma)=\frac{[\mathrm{sign}((a_i^Tv)^2-\gamma)]_+a_i^Tv}{\sqrt{\sum_{k=1}^{p}[\mathrm{sign}((a_k^Tv)^2-\gamma)]_+(a_k^Tv)^2}},\quad i=1,\cdots,p. \tag{6}$$

According to [20], eq.(5) can be cast in the following form:

$$\vartheta_{l_0}(\gamma)=\max_{x\in B^n}\sum_{i=1}^{p}\left[\left(a_i^Tv\right)^2-\gamma\right]_+. \tag{7}$$

Table 5. Response to stress (GO: 0006950) in root samples.

Stress	SPCA-1		SPCA-2		WPCS		PLS	
Type	SF	PV	SF	PV	SF	PV	SF	PV
Cold	84/300 28.0%	7.21E-23	81/300 27.0%	9.01E-21	**98/300 32.7%**	6.51E-33	79/300 26.3%	1.48E-19
Drought	100/300 33.3%	1.68E-34	99/300 33.0%	1.03E-33	**100/300 33.3%**	1.58E-34	79/300 26.3%	1.34E-19
Salt	92/300 30.7%	1.79E-28	99/300 33.0%	9.23E-34	**103/300 34.3%**	6.50E-37	91/300 30.3%	9.12E-28
UV-B	50/300 16.7%	1.30E-04	66/300 22.0%	5.55E-12	**88/300 29.3%**	1.33E-25	55/300 18.3%	1.09E-06
Heat	71/300 23.7%	1.11E-14	78/300 26.0%	4.21E-19	**89/300 29.7%**	2.14E-26	75/300 25.0%	3.20E-17
Osmotic	87/300 29.0%	6.81E-25	96/300 32.0%	1.89E-31	**101/300 33.7%**	2.52E-35	81/300 27.0%	6.93E-21

doi:10.1371/journal.pone.0038873.t005

Table 6. The numbers of response to cold (GO: 0009410) in shoot samples.

Method	SPCA-1	SPCA-2	WPCS	PLS
Number and percent	38 genes,12.7%	45 genes,15.0%	48 genes,16.0%	33 genes,11.0%
P-value	1.45E-27	2.43E-36	2.73E-40	1.18E-21

doi:10.1371/journal.pone.0038873.t006

Here, for$t\in R$, sign(t) denotes the sign of the argument and $t_+ = \max\{0,t\}$.

For large enoughγ, $z^*(\gamma)=0$. Since.

$$\begin{aligned}\max_{z\neq 0}\frac{\|Az\|_2^2}{\|z\|_0} &= \max_{z\neq 0}\frac{\left\|\sum_i z_i a_i\right\|_2^2}{\|z\|_0}\\ &\leq \max_{z\neq 0}\frac{\sum_i |z_i|^2\|a_i\|_2^2}{\sum_i |z_i|^2}\\ &= \max_i\|a_i\|_2^2 = \|a_i{}^*\|_2^2,\end{aligned} \tag{8}$$

we get.

$$\|Az\|_2^2 - \gamma\|z\|_0 < 0, \tag{9}$$

for all nonzero vector z, when $\gamma > \|a_i{}^*\|_2^2$.So, this derivation assumes that $\gamma < \|a_i{}^*\|_2^2$, then.

$$\left(a_i^T x\right)^2 - \gamma > 0. \tag{10}$$

Otherwise$\gamma \geq \|a_i\|_2^2$, so

$$z_i^*(\gamma) = 0, \qquad i = 1,\ldots,p. \tag{11}$$

The Idea of the Proposed Method

The residual sum of squares (RESS) can be used for evaluating the quality of the reconstitution of $\mathbf{A}$ with M PCs of SPCA. According to [34], it can be computed as follows:

$$\begin{aligned}\mathrm{RESS}_M &= \left\|\mathbf{A} - \hat{\mathbf{A}}^{[M]}\right\|^2\\ &= \Psi - \sum_{l=1}^{M}\lambda_l,\end{aligned} \tag{12}$$

where $\hat{\mathbf{A}}^{[M]}$ is a reconstitution of $\mathbf{A}$, Ψ denotes the sum of all the squared elements, and λ_l is the singular value of the l-th component. The smaller the value of RESS is, the better the SPCA model is. From eq.(12), we can see that a larger M may give a better estimation of $\hat{\mathbf{A}}^{[M]}$. If M takes value r (rank of $\mathbf{A}$),

Table 7. Different genes of response to cold (GO: 0009410) in shoot samples.

Gene No.	Function of Gene
At1g21910	Participates in plant developmental processes as well as biotic and/or abiotic stress signaling.
At1g22770	Regulates several developmental processes, such as circadian clock, carbohydrate metabolism, and cold stress response.
At1g29395	Expression is induced by short-term cold-treatment, water deprivation, and abscisic acid (ABA) treatment.
At2g19450	Role in senescence and seed development induced by cold-stress.
At2g25930	Temperature stress reduced the pyk20 transcript level.
At2g28900	Predominantly expressed in leaves and is also inducible by cold treatment.
At2g33380	Plays a role as a peroxygenase involved in oxylipin metabolism during biotic and abiotic stress.
At2g38470	Involved in response to various abiotic stresses
At2g47180	Increases tolerance to chilling stress
At3g05880	Induced by low temperatures, dehydration and salt stress.
At3g48360	Mediates multiple responses to nutrients, stresses, and hormones.
At3g53990	Low temperature and salt responsive protein family.
At4g30650	Low temperature and salt expression protein homologous.
At4g30660	Putative low temperature and salt responsive protein.
At4g37610	Under cold stress indicates increased expression.
At5g52300	Induced by low temperature, exogenous abscisic acid (ABA) and drought.
At5g57560	Controlling tolerance to cold stress

doi:10.1371/journal.pone.0038873.t007

Table 8. The numbers of response to light stimulus (GO: 0009416) in root samples.

Method	SPCA-1	SPCA-2	WPCS	PLS
Number and percent	17 genes, 5.7%	20 genes, 6.7%	24 genes, 8.0%	17 genes, 5.7%
P-value	1.74E-02	2.90E-04	7.42E-07	1.55E-02

doi:10.1371/journal.pone.0038873.t008

$$\begin{aligned}\mathrm{RESS}_r &= \left\| \mathbf{A} - \hat{\mathbf{A}}^{[r]} \right\|^2 \\ &= \Psi - \sum_{l=1}^{r} \lambda_l,\end{aligned} \tag{13}$$

where the matrix **A** can be perfectly reconstituted. Let $M=1$, the RESS_1 can be expressed as follows:

$$\begin{aligned}\mathrm{RESS}_1 &= \left\| \mathbf{A} - \hat{\mathbf{A}}^{[1]} \right\|^2 \\ &= \Psi - \lambda_1 \\ &= \Psi - \sum_{l=1}^{r} \lambda_l + \sum_{l=2}^{r} \lambda_l,\end{aligned} \tag{14}$$

Substituting eq.(13) into eq.(14), the RESS_1 can be obtained as follows:

$$\mathrm{RESS}_1 = \mathrm{RESS}_r + \sum_{l=2}^{r} \lambda_l. \tag{15}$$

As eq.(15) shown, if only one PC is used to reconstitute the matrix **A**, RESS_1 may be larger $\sum_{l=2}^{r} \lambda_l$ than RESS_r. So, if only the first PC is used for characteristic gene selection, some important information may be lost, especially when the second one or two SVs are approximately equal to the first one. In order to obtain a better reconstitution, all the PCs of SPCA need to be utilized.

In SPCA, **F** is the matrix of factor scores and **Z** is a loading matrix of the principal components (PCs), which transforms the original data matrix into factor scores. The data matrix **A**, factor scores matrix **F** and PCs **Z** are shown in Figure 4.

As Figure 4 shown, the PCs **Z** give the coefficients of the linear combinations used to compute the factors scores **F**. So the bigger the absolute value of the elements in PCs **Z** is, the more contribution it gives for the factor scores matrix, the more important the corresponding gene in **A** is. So the characteristic genes can be selected according to the PCs **Z**.

Let.

$$Z_i = \left[z_{1i}, z_{2i}, \cdots, z_{pi} \right]^T, i = 1, \cdots, r \tag{16}$$

denote the *i*-th PC, and then the PCs can be given as the follows:

$$\mathbf{Z} = [Z_1, Z_2, \cdots, Z_r]. \tag{17}$$

Substituting $\mathbf{F} = \mathbf{UD}$ into eq.(3), it can be reformed as follows:

$$\mathbf{A} = \mathbf{UDZ}^T, \tag{18}$$

where **D** is the diagonal matrix of singular values. Let

$$\hat{\mathbf{Z}}^T = \mathbf{DZ}^T, \tag{19}$$

eq.(18) can be reformed as follows:

$$\mathbf{A} = \mathbf{U}\hat{\mathbf{Z}}^T. \tag{20}$$

Substituting eq.(17) into eq.(19),

Table 9. Different genes of response to light stimulus (GO:0009416) in root samples.

Gene No.	Function of Gene
At2g29500	HSP20-like chaperones superfamily protein.
At3g54890	Encodes a component of the light harvesting complex associated with photosystem I.
At3g55120	Catalyzes the conversion of chalcones into flavanones.
At5g02810	Acts as transcriptional repressor of CCA1 and LHY.
At5g12030	Encodes a cytosolic small heat shock protein with chaperone activity that is induced by heat and high light intensity stress.
At5g15960	stress-responsive protein (KIN1).
At5g24470	Encodes a pseudo-response regulator whose mutation affects various circadian-associated biological events such as red light sensitivity of seedlings during early photomorphogenesis.
At5g45340	abscisic acid 8′-hydroxylase 3.

doi:10.1371/journal.pone.0038873.t009

Table 10. The number of each stress type in the raw data.

Stress Type	cold	drought	salt	UV-B	heat	osmotic	control
Number	6	7	6	7	8	6	8

doi:10.1371/journal.pone.0038873.t010

$$\begin{aligned}\hat{\mathbf{Z}} &= \left[\mathbf{D}\mathbf{Z}^T\right]^T \\ &= \mathbf{Z}\mathbf{D}^T \\ &= [Z_1, Z_2, \cdots, Z_r]\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_r \end{bmatrix} \\ &= [d_1 Z_1, d_2 Z_2, \cdots, d_r Z_r].\end{aligned} \tag{21}$$

The matrix $\hat{\mathbf{Z}}$ is referred to as weighted PCs (WPC), which can be obtained via weighting PCs by the diagonal matrix of SVs (WPCS).

Substituting eq.(16) into eq.(21), the WPC can be reformed as follows:

$$\hat{\mathbf{Z}} = \begin{bmatrix} d_1 z_{11} & d_2 z_{12} & \cdots & d_r z_{1r} \\ d_1 z_{21} & d_2 z_{22} & \cdots & d_r z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ d_1 z_{p1} & d_2 z_{p2} & \cdots & d_r z_{pr} \end{bmatrix}. \tag{22}$$

As the absolute value of the *i*-th row of WPC $\hat{\mathbf{Z}}$ somewhat denotes the importance of the *i*-th gene, the absolute value sum of all the entries in the *i*-th row as the evaluating vector EV, which can be expressed as follows:

$$EV = \begin{bmatrix} \sum_{i=1}^{r} |d_1 z_{1i}| \\ \sum_{i=1}^{r} |d_2 z_{2i}| \\ \vdots \\ \sum_{i=1}^{r} |d_p z_{pi}| \end{bmatrix}. \tag{23}$$

In particular, if the dimensionality of the gene data is p, the EV has p entries. After sorting the evaluating vector EV, the genes corresponding to the first m largest entries can be selected as characteristic genes.

In summary, the main steps of WPCS method are shown as follows.

(1) Given the observation matrix $\mathbf{A} \in R^{n \times p}$, $\gamma \geq 0$, $num > 0$.

(2) To obtain the PCs $\mathbf{Z}$, execute SPCA on the $\mathbf{A}$.

(3) To obtain the SVs, execute the SVD.

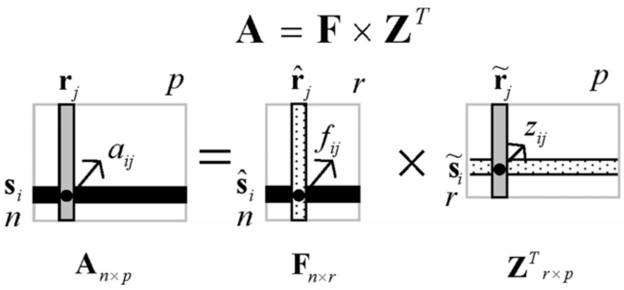

Figure 4. The graphical depiction of SPCA of a matrix A with factor scores F and PCs Z. In this figure, with factor scores $\mathbf{F}$ and PCs $\mathbf{Z}$. $\tilde{\mathbf{r}}_j$ is the row vector of PCs $\mathbf{Z}$ the *j*-th gene, which transforms the original data vector $\mathbf{r}_j$ into factor scores $\hat{\mathbf{r}}_j$. Correspondingly, $\tilde{\mathbf{s}}_i$ is the column vector of PCs $\mathbf{Z}$, which transforms the original data vector $\mathbf{s}_i$ into factor scores $\hat{\mathbf{s}}_i$.
doi:10.1371/journal.pone.0038873.g004

(4) Obtain the WPC $\hat{\mathbf{Z}}$ via multiplying PCs by the diagonal matrix of SVs.

(5) Obtain the evaluating vector EV by summing the absolute value of WPC $\hat{\mathbf{Z}}$ in row.

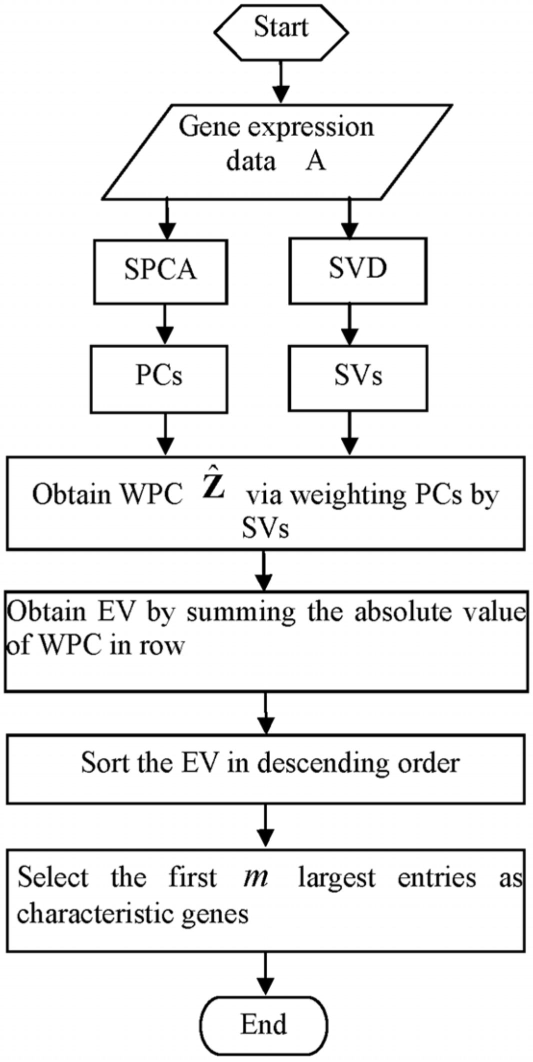
Start
Gene expression data A
SPCA
SVD
PCs
SVs
Obtain WPC $\hat{\mathbf{Z}}$ via weighting PCs by SVs
Obtain EV by summing the absolute value of WPC in row
Sort the EV in descending order
Select the first m largest entries as characteristic genes
End

Figure 5. Workflow diagram of WPCS.
doi:10.1371/journal.pone.0038873.g005

(7) Sort the EV in descending order.

(7) Select the genes corresponding to the first m largest entries as characteristic genes.

The workflow diagram of our method is shown in Figure 5.

Acknowledgments

WPCS source codes and data set are also available at http://www.yongxu.org/code/LJX_PONE.rar.

Author Contributions

Conceived and designed the experiments: JXL. Performed the experiments: YX CHZ. Analyzed the data: YW JYY. Wrote the paper: JXL YX CHZ.

References

1. Hirayama T, Shinozaki K (2010) Research on plant abiotic stress responses in the post-genome era: past, present and future. The Plant Journal 61: 1041–1052.
2. Liu H, Rahman A, Semino-Mora C, Doi SQ, Dubois A (2008) Specific and sensitive detection of H. pylori in biological specimens by real-time RT-PCR and in situ hybridization. Plos One 3: e2689.
3. Maan NS, Maan S, Nomikou K, Johnson DJ, El Harrak M, et al. (2010) RT-PCR assays for seven serotypes of epizootic haemorrhagic disease virus & their use to type strains from the Mediterranean Region and North America. Plos One 5: e12782.
4. Blevins T (2010) Northern blotting techniques for small RNAs. Methods Mol Biol 631: 87–107.
5. Josefsen K, Nielsen H (2011) Northern blotting analysis. Methods in molecular biology (Clifton, NJ) 703: 87–105.
6. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.
7. Heller MJ (2002) DNA microarray technology: devices, systems, and applications. Annual Review of Biomedical Engineering 4: 129–153.
8. Sato F, Tsuchiya S, Terasawa K, Tsujimoto G (2009) Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. Plos One 4: e5540.
9. Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, et al. (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. The Plant Journal 31: 279–292.
10. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, et al. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. The Plant Journal 50: 347–363.
11. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Statistical Science 18: 71–103.
12. Saeys Y, Inza I, Larra aga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23: 2507–2517.
13. Sampson DL, Parker TJ, Upton Z, Hurst CP (2011) A Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning Approaches. Plos One 6: e24973.
14. Park MS, Choi JY (2009) Theoretical analysis on feature extraction capability of class-augmented PCA. Pattern recognition 42: 2353–2362.
15. Ma S, Kosorok MR (2009) Identification of differential gene pathways with principal component analysis. Bioinformatics 25: 882–889.
16. De Haan J, Piek E, van Schaik R, De Vlieg J, Bauerschmidt S, et al. (2010) Integrating gene expression and GO classification for PCA by preclustering. BMC bioinformatics 11: 158.
17. Musumarra G, Barresi V, Condorelli DF, Fortuna CG, Scire S (2004) Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysis. Journal of Chemometrics 18: 125–132.
18. Boulesteix AL, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8: 32–44.
19. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. Journal of computational and graphical statistics 15: 265–286.
20. Journée M, Nesterov Y, Richtarik P, Sepulchre R (2010) Generalized power method for sparse principal component analysis. The Journal of Machine Learning Research 11: 517–553.
21. Lai ZH, Wan MH, Jin Z, Yang JA (2011) Sparse two-dimensional local discriminant projections for feature extraction. Neurocomputing 74: 629–637.
22. Luss R, d'Aspremont A (2010) Clustering and feature selection using sparse principal component analysis. Optimization and Engineering 11: 145–157.
23. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10: 515–534.
24. Liu JX, Zheng CH, Xu Y (2012) Extracting plants core genes responding to abiotic stresses by penalized matrix decomposition. Computers in Biology and Medicine: DOI = '10.1016/j.compbiomed.2012.1002.1002'.
25. Le Cao KA, Boitard S, Besse P (2011) Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC bioinformatics 12: 253.
26. Liu JX, Zheng CH, Xu Y (2011) Lasso Logistic Regression based Approach for Extracting Plants Coregenes Responding to Abiotic Stresses. IWACI. Wuhan, CN. 463∼466.
27. Yang H, Yang JY (2003) Why can LDA be performed in PCA transformed space? Pattern recognition 36: 563–566.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25: 25–29.
29. Boyle EI, Weng SA, Gollub J, Jin H, Botstein D, et al. (2004) TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.
30. Shen H, Huang JZ (2008) Sparse principal component analysis via regularized low rank matrix approximation. Journal of multivariate analysis 99: 1015–1034.
31. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, et al. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Research 32: D575–D577.
32. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 99: 909–917.
33. Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. Psychometrika 1: 211–218.
34. Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdisciplinary reviews: Computational Statistics 2: 433–459.