HB-PLS: A statistical method for identifying biological process or pathway regulators by integrating Huber loss and Berhu penalty with partial least squares regression

Wenping Deng¹, Kui Zhang², Cheng He³, Sanzhen Liu³ and Hairong Wei¹

INTRODUCTION

In a gene regulatory network (GRN), a node corresponds to a gene and an edge represents a directional regulatory relationship between a transcription factor (TF) and a target gene. Understanding the regulatory relationships among genes in GRNs can help elucidate the various biological processes and underlying mechanisms in a variety of organisms. Although experiments can be conducted to acquire evidence of gene regulatory interactions, these are labor-intensive and time-consuming. In the past two decades, the advent of high-throughput technologies including microarray and RNA-Seq, have generated an enormous wealth of transcriptomic data. As the throughput technologies including microarray and RNA-Seq, have generated an enormous wealth of transcriptomic data. As the throughput technologies including microarray and RNA-Seq, have generated an enormous wealth of transcriptomic data. As the throughput technologies including microarray and RNA-Seq, have generated an enormous wealth of transcriptomic data.
empirical distribution of the mutual information for each pair of
genes and eliminates edges with scores that are not statistically
significant. ARACNE is similar to RN; however, ARACNE makes
use of the data processing inequality (DPI) to eliminate the least
significant edge of a triplet of genes, which decreases the false
positive rate of the inferred network. MRNET\cite{30} employs the
maximum relevance and minimum redundancy feature selection
method to infer GRNs. Finally, triple-gene mutual
interaction (TGMI) uses condition mutual information to
evaluate triple gene blocks to infer GRNs\cite{21}. Information
theory-based methods are used extensively for constructing
GRNs and for building large networks because they have a low
computational complexity and are able to capture nonlinear
dependencies. However, there are also disadvantages in using
mutual information, including high false-positive rates\cite{22} and
the inability to differentiate positive (activating), negative
(inhibiting), and indirect regulatory relationships.
Reconstruction of the transcriptional regulatory network can be
implemented by the neighborhood selection method.
Neighborhood selection\cite{23} is a sub-problem of covariation
selection. Assume $\Gamma$ is a set containing all of the variables
(genes), the neighborhood $ne_a$ of a variable $a \in \Gamma$ is the smallest
subset of $\Gamma \setminus \{a\}$ such that, given all variables in $ne_a$, variable $a$
is conditionally independent of all remaining variables. Given $n$
i.i.d. observations of $\Gamma$, neighborhood selection aims to
estimate the neighborhood of each variable in $\Gamma$ individually.
The neighborhood selection problem can be cast as a multiple
linear regression problem and solved by regularized methods.

Following the differential equation in\cite{24}, the expression
levels of a target gene $y$ and the expression levels of the TF
genes $x$ form a linear relationship:

$$y_i = \beta_0 + x_i^T \beta + \epsilon_i \quad i = 1, 2, …, n$$

where $n$ is the number of samples, $x_i = (x_{i1}, …, x_{ip})^T$ is the
expression level of $p$ TF genes, and $y_i$ is the expression level of the
target gene in sample $i$. $\beta_0$ is the intercept and $\beta = (\beta_1, …, \beta_p)^T$
are the associated regression coefficients; if any $\beta_j \neq 0$
($j = 1, …, p$) then TF gene $j$ regulates target gene $i$. $\epsilon_i$ are
independent and identically distributed random errors with mean 0
and variance $\sigma^2$. The method to get an estimate of $\beta$ and $\beta_0$ is to
transform the parameter estimate to a convex optimization problem:

$$\hat{\beta} = \text{argmin}_\beta L(\beta) = \text{argmin}_\beta \sum_{i=1}^{n} L(y_i - \beta_0 - x_i^T \beta) + \lambda P(\beta)$$

where $L(\cdot)$ is a loss function, $P(\cdot)$ is a penalization function, and
$\lambda > 0$ is a tuning parameter which determines the importance
of penalization. Different loss functions, penalization functions, and
methods for determining $\lambda$ have been proposed in the literature.
Ordinary least squares (OLS) is the simplest method with a square
loss function $L(y_i - \beta_0 - x_i^T \beta) = (y_i - \beta_0 - x_i^T \beta)^2$ and no
penalization function. The OLS estimator is unbiased\cite{25}. However,
since it is common for the number of genes, $p$, to be much larger
than the number of samples, $n$ (i.e. $p \gg n$) in any given gene
expression data set, there is no unique solution for OLS. Even
when $n > p$, OLS estimation features high variance. To tackle
these problems, ridge regression\cite{26} adds a $\ell_2$ penalty,
$P(\beta) = \sum_p \beta_j^2$, on the coefficients which introduces a bias but
reduces the variance of the estimated $\hat{\beta}$. In ridge regression, there
is a unique solution even for the $p \gg n$ case. Least absolute
shrinkage and selection operator (LASSO)\cite{27} is similar to ridge
regression, except the $\ell_2$ penalty in ridge regression is replaced by
the $\ell_1$ penalty, $P(\beta) = \sum_p |\beta_j|$.\footnote{\cite{28}}

The main benefit of least absolute shrinkage and selection
operator (LASSO) is that it performs variable selection and
regularization simultaneously thereby generating a sparse
solution, a desirable property for constructing GRNs. When
LASSO is used for selecting regulatory TFs for a target gene, there
are two potential limitations. First, if several TF genes are
correlated and have large effects on the target gene, LASSO has a
tendency to choose only one TF gene while zeroing out the
other TF genes. Second, some studies\cite{29} state that LASSO does
not have oracle properties; that is, it does not have the
capability to identify the correct subset of true variables or to
have an optimal estimation rate. It is claimed that there are
cases where a given $\lambda$ that leads to optimal estimation rate
ends up with an inconsistent selection of variables. For the first
limitation, Zou and Hastie\cite{30} proposed elastic net, in which the
penalty is a mixture of LASSO and ridge regressions:

$$P(\beta) = a \sum_p |\beta_j| + \frac{1-a}{2} \sum_p \beta_j^2$$

where $a(0 < a < 1)$ is called the elastic net mixing parameter. When $a = 1$, the elastic net
penalty becomes the LASSO penalty; when $a = 0$, the elastic
net penalty becomes the ridge penalty. For the second
limitation, adaptive LASSO\cite{28} was proposed as a regularization
method, which enjoys the oracle properties. The penalty function
for adaptive LASSO is:

$$P(\beta) = \sum_p \beta_j |\hat{\beta}_j|$$

where adaptive weight $\hat{\beta}_j$ is an initial estimate of the
coefficients obtained through ridge regression or LASSO; $\gamma$ is a
positive constant, and is usually set to 1. It is evident that
adaptive LASSO penalizes more those coefficients with lower
initial estimates.

It is well known that the square loss function is sensitive to
heavy-tailed errors or outliers. Therefore, adaptive LASSO may
fail to produce reliable estimates for datasets with heavy-tailed
errors or outliers, which commonly appear in gene expression
datasets. One possible remedy is to remove influential
observations from the data before fitting a model, but it is
difficult to differentiate true outliers from normal data. The
other method is to use robust regression. Wang et al.\cite{29}
combined the least absolute deviation (LAD) and weighted
LASSO penalty to produce the LAD-LASSO method. The
objective function is:

$$\sum_{i=1}^{n} \left[|y_i - \beta_0 - x_i^T \beta| + \lambda L(\beta_j)\right]$$

With this LAD loss, LAD-LASSO is more robust than OLS to
unusual $y$ values, but it is sensitive to high leverage outliers.
Moreover, LAD estimation degrades the efficiency of the
resulting estimation if the error distribution is not heavy
tailed\cite{31}. To achieve both robustness and efficiency, Lambert-
Lacroix and Zwald 2011\cite{32}, proposed Huber-LASSO, which
combined the Huber loss function and a weighted LASSO
penalty. The Huber function (see Materials and Methods) is a
hybrid of squared error for relatively small errors and absolute
error for relatively large ones. Owen 2007\cite{33} proposed the use
of the Huber function as a loss function and the use of a
reversed version of Huber’s criterion, called Berhu, as a penalty
function. For the Berhu penalty (see Materials and Methods),
relatively small coefficients contribute their $\ell_2$ norm to the
penalty while larger ones cause it to grow quadratically. This
Berhu penalty sets some coefficients to 0, like LASSO, while
shrinking larger coefficients in the same way as ridge

Hube...
Huber-Berhu partial least squares regression

regression. In\(^{34}\), the authors showed that the combination of the Huber loss function and an adaptive Berhu penalty enjoys oracle properties, and they also demonstrated that this procedure encourages a grouping effect. In previous research, the authors solved a Huber-Berhu optimization problem using CVX software\(^{33-35}\), a Matlab-based modeling system for convex optimization. CVX turns Matlab into a modeling language, allowing constraints and objectives to be specified using standard Matlab expression syntax. However, since CVX is slow for large datasets, a proximal gradient descent algorithm was developed for the Huber-Berhu regression in this study, which runs much faster than CVX.

Reconstruction of GRNs often involves ill-posed problems due to high dimensionality and multicolinearity. Partial least squares (PLS) regression has been an alternative to ordinary regression for handling multicolinearity in several areas of scientific research. PLS couples a dimension reduction technique and a regression model. Although PLS has been shown to have good predictive performance in dealing with ill-posed problems, it is not particularly tailored for variable selection. Sæbø et al. 2007\(^{36}\) first proposed the soft-threshold-PLS (ST-PLS), in which the \(\ell_1\) penalty is used for PLS loading weights of multiple latent components. Such a method is especially applicable for classification and variable selection when the number of variables is greater than the number of samples. Chun and Keleg 2010\(^{37}\) proposed a similar sparse PLS regression for simultaneous dimension reduction and variable selection. Both the methods from Sæbø et al. 2007 and Chun and Keleg 2010 used the same \(\ell_1\) penalty for PLS loading weights. Lê Cao et al. 2008\(^{38}\) also proposed a sparse PLS method for variable selection when integrating omics data. They added sparsity into PLS with a LASSO penalization combined with singular value decomposition (SVD) computation. In this study, the Huber loss function and the Berhu penalty function were embedded into a PLS framework. Real gene data was used to demonstrate that this approach is applicable for the reconstruction of GRNs.

MATERIALS AND METHODS

High-throughput gene expression data

The lignin pathway analysis used an Arabidopsis wood formation compendium dataset containing 128 Affymetrix microarrays pooled from six experiments (accession identifiers: GSE607, GSE6153, GSE18985, GSE2000, GSE24781, and GSE5633) in NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/). These datasets were originally obtained from hypocotyledoned stems under short-day conditions known to induce secondary wood formation\(^{38}\). The original CEL files were downloaded from GEO and preprocessed using the affy package in Bioconductor (https://www.bioconductor.org/) and then normalized with the robust multi-array analysis (RMA) algorithm in affy package. This compendium data set was also used in our previous studies\(^{40}\). The maize B73 compendium data set used for predicting photosynthesis light reaction (PLR) pathway regulators was downloaded from three NCBI databases: (1) the sequence read archive (SRA) (https://www.ncbi.nlm.nih.gov/sra), 39 leaf samples from ERP011838; (2) Gene Expression Omnibus (GEO), 24 leaf samples from GSE61333, and (3) BioProject (https://www.ncbi.nlm.nih.gov/bioproject/), 36 seedling samples from PRJNA483231. This compendium is a subset of that used in our earlier co-expression analysis\(^{41}\). Raw reads were trimmed to remove adaptors and low-quality base pairs via Trimmomatic (v3.3). Clean reads were aligned to the B73Ref3 with STAR, followed by the generation of normalized FPKM (fragments per kb of transcript per million reads) using Cufflinks software (v2.1.1)\(^{42}\).

Huber and Berhu functions

In estimating regression coefficients, the square loss function is well suited if \(y_i\) follows a Gaussian distribution, but it gives a poor performance when \(y_i\) follows a heavy-tailed distribution or there are outliers. On the other hand, the least absolute deviation (LAD) loss function is more robust to outliers, but the statistical efficiency is low when there are no outliers in the data. The Huber function, introduced in\(^{43}\), is a combination of linear and quadratic loss functions. For any given positive real \(M\) (called shape parameter), the Huber function is defined as:

\[
H_M (z) = \begin{cases} 
\frac{z^2}{2M} & |z| \leq M \\
\frac{M^2}{2M} & |z| > M 
\end{cases}
\]

This function is quadratic for small \(z\) values but grows linearly for large values of \(z\). The parameter \(M\) determines where the transition from quadratic to linear takes place (Fig. 1a). In this study, the default value of \(M\) was set to be one tenth of the interquartile range (IRQ), as suggested by\(^{46}\). The Huber function is a smooth function with a derivative function:

\[
H_M'(z) = \begin{cases} 
\frac{z}{M} & |z| \leq M \\
\frac{M}{2M} & |z| > M 
\end{cases}
\]

The ridge regression uses the quadratic penalty on regression coefficients, and it is equivalent to putting a Gaussian prior on the coefficients. LASSO uses a linear penalty on regression coefficients, and this is equivalent to putting a Laplace prior on the coefficients. The advantage of LASSO over ridge regression is that it implements regularization and variable selection simultaneously. The disadvantage is that, if a group of predictors is highly correlated, LASSO picks only one of them and shrinks the others to zero. In this case, the prediction performance of ridge regression dominates the LASSO. The Berhu penalty function, introduced in Owen 2007\(^{33}\), is a hybrid of the quadratic penalty and LASSO. It gives a quadratic penalty to large coefficients while giving a linear penalty to small coefficients, as shown in Fig. 1b. The Berhu function is defined as:

\[
B_M (z) = \begin{cases} 
\frac{|z|^2}{2M} & |z| \leq M \\
\frac{M^2}{2M} & |z| > M 
\end{cases}
\]

The shape parameter \(M\) was set to be the same as that in the Huber function. As shown in Fig. 1b, the Berhu function is a convex function, but it is not differentiable at \(z = 0\). The 2D contours of Huber and Berhu functions are shown in Fig. 1c and Fig. 1d, respectively. When the Huber loss function and the Berhu penalty were combined, an objective function, as referred to as the Huber-Berhu function, was obtained, as shown below.

\[
f(\beta) = \sum_{i=1}^{n} H_M(y_i - \beta_0 - x_i^T \beta) + \lambda \sum_{j=1}^{p} B_M (\beta_j)
\]

The estimation of coefficients using the Huber-Berhu objective (Fig. 2a), LASSO (Fig. 2b), and the ridge (Fig. 2c) regressions provided some insights. The Huber loss corres-
ponds to the rotated, rounded rectangle contour in the top right corner, and the center of the contour is the solution of the un-penalized Huber regression. The shaded area is a map of the Berhu constraint where a smaller $\lambda$ corresponds to a larger area. The estimated coefficient of the Huber-Berhu regression is the first place the contours touch the shaded area; when $\lambda$ is small, the touch point is not on the axes, which means the Huber-Berhu regression behaves more like the ridge regression, which does not generate a sparse solution. When $\lambda$ increases, the correspondent shaded area changes to a diamond, and the touch point is more likely to be located on the axes. Therefore, for large $\lambda$, the Huber-Berhu regression behaves like LASSO, which can generate a sparse solution.

**The algorithm to solve the Huber-Berhu regression**

Since the Berhu function is not differentiable at $z = 0$, it is difficult to use the gradient descent method to solve equation (4). Although we can use the general convex optimization solver CVX[35] for a convex optimization problem, it is too slow for real biological applications. Therefore, a proximal gradient descent algorithm was developed to solve equation (4). Proximal gradient descent is an effective algorithm to solve an optimization problem with decomposable objective function. Suppose the objective function can be decomposed as $f(z) = g(z) + h(z)$, where $g(z)$ is a convex differentiable function and $h(z)$ is a convex non-differentiable function. The idea behind the proximal gradient descent[45] method is to make a quadratic approximation to $g(z)$ and leave $h(z)$ unchanged. That is:

$$f(z) = g(z) + h(z) = g(z) + \nabla g(x)^T (z - x) + \frac{1}{2\lambda} \|z - x\|^2 + h(z)$$

At each step, $x$ is updated by the minimum of the right side of above formula.

$$x^* = \arg\min_x g(x) + \nabla g(x)^T (z - x) + \frac{1}{2\lambda} \|z - x\|^2 + h(z)$$

$$= \arg\min_x \frac{1}{2\lambda} \|z - (x - \lambda \nabla g(x))\|^2 + h(z)$$

The operator $\text{Prox}_{\lambda h}(x) = \arg\min \frac{1}{2\lambda} \|z - x\|^2 + h(z)$ is called

Fig. 1 Huber loss function (a) and Berhu penalty function (b); The 2D contours of Huber loss function (c) and Berhu penalty function (d).

Fig. 2 Estimation picture for the Huber-Berhu regression (a) when least absolute shrinkage and selection operator (LASSO) (b) and ridge (c) regressions are used as a comparison.
Huber-Berhu partial least squares regression

proportional mapping for $h$. To solve (7), the key is to compute the proximal mapping for the Berhu function:

$$AB_h(z) = \frac{1}{4} \left( [z]_M + 2 \frac{z^2 + M^2}{2M} 1_{|z| > M} - [z]_M + 2 \frac{([z] - M)^2}{2M} 1_{|z| > M} \right)$$

Let $u(z) = \frac{1}{4} \left( [z]_M + 2 \frac{z^2 + M^2}{2M} 1_{|z| > M} - [z]_M + 2 \frac{([z] - M)^2}{2M} 1_{|z| > M} \right)$ satisfies theorem 4 in [48],

$$\text{Prox}_{x,\beta u} (x) = \text{Prox}_{x,\beta u} (x) \circ \text{Prox}_{x,\beta l} (x)$$

It is not difficult to verify:

$$\text{Prox}_{x,\beta u} (x) = \text{sign}(x) \min \left( \frac{1}{M + t} |x| + t \lambda, 0 \right)$$

Finding $\beta_0$ and $\beta$ that minimize $f(\beta)$ in (7) is detailed in Algorithm 1.

Algorithm 1 uses the accelerated proximal gradient descent method to solve (7). Line 3 implements the acceleration of [47]. Lines 6–7 compute the proximal mapping of the Berhu function. Lines 5–10 use a backtracking method to determine the step size.

**Embedding the Huber-Berhu objective function into PLS**

Let $X(n \times p)$ and $Y(n \times q)$ be the standardized predictor variables (gene expression of TF genes) and dependent variables (gene expression of pathway genes), respectively. PLS [48] looks for a linear combination of $X$ and a linear combination of $Y$ such that their covariance reaches a maximum:

$$\max_{l_0 \leq 1, l_1 \leq 1} \text{cov}(Xu, Yv)$$

Here, the linear combination $\xi = Xu$ and $\eta = Yv$ are called component scores (or latent variables) which are generated through the $p$ and $q$ dimensional weight vectors $u$ and $v$.

**Algorithm 1:** Accelerated proximal gradient descent method to minimize $f(\beta)$ in equation (7) respected to $\beta_0$ and $\beta$

| Step | Description |
|------|-------------|
| 1    | Output: regression coefficient ($\beta$) |
| 2    | Input: predictor matrix ($X$), dependent vector ($y$), and penalty constant ($\lambda$) |
| 3    | Compute the gradient of Huber loss at $v$ using (5), denoted as $G_v$ |
| 4    | Compute the gradient of Huber loss at $u$ using (6), denoted as $G_u$ |
| 5    | Update $\beta_0$ and $\beta$ |

Equations (14) and (15) can be solved using Algorithm 1. Therefore (13) can be solved iteratively by updating $u$ and $v$ alternately. Note, it is not cost-efficient to spend a lot of effort optimizing over $u$ in line 6 before a good estimate for $v$ is computed. Since Algorithm 1 is an iterative algorithm, it may make sense to stop the optimization over $u$ early before updating $v$. In the implementation, one step of proximal mapping was used to update $u$ and $v$. That is:

$$u = \text{Prox}_{x,\beta u} \left( x - t \frac{\partial H(M - uv')}{\partial u} \right)$$

$$v = \text{Prox}_{x,\beta v} \left( v - t \frac{\partial H(M - uv')}{\partial v} \right)$$

The algorithm for finding the solution of the Huber-Berhu PLS regression in (13) is detailed in Algorithm 2.

**Tuning criteria and choice of the PLS dimension**

The Huber-Berhu PLS regression has two tuning parameters, namely, the penalization parameter $\lambda$ and the number of hidden components $K$. To select the best penalization parameter, $\lambda$, a common $k$-fold cross-validation (CV) procedure that
minimizes the overall prediction error is applied using a grid of possible values. If the sample size is too small, CV can be replaced by leave-one-out validation; this procedure is also used in for tuning penalization parameters\cite{37, 49}.

To choose the dimension of PLS, the $Q^2_h$ criteria were adopted. $Q^2_h$ criteria were first proposed by Tenenhaus\cite{50}. These criteria characterize the predictive power of the PLS model by performing cross-validation computation. $Q^2_h$ is defined as:

$$Q^2_h = 1 - \frac{\sum_{i=1}^{n} PRESS_{h}^{k}}{\sum_{i=1}^{n} RS_{h}^{k}}$$

where $PRESS_{h}^{k} = \sum_{i=1}^{n} (y_{ki} - \hat{y}_{ki}^{(h)})^2$ is the Prediction Error Sum of Squares, and $RS_{h}^{k} = \sum_{i=1}^{n} (y_{ki}^{(h)} - \bar{y}_{k})^2$ is the Residual Sum of Squares for the variable $k$ and the PLS dimension $h$. The criterion for determining if $h$ contributes significantly to the prediction is:

$$Q^2_h \geq (1 - 0.952) = 0.0975$$

This criterion is also used in SIMCA-P software\cite{51} and sparse PLS\cite{38}. However, the choice of the PLS dimension still remains an open question. Empirically, there is little biological meaning when $h$ is large and good performance appears in 2--5 dimensions.

\section*{RESULTS}

\subsection*{The efficiency of the proximal gradient descent algorithm}

We developed the proximal gradient descent algorithm (Algorithm 1) to solve Huber-Berhu regression. As compared to CVX, it could reduce the running time to at least 10 times, but up to 90 times in a desktop computer with 2.2 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 memory for a setting of $m$ and $p$ based on 30 replications. For different $m$, the patterns are similar (Fig. 3). More details can be found in the Deng 2018\cite{52}.

\subsection*{Validation of Huber-Berhu PLS with lignin biosynthesis pathway genes and regulators}

The HB-PLS algorithm was examined for its accuracy in identifying lignin pathway regulators using the Arabidopsis thaliana microarray compendium data set produced from stem tissues\cite{40}. TFs identified by HB-PLS were compared to those identified by SPLS. The 50 top TFs that were ranked based on their connectivities with the lignin biosynthesis pathway genes were identified using HB-PLS (Fig. 4a) and compared to those identified by SPLS (Fig. 4b), respectively. The lignin biosynthesis pathway genes are shown in Fig. 4c. The positive lignin biosynthesis pathway regulators, which are supported by literature evidence, are shown in coral color. The HB-PLS algorithm identified 15 known lignin pathway regulators. Of these, MYB63, SND3, MYB46, MYB85, LBD15, SND1, SND2, MYB103, MYB58, MYB43, NST2, GATA12, VND4, NST1, MYB52, are positive known transcriptional activators of lignin biosynthesis in the SND1-mediated transcriptional regulatory network\cite{53}, and LBD15\cite{54} and GATA12\cite{55} are also involved in regulating various aspects of secondary cell wall synthesis. Interestingly, SPLS identified the same set of positive pathway regulators as HB-PLS though their ranking orders are different.

\subsection*{Prediction of photosynthetic pathway regulators in Arabidopsis thaliana using Huber-Berhu PLS}

Photosynthesis is mediated by the coordinated action of
approximately 3,000 different proteins, commonly referred to as photosynthesis proteins. In this study, we used genes from the photosynthesis light reaction pathway and Calvin cycle pathway to study which regulatory genes can potentially control photosynthesis. Analysis was performed using HB-PLS, with SPLS as a comparative method. The compendium data set we used is comprised of 238 RNA-seq data sets from Arabidopsis thaliana leaves that were under normal/untreated conditions. Expression data for 1389 TFs and 130 pathway genes were extracted from the above compendium data set and used for analyses. The results of HB-PLS and SPLS methods are shown in Fig. 5a and 5b, respectively, where 33 rather than 50 TFs were shown because the SPLS method only identified 33 TFs. Of the top 33 candidate TFs in the lists, HB-PLS identified 11 positive known TFs while SPLS identified 6 positive known TFs. IAA7, also known as AXR2, is regulated by HY5, which binds G-box in LIGHT-HARVESTING CHLOROPHYLL A/B proteins. STO, also known as BBX24, whose protein physically interacts with photosynthesis regulator HY5 to control photomorphogenesis; PHYTOCHROME-INTERACTING FACTOR (PIF) family have been shown to affect the expression of photosynthesis-related genes, including genes encoding LHCA, LHCB, and PsaD proteins. PIFs repress chloroplast development and photomorphogenesis; PIF7, together with PIF3 and PIF4, regulates responses to prolonged red light by modulating phyB levels. PIF7 is also involved in the regulation of circadian rhythms. GLK2 directly regulate the expression of a series of photosynthetic genes including the genes encoding the PSI-LHCl complex and PSII-LHCII complex. The plastid sigma-like transcription factor SIG1 regulate psaA respectively; TOC1 is a member of the PRR (PSEUDO-RESPONSE REGULATOR) family that includes PRR9, PRR7, PRR5, PRR3, and PRR1/TOC1. HY5 also binds and regulates the circadian clock gene PRR7, which affects the operating efficiency of PSII under blue light. GATA transcription factors have implicated some proteins in light-mediated and circadian-regulated gene expression, GATAs can bind to XXIII box, a cis-acting elements involved in light-
regulated expression of the nuclear gene GAPB, which encodes the B subunit of chloroplast glyceraldehyde-3-phosphate dehydrogenase in *A. thaliana* [70]. In addition, GATA interacts with SORLIP motifs in the 3-hydroxy-3-methylglutaryl-CoA reductase (*HMGR*) promoter of *Picrorhiza kurrooa*, a herb plant, for the control of light-mediated expression; upstream sequences of *HMGR* of *P. kurrooa* (*PropkHMGR*)-mediated gene expression was higher in the dark as compared to that in the light in *A. thaliana* across four temperatures studied [71]. GATA phytochrome interacting factor transcription factors regulate light-induced vindoline biosynthesis in *Catharanthus roseus* [72].

A number of genes show greater than 2-fold higher expression in light-grown than dark-grown seedlings with the greatest differences observed for *GATA6*, *GATA7*, *GATA21*-23 [68], with *GATA6* and 7 showing about 6- and 4-fold difference in expression levels. *GATA11* is found to be a hub regulator of photosynthesis and Chlorophyll biosynthesis [73]. The GLK transcription factors promote the expression of many nuclear-encoded photosynthetic genes that are associated with chlorophyll biosynthesis and light-harvesting functions [74]; HSFA1, a master regulator of transcriptional regulation under heat stress, regulates photosynthesis by inducing the expression of downstream transcription factors [75]. *BEH1* is a homolog of *BZR1*, genetic analysis indicates that the *BZR1*-PIF4 interaction controls a core transcription network by integrating brassinosteroids and light response [76].

### The performance and sensitivity of HB-PLS using SPLS as a comparison

We tested the HB-PLS method in comparison with SPLS using two metabolic pathways, lignin biosynthesis pathway and a unified photosynthesis pathway whose regulatory genes are largely and partially known, respectively. We found that HB-PLS could identify more positive known TFs that are supported by existing literature in the output lists. To examine which methods can rank relatively more positive known TFs to the top of output regulatory gene lists, we plotted receiver operating characteristic curves (ROC) and calculated the area under the ROC curve (AuROC), which reflects the sensitivity versus 1-specificity of a method. The results are shown in Fig. 6. For lignin biosynthesis pathway, HB-PLS was capable of ranking more positive known pathway regulators to the top in the inferred regulatory gene list. As a result, the AuROC of HB-PLS (0.94) (Fig. 6a) is much larger than that of SPLS (0.73) (Fig. 6b). For the unified light reaction and Calvin cycle pathway, the true pathway regulators have not been fully identified, and they are only partially known. Although SPLS only identified the 6 positive known pathway regulators in comparison with 10 identified by HB-PLS, SPLS ranked 4 out 6 positive known pathway regulators to the top 8 positions, resulting in slightly higher sensitivity versus 1-specificity. HB-PLS identified 10 positive known regulators among the top 33 regulatory genes, which are more evenly distributed in the list, resulting in relatively smaller AuROC (0.49) as compared to the AuROC of SPLS (0.64). The overall lower AuROC values for both methods for photosynthesis pathway are probably owing to the low number of positive known regulatory genes for this pathway.

Given the fact that lignin biosynthesis pathway regulators have been well identified and characterized experimentally [77], they are specifically suited for examining the efficiency of the HB-PLS method for each pathway gene. We selected two methods, SPLS and PLS, as comparisons. For each output TF list to a pathway gene yielded from one of three methods, we applied a series of cutoffs, with the number of TFs retained varying from 1 to 40 in a shifting step of 1 at a time, and then counted the number of positive regulatory genes in each of the retained lists. The results are shown in Supplementary Fig. S1. It is obvious that for almost every pathway gene, HB-PLS has higher sensitivity versus specificity.

The results indicate that the HB-PLS and SPLS regressions, in

---

**Fig. 5** The implementation of Huber-Berhu-Partial Least Squares (HB-PLS) to identify candidate regulatory genes (purple and coral nodes) controlling photosynthesis and related pathway genes. (a) was compared with the sparse partial least squares (SPLS) method (b) in identifying regulators that affects maize photosynthesis light reaction and Calvin cycle pathway genes. The green and yellow nodes within the cycles represent photosynthesis light reaction pathway genes and Calvin cycle pathway genes, respectively. Coral nodes in the circles represent positive predicted biological process or pathway regulators that are supported by existing literature, and shallow purple nodes contain other predicted TFs that do not have experimentally validated supporting evidence at present.
many cases, are much more efficient in recognizing positive regulators to a pathway gene compared to the PLS regression (Supplementary Fig. S1). For most pathway genes like PAL1, C4H, CCR1, C3H, and COMT1, HB-PLS method could identify more positive regulators in the top 20 regulators as compared to the SPLS method. For HCT, CCoAOMT1, CAD8, and FSH, HB-PLS was almost always more efficient than SPLS when the top cut-off lists contained fewer than 40 regulators. For pathway gene CAD8, both SPLS and PLS both failed to identify positive regulators while HB-PLS performed more efficiently.

DISCUSSION

The identification of gene regulatory relationships through constructing GRNs from high-throughput expression data sets has some inherent challenges due to high dimensionality and multicollinearity. High dimensionality is caused by a multitude of gene variables while multicollinearity largely results from a large number of genes versus a relatively small sample size. In this study, we combined three types of computational approaches, statistics (PLS), machine learning (Semi-unsupervised learning) and convex optimization (Berhu and Huber) for simulating gene regulatory relationships, as illustrated in Fig. 7, and our results showed this integrative approach is viable and efficient.

One method that we frequently use to deal with dimensionality and multicollinearity is partial least squares (PLS), which couples dimension reduction with a regression model. However, because PLS is not particularly suited for variable/feature selection, it often produces linear combinations of the original predictors that are hard to interpret due to high dimensionality[78]. To solve this problem, Chun and Keles developed an efficient implementation of sparse PLS, referred to as the SPLS method, based on the least angle regression[79]. SPLS was then benchmarked by means of comparisons to well-known variable selection and dimension reduction approaches via simulation experiments[78]. We used the SPLS method in our previous study[61] and found that it was highly efficient in identifying pathway regulators and thus used it as a benchmark for evaluating the new methods.

In this study, we developed a PLS regression that incorporates the Huber loss function and the Berhu penalty for identification of pathway regulators using high-throughput gene expression data (with dimensionality and multicollinearity). Although the Huber loss function and the Berhu penalty have been proposed in regularized regression models[43,80], this is the first time that both of them were combined with the PLS regression at the same time. The Huber function is a combination of linear and quadratic loss functions. In comparison with other loss functions (e.g., square loss and least absolute deviation loss), Huber loss is more robust to outliers and has higher statistical efficiency than the LAD loss function in the absence of outliers. The Berhu function[33] is a hybrid of the $\ell_2$ penalty and the $\ell_1$ penalty. It gives a quadratic penalty to large coefficients and a linear penalty to small coefficients. Therefore, the Berhu penalty has advantages of both the $\ell_2$ and $\ell_1$ penalties: smaller coefficients tend to shrink to zero while the coefficients of a group of highly correlated predictive variables are not changed much if their coefficients are large.

A comparison of HB-PLS with SPLS and also PLS suggests that HB-PLS can identify more true pathway regulators. This is an advantage over either SPLS or PLS (Supplementary Fig. S1) when experimental validation is concerned. The application of HB-PLS and SPLS methods to identification of lignin biosynthesis pathway regulators in A. thaliana led to the identification of 15 and 15 positive pathway regulators, respectively, while application of the HB-PLS and SPLS methods to identification of photosynthesis pathway regulators in A. thalian resulted in 10 and 6 positive pathway regulators, respectively. The outperformance of HB-PLS over SPLS (Fig. 6a) and PLS (Supplementary Fig. S1) implicates that the use of Huber loss function and Berhu penalty function for convex optimization contributed to the recognition of true pathway regulators and rank them at the top of the output lists. It also suggests the viability and the increased power of combination of statistics (PLS), machine learning (Semi-unsupervised learning) and convex optimization (Berhu and Huber) for recognition of regulatory relationships. In addition, the ROC plotting suggests that HB-PLS method has comparable sensitivity versus 1-specificity compared to SPLS because HB-PLS achieved a higher AuROC for lignin biosynthesis pathway but a lower AuROC for the unified photosynthesis pathway as
compared to SPLS (Fig. 6). However, the fact that the HB-PLS identified the same or higher number of positive true regulators than SPLS for the two pathways we analyzed, and the sensitivity of HB-PLS is much better than that of SPLS for lignin pathway whose regulatory genes are more complete, and slightly worse than that of HB-PLS for photosynthesis light reaction and Calvin cycle pathway (Fig. 5 and Supplementary Fig. S1) whose regulatory genes are only partially known. Therefore, HB-PLS has an overall larger advantage. Unfortunately, except the two pathways we evaluated, there are almost no other metabolic pathways whose regulatory genes have been mostly identified. Our analysis showed that the two methods could empower the recognition of pathway regulators including some unique pathway regulators, and thus are useful in continued research.

**CONCLUSIONS**

A new method called the HB-PLS regression was developed for identifying biological process or pathway regulators by integration of statistics, machine learning and convex optimization approaches. In HB-PLS, an accelerated proximal gradient descent algorithm was specifically developed to solve Huber and Berhu optimization, which can estimate the regression parameters by optimizing the objective function based on the Huber and Berhu functions. Characteristic analysis of the Huber-Berhu regression indicated it could identify sparse solution. When modeling the gene regulatory relationships from regulatory genes to pathway genes, HB-PLS is capable of dealing with the high multicollinearity of both regulatory genes and pathway genes. Application of the HB-PLS to real A. thaliana high-throughput data showed that HB-PLS could identify majority positive known regulatory genes that govern two pathways. Sensitivity verse 1-specificity plotting showed that HB-PLS could rank more positive known regulators to the top of output regulatory gene lists for lignin biosynthesis pathways while SPLS can rank more for the unified photosynthesis pathway. Our study suggests that the overall performance of HB-PLS exceeds that of SPLS but both methods may have comparable sensitivity/specificity and are instrumental for identifying real biological process and pathway regulators from high-throughput gene expression data.

**ACKNOWLEDGMENTS**

NSF Plant Genome Program [1703007 to SL and HW]; NSF Advances in Biological Informatics [dbi-1458130 to HW]; USDA McIntire-Stennis Fund to HW.

**Availability of R Package:** The R code and sample data for HB-PLS is available at github https://github.com/hwei0805/HB-PLS. For the R code of SPLS, please write to Dr. Wei (hairong@mtu.edu) to request.

**Conflict of interest**

The authors declare that they have no conflict of interest.

**Supplementary Information** accompanies this paper at (http://www.maxapress.com/article/doi/10.48130/FR-2021-0006)

**Dates**

Received 11 January 2021; Accepted 11 March 2021; Published online 30 March 2021

**REFERENCES**

1. Chen T, He HL, Church GM. 1999. Modeling gene expression with differential equations. In *Proceeding of the Pacific Symposium on*
Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U. S. A.*, 97:12182–6

19. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5:e8

20. Meyer PE, Kontos K, Lafitte F, Bontempi G. 2007. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology* 2007:98787 https://rdcu.be/chDK7

21. Gunasekara C, Zhang K, Deng W, Brown L, Wei H. 2018. TGMi: an efficient algorithm for identifying pathway regulators through evaluation of triple-gene mutual interaction. *Nucleic Acids Res.* 46:e67

22. Zhang X, Zhao X, He K, Lu L, Cao Y, et al. 2012. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 28:98–104

23. Meinhausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the Lasso. *Annals of statistics* 34:1436–62

24. Zhang X, Liu K, Liu Z, Duval B, Richer JM, et al. 2013. NARROMi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* 29:106–13

25. Hayes AF, Cai L. 2007. Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation. *Behav. Res. Methods* 39:79–22

26. Hoerl AE, Kennard RW. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67

27. Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58:267–88

28. Zou H. 2006. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101:1418–29

29. Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67:301–20

30. Wang H, Li G, Jiang G. 2007. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics* 25:347–55

31. Yu C, Yao W. 2017. Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation* 46:6261–82

32. Lambert-Lacroix S, Zwald L. 2011. Robust regression through the Huber’s criterion and adaptive lasso penalty. *Electronic Journal of Statistics* 5:1015–53

33. Owen AB. 2007. A robust hybrid of lasso and ridge regression. *Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference on Machine and Statistical Learning: Prediction and Discovery, Snowbird, UT*, 2006, Contemporary Mathematics 443:39–72 Providence RI: American Mathematical Society http://www.ams.org/books/contm/443/

34. Zwald L, Lambert-Lacroix S. 2012. The BerHu penalty and the grouped effect. arXiv preprint arXiv:1207.8668

35. Grant M, Boyd S, Ye Y. 2008. CVX: Matlab software for disciplined convex programming. http://cvxr.com/cvx/

36. Sarbse A, Almey A, Aaree J, Aastveit AH. 2007. ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. *Chemometrics and Intelligent Laboratory Systems* 22:54–62

37. Chun H, Keleş S. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72:3–25

38. Lé Cao K-A, Rossouw D, Robert-Granié C, Besse P. 2008. A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology* 7:Artic135

Deng et al. *Forestry Research* 2021, 1:6
1:58−67
physically interact to regulate gene expression in
*Plant Physiol.* 176:2963−76

216:32−45
New Phytol.

photosynthetic capacity: conservation and divergence from
*Physiol.* 129:747−61

47:1015−34
Semismooth newton coordinate descent

1248−51
Springer. pp. 101, ed. Lovric M. Berlin, Heidelberg: Springer. pp. 1248−51 https://doi.org/10.1007/978-3-642-04898-2_594

41:91−9.
New York: Curran Associates Inc. https://proceedings.
Proceedings of the 26th International Conference on Neural
NIPS’13:

60:106:7660−5
Phytochromes promote seedling light responses by inhibiting four
negatively-acting phytochrome-interacting factors.
*Proc. Natl. Acad. Sci. U. S. A.* 106:7660−5

60:177−9
Physiol.

77:251−62
IEEE/ACM Trans Comput Biol Bioinform

7:251−62
SIMCA-P. 2002. SIMCA-P+ 10 Manual. Umetrics AB

20:2763−82
*Plant Cell*

14:802−9
*cell biology*

38:332−47
transcription factor HY5 integrates light and hormone signaling

58:177−9
*Physiol.*

39: Chaffey N, Cholewa E, Regan S, Sundberg B. 2002. Secondary
xylem development in Arabidopsis: a model for wood formation.
*Physiolgia plantarum* 114:594−600.

40: Kumari S, Deng W, Gunasekara C, Chiang V, Chen HS, et al. 2016.
Bottom-up GGM algorithm for constructing multilayered
hierarchical gene regulatory networks that govern biological
pathways or processes. *BMC Bioinformatics* 17:132

41: Zheng J, He C, Qin Y, Lin G, Park WD, et al. 2019. Co-expression
analysis aids in the identification of genes in the cuticular wax
pathway in maize. *Plant J.* 97:530−42

42: Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. 2010.
Transcript assembly and quantification by RNA-Seq reveals
unannotated transcripts and isoform switching during cell
differentiation. *Nat. Biotechnol.* 28:511−5

43: Huber PJ. 2011. Robust statistics. In *International Encyclopedia of
Statistical Science*, ed. Lovric M. Berlin, Heidelberg: Springer. pp. 1248−51 https://doi.org/10.1007/978-3-642-04898-2_594

44: Yi C, Huang J. 2017. Semismooth Newton coordinate descent
algorithm for elastic-net penalized Huber loss regression and
quantile regression. *Journal of Computational and Graphical
Statistics* 26:547−57

45: Parikh N, Boyd S. 2014. Proximal algorithms. *Foundations and
Trends® in Optimization* 1:127−239

46: Yu YL. 2013. On decomposing the proximal map. NIPS’13:
Proceedings of the 26th International Conference on Neural
Information Processing Systems, Lake Tahoe, Nevada, 2013, vol
1:91−9. New York: Curran Associates Inc. https://proceedings.
neurips.cc/paper/2013/file/98dec83da57b0395e163467c9da5e521
b-Paper.pdf

47: Beck, A. and M. Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2(1):183−202

48: Vinzi VE, Russolillo G. 2013. Partial least squares algorithms and
methods. *WIREs Computational Statistics* 5:1−19

49: Shen H, Huang JZ. 2008. Sparse principal component analysis via
regularized low rank matrix approximation.

50: Shen H, Huang J. 2017. Semismooth newton coordinate descent
algorithm for elastic-net penalized Huber loss regression and
quantile regression. *Journal of Computational and Graphical
Statistics* 26:547−57

51: Tenenhaus A, Guillemot V, Gidrol X, Frouin V. 2010. Gene
association networks from microarray data using a regularized
estimation of partial correlation based on PLS regression.
*IEEE/ACM Trans Comput Biol Inform Bioinf* 7:251−62

52: Simca P. 2002. SIMCA-P+ 10 Manual. Umetrics AB

53: Deng W. 2018. Algorithms for reconstruction of gene regulatory
networks from high-throughput gene expression data. PhD. Open
Access Dissertation. Michigan Technological University. pp. 101 https://digitalcommons.mtu.edu/etdr/722/

54: Zhou J, Lee C, Zhong R, Ye ZH. 2009. MYB58 and MYB63 are
transcriptional activators of the lignin biosynthetic pathway

55: Job N, Yadukrishnan P, Bursch K, Datta S, Johansson H. 2018. Two
B-Box Proteins Regulate Photomorphogenesis by Oppositely
Modulating HY5 through their Diverse C-Terminal Domains. *Plant
Physiol.* 176:2963−76.

56: Jiang Y, Yang C, Huang S, Xie F, Xu Y, et al. 2019. The ELF3-PIF7
Interaction Mediates the Circadian Gating of the Shade Response
in Arabidopsis. *Science* 22:288−98

57: Kim K, Jeong J, Kim J, Lee N, Kim ME, et al. 2016. PIF1 Regulates
Plastid Development by Repressing Photosynthetic Genes in the
Endodermis. *Molecular plant* 9:1415−27

58: Shin J, Kim K, Kang H, Zulfugarov IS, Bae G, et al. 2009.
Phytochromes promote seedling light responses by inhibiting four
negatively-acting phytochrome-interacting factors.
*Proc. Natl. Acad. Sci. U. S. A.* 106:7660−5

59: Leivar P, Monte E, Al-Sady B, Carle C, Storer A, et al. 2008. The
Arabidopsis phytochrome-interacting factor PIF7, together with
PIF3 and PIF4, regulates responses to prolonged red light
by modulating phyB levels. *Plant Cell* 20:337−52

60: Waters MT, Wang P, Korkaric M, Capper RG, Saunders NJ, Langdale
JA. 2009. GLK transcription factor coordinate expression of
the photosynthetic apparatus in *Arabidopsis.* The *Plant Cell* 21:1109−28

61: Zubo YO, Blakley IC, Franco-Zorrilla JM, Yamburenko MV, Solano R,
et al. 2018. Coordination of Chloroplast Development through the
Action of the GNC and GLK Transcription Factor Families. *Plant
physiology* 178:130−47

62: Privat I, Hakimi MA, Buhot L, Favory JY, Mache-Lerbs S. 2003.
Characterization of Arabidopsis plastid sigma-like transcription factors SIG1, SIG2 and SIG3. *Plant Mol. Biol.* 51:385−99

63: Littauer S, Battle MW, Lawson T, Jones MA. 2015. Phototropins
maintain robust circadian oscillation of PSII operating efficiency
under blue light. *Plant J.* 83:1034−45

64: Manfield IW, Devlin PF, Jen CH, Westerdraad DR, Gilmartin PM. 2007.
Conservation, convergence, and divergence of light-responsive,
circadian-regulated, and tissue-specific expression patterns during
evolution of the Arabidopsis GATA gene family. *Plant Physiol*
143:941−58

65: Zhang Z, Ren C, Zou L, Wang Y, Li S, et al. 2018. Characterization of
the GATA gene family in *Vitis vinifera*: genome-wide analysis,
expression profiles, and involvement in light and phytohormone
response. *Genome* 61:713−23

66: Jeong MJ, Shih MC. 2003. Interaction of a GATA factor with cis-
element factors involved in light regulation of nuclear genes
encoding chloroplast glycerolaldehyde-3-phosphate
dehydrogenase in *Arabidopsis.* *Biochem. Biophys. Res. Commun.*
300:555−62

67: Kawoosa T, Gahlan P, Devi AS, Kumar S. 2014. The GATA and
SORKIP motifs in the 3-hydroxy-3-methylglutaryl-CoA reductase
promoter of *Picroirhiza kurrooa* for the control of light-mediated
expression. *Funct. Integr. Genomics* 14:191−203

68: Liu Y, Patra B, Pattanaik S, Wang Y, Yuan L. 2019. GATA and
Phytochrome Interacting Factor Transcription Factors Regulate
Light-Induced Vindoline Biosynthesis in *Catharanthus roseus.*
*Plant Physiol.* 180:1336−50

69: Gargouri M, Park JJ, Holguin FO, Kim MJ, Wang H, et al. 2015.
Identification of regulatory network hubs that control lipid
metabolism in *Chlamydomonas reinhardtii.* *J. Exp. Bot.* 66:4551−66

70: Waters MT, Langdale JA. 2009. The making of a chloroplast. *EMBO
J.* 28:2861−73

71: Yoshida T, Ohama N, Nakajima J, Kidokoro S, Mizoi J, et al. 2011.
*Arabidopsis* HsfA1 transcription factors function as the main
positive regulators in heat shock-responsive gene expression.
*Mol. Genet. Genomics* 286:321−32

72: Oh E, Zhu JY, Wang ZY. 2012. Interaction between BZR1 and PIF4
integrates brassinosteroid and environmental responses. *Nature
cell biology* 14:802−9

73: Zhong R, Lee C, Zhou J, McCarthy RL, Ye ZH. 2008. A battery of
transcription factors involved in the regulation of secondary cell
wall biosynthesis in Arabidopsis. *Plant Cell* 20:2763−82

74: Deng et al. Forestry Research 2021, 1:6
78. Chun H, Keleş S. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol* 72:3–25
79. Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Annals of Statistics* 32:407–99
80. Xie Y, Liu Y, Valdar W. 2016. Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. *Biometrika* 103:493–511