Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Partial least squares based identification of Duchenne muscular dystrophy specific genes[#]

Hui-bo AN[†1], Hua-cheng ZHENG[2], Li ZHANG[1], Lin MA[1], Zheng-yan LIU[1]

([1]*Department of Pathology, Children's Hospital, Shijiazhuang 050031, China*)

([2]*Department of Neurology, Children's Hospital, Shijiazhuang 050031, China*)

[†]E-mail: huiboan1@sina.com

Received Mar. 3, 2013; Revision accepted Aug. 29, 2013; Crosschecked Oct. 16, 2013

Abstract: Large-scale parallel gene expression analysis has provided a greater ease for investigating the underlying mechanisms of Duchenne muscular dystrophy (DMD). Previous studies typically implemented variance/regression analysis, which would be fundamentally flawed when unaccounted sources of variability in the arrays existed. Here we aim to identify genes that contribute to the pathology of DMD using partial least squares (PLS) based analysis. We carried out PLS-based analysis with two datasets downloaded from the Gene Expression Omnibus (GEO) database to identify genes contributing to the pathology of DMD. Except for the genes related to inflammation, muscle regeneration and extracellular matrix (ECM) modeling, we found some genes with high fold change, which have not been identified by previous studies, such as *SRPX*, *GPNMB*, *SAT1*, and *LYZ*. In addition, downregulation of the fatty acid metabolism pathway was found, which may be related to the progressive muscle wasting process. Our results provide a better understanding for the downstream mechanisms of DMD.

Key words: Partial least squares (PLS), Gene expression profile, Duchenne muscular dystrophy (DMD)
doi:10.1631/jzus.B1300060 **Document code:** A **CLC number:** R746.2; R-3

1 Introduction

Duchenne muscular dystrophy (DMD) is a devastating inherited neuromuscular disorder that affects one in 3600–6000 live male births (Bushby *et al.*, 2010). DMD is caused by mutations or deletions in the X-linked dystrophin gene. The dystrophin protein locates underneath the sarcolemma and assembles with sarcolemmal proteins to form the dystrophin-associated protein complex (DAPC), which links the cytoskeleton to the extracellular matrix (ECM). Disruption of DAPC may break the mechanical linkage crucial for sarcolemmal integrity, leading to calcium leak channel openings, activation of calcium-dependent protease and fiber necrosis (Straub and Campbell, 1997). In addition, DMD is a progressive disease in which DMD patients gradually lose most of their muscle due to increasing rate of fibrosis and fatty tissue infiltration along with the aging process.

Although the responsible gene for DMD has been characterized for more than 20 years (Koenig *et al.*, 1987; Kunkel *et al.*, 1987), a comprehensive understanding of the downstream mechanisms caused by the absence of dystrophin is still lacking. Previous studies on the secondary changes in DMD have revealed possible involvement of calcium homeostasis (Head, 2010), nitric oxide synthase (Altamirano *et al.*, 2012), inflammation (Monici *et al.*, 2003), and mast cell degranulation (Gorospe *et al.*, 1994), suggesting that the pathological process of DMD is highly complicated. Current available large-scale parallel gene expression analysis has provided greater ease for investigating the underlying molecular pathophysiological mechanisms of DMD. Several gene expression profiling studies (Chen *et al.*, 2000; 2005; Haslett

[#] Electronic supplementary materials: The online version of this article (doi:10.1631/jzus.B1300060) contains supplementary materials, which are available to authorized users

et al., 2002; Pescatori *et al.*, 2007; Wong *et al.*, 2009) have been carried out, providing insights into the pathology of DMD. These studies typically implemented standard analysis of variance/regression to identify the differentially expressed genes over two types of samples. However, this analysis procedure becomes fundamentally flawed when there are unaccounted array-specific factors which are not detectable and cannot be removed by any standard normalizing method. For example, some genes are very highly expressed or depressed as a result of certain demographic profiles. Chakraborty and Datta (2012) proposed that partial least squares (PLS) based analysis was robust in selecting disease specific genes with expression profile data, yielding higher sensitivity while maintaining reasonable specificity, false discovery rate (FDR), and false non-discovery rate (FNR) compared with variance/regression analysis. In this study, to identify genes truly differentially expressed between DMD and normal skeletal muscles, we performed a PLS-based analysis using two datasets downloaded from the Gene Expression Omnibus (GEO) database. A multivariate linear model based on PLS was used to describe the relationship between genes expression and DMD status. Meanwhile, age was also considered as a possible independent variable that may contribute to the status of the samples. Our results provide a better understanding on the downstream mechanisms of DMD and will facilitate further research of new adjuvant treatments.

2 Materials and methods

2.1 Simulation studies

A simulation study was performed to compare the performance of the PLS-based method and a commonly used tool, linear models for microarray data (limma), which is implemented in the R statistical software (http://www.r-project.org/).

We envisaged an expression profiling study with 5 000 genes, including 300 differentially expressed genes and 4 700 non-differentially expressed genes, in a 2×15 sample design. The simulation study is divided into two settings: (1) assuming the genes to be independent of each other; (2) assuming dependence within different groups of genes.

The log transformed gene expression values ($\boldsymbol{Y}$) for all genes are generated using a linear model and expressed as

$$\boldsymbol{Y}=\boldsymbol{X\beta}+e, \quad (1)$$

where $\boldsymbol{X}$ denotes the design matrix corresponding to the above linear model and $\boldsymbol{\beta}$ denotes the corresponding vector of regression coefficients. The error term e is assumed to be independently distributed as $N(0, \sigma^2)$, where the choice of σ^2 is: (1) For each gene, when $\beta=0$, σ is the standard deviation from the control group of the real array data, so different genes are identified with different error scales; (2) Define the noise to signal ratio $\eta=\sigma^2/\mathrm{Var}(\boldsymbol{X\beta})$, where σ^2 is the random error variance and $\mathrm{Var}(\boldsymbol{X\beta})$ is the variance of the signal $\boldsymbol{X\beta}$. This quantity measures the relative intensity of the noise coming from the random error and the confounded primary variable signal depicting the expression effect of the genes over the two groups. We consider three different values 0.1, 0.5, and 1.0 for η to incorporate the cases of strong, moderate, and weak primary signal intensity, respectively. From the three choices of noise to signal ratio, we calculated the corresponding values of σ^2 and used them to simulate the values of e in the model, at different $\boldsymbol{\beta}\sim U[-1, 0)$ and $U(0, -1]$ (assuming different genes have different effects). The simulation study was performed 100 times and the average values of the four performance measures were computed.

2.2 Microarray data

Microarray data were downloaded from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) database. Two datasets (GSE6011 and GSE3307), which included 27 DMD patients and 14 healthy controls, were used in subsequent analysis. Dystrophin protein was absent in all patients (Chen *et al.*, 2000; Pescatori *et al.*, 2007). Sample information for all subjects is listed in Table 1. The two datasets were based on the GPL96 platform: [HG-U133A] Affymetrix Human Genome U133A Array.

2.3 Detection of differentially expressed genes

Entire datasets including CEL and simple omnibus format in text (SOFT) formatted family files for all samples were downloaded. The CEL files were generated by Affymetrix DNA microarray image analysis software and they contained the information

Table 1 Characteristics of the samples

GEO accession	Type	ID	Age (year)	Gender
GSM139515	DMD	D01	0.1	Male
GSM139516	DMD	D02	0.2	Male
GSM139517	DMD	D03	0.3	Male
GSM139519	DMD	D04	0.4	Male
GSM139520	DMD	D05	0.4	Male
GSM139521	DMD	D06	0.5	Male
GSM139522	DMD	D07	0.5	Male
GSM139523	DMD	D08	0.6	Male
GSM139524	DMD	D09	0.7	Male
GSM139525	DMD	D10	0.7	Male
GSM139526	DMD	D11	0.9	Male
GSM139527	DMD	D12	1.0	Male
GSM139528	DMD	D13	1.0	Male
GSM139529	DMD	D14	1.2	Male
GSM139530	DMD	D15	1.2	Male
GSM139531	DMD	D16	1.3	Male
GSM139532	DMD	D17	1.7	Male
GSM139533	DMD	D18	1.7	Male
GSM139534	DMD	D19	1.8	Male
GSM121357	DMD	D20	5.0	Male
GSM121361	DMD	D21	8.0	Male
GSM121363	DMD	D22	7.0	Male
GSM121368	DMD	D23	8.0	Male
GSM121369	DMD	D24	7.0	Male
GSM139535	DMD	D25	2.3	Male
GSM139536	DMD	D26	3.9	Male
GSM139537	DMD	D27	5.1	Male
GSM139501	Control	C28	0.4	Male
GSM139502	Control	C29	0.5	Female
GSM139503	Control	C30	0.5	Female
GSM139504	Control	C31	0.5	Male
GSM139505	Control	C32	0.5	Male
GSM139506	Control	C33	0.6	Male
GSM139507	Control	C34	0.7	Female
GSM139508	Control	C35	0.9	Male
GSM139509	Control	C36	1.5	Male
GSM139510	Control	C37	2.8	Male
GSM139511	Control	C38	3.0	Male
GSM139512	Control	C39	4.2	Male
GSM139513	Control	C40	5.0	Male

about each probe on the chip. For raw intensity value normalization, robust multiarray analysis (RMA) (Irizarry *et al.*, 2003) was used as follows: firstly, a model-based background correction was used to neutralize the effects of background noise and the processing artifacts; secondly, quantile normalization was used to align expression values to a common scale; finally, an iterative median polishing procedure was used to generate a single expression value for each probe set. The resulting RMA expression value (log2-transformed) was then used for further analysis. A multivariate linear model was used to describe the relationship between the gene expression and DMD disease status. Age was also introduced as an independent variable to discover the disease-related genes comprehensively. For each sample, the model is expressed as

$$y=\sum_{i=1}^{p}\alpha_i x_i+\beta\cdot\text{age}+b, \tag{2}$$

where y is the binary variable of disease status (0 coded as control and 1 coded as DMD), and p is the total number of genes in the chip array. Obviously, in our dataset, the number of genes (p=22 283) was much greater than the sample number (n=41). PLS (Helland, 1988; 1990), a dimension reduction procedure for modeling without imposing strong assumptions, were then used to estimate the effect for each gene. The main idea of PLS regression was to build orthogonal components, called "latent variables", and it is expressed as

$$\text{COV}(\boldsymbol{t}_k,\boldsymbol{u}_k)\rightarrow\max, \tag{3}$$

$$\text{s.t. }\|\boldsymbol{t}_k\|=1,\ \|\boldsymbol{u}_k\|=1, \tag{4}$$

where $\boldsymbol{t}_k$ is the kth latent variable decomposes from all individuals' gene expression data $\boldsymbol{X}$ (the matrix of $n\times p$, n is number of individuals, each column was normalized), $\boldsymbol{u}_k$ is the kth latent variable decomposed from the target trait data $\boldsymbol{Y}$ ($n\times 1$). The nonlinear iterative partial least squares (NIPALS) algorithm (Martins *et al.*, 2010) was used to calculate the PLS latent variables derived from the expression profile on the target trait, as

(1) Randomly initialize $\boldsymbol{u}_0=\boldsymbol{Y}$;
(2) $\boldsymbol{w}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}_0$, $\boldsymbol{w}=\boldsymbol{w}/\|\boldsymbol{w}\|$;
(3) $\boldsymbol{t}=\boldsymbol{X}\boldsymbol{w}$;
(4) $\boldsymbol{c}=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{t}$, $\boldsymbol{c}=\boldsymbol{c}/\|\boldsymbol{c}\|$;
(5) $\boldsymbol{u}=\boldsymbol{Y}\boldsymbol{c}$;
(6) If $\boldsymbol{u}-\boldsymbol{u}_0<10^{-8}$, then go to Step (7);
else $\boldsymbol{u}_0=\boldsymbol{u}$, repeat Steps (2)–(5);
(7) $\boldsymbol{X}=\boldsymbol{X}-\boldsymbol{t}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{X}$, $\boldsymbol{Y}=\boldsymbol{Y}-\boldsymbol{t}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{Y}$. Then go back to Step (2) to calculate the next latent variable.

To evaluate the importance of the expressed genes on disease, the statistics of variable importance on the projection (VIP) (Gosselin *et al.*, 2010) was calculated by

$$\mathrm{VIP}_j = \sqrt{p\sum_{k=1}^{h}\mathrm{Cor}^2(\boldsymbol{Y},\boldsymbol{t}_k)\boldsymbol{w}_{kj}^2 \Big/ \sum_{k=1}^{h}\mathrm{Cor}^2(\boldsymbol{Y},\boldsymbol{t}_k)}, \quad (5)$$

where Cor( ) operator is the Pearson correlation coefficient, and for each $\boldsymbol{w}_k$, it should be normalized by dividing $\|\boldsymbol{w}_k\|$, and h is the number of latent variables used in the model.

To avoid the model over fitting, the best number of latent variables h was determined by $P<0.05$ based on the logistic regression model with maximum likelihood estimation. The VIP for each gene was then calculated with the significant h latent variables to show genes associated with DMD. Additionally, the false discovered rate (FDR) procedures are used to control the expected proportion of incorrectly rejected null hypotheses. In our study, we used permutation tests to do the FDR control. The permutation procedure (N=10 000 times) was implemented to obtain the empirical distribution of PLS-based VIP in each replicate. The FDR for each gene could be evaluated according to the obtained empirical distribution as

$$\mathrm{FDR}_i = \frac{\sum_j^{10000}\sum_i^{p}\mathrm{Bool}(\mathrm{VIP}_{i,j} > \mathrm{VIP}_i)}{10000p}, \quad (6)$$

where the Bool( ) is the logical value of the expression while "True" is 1 and "False" is 0. The significant genes were selected with a cutoff of FDR<0.05.

2.4 Pathway enrichment analysis

Selected probes were annotated according to the SOFT files. All genes were mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (http://www.genome.jp/kegg/). A hyper geometric distribution test was carried out to identify pathways in which differentially expressed genes are significantly enriched.

3 Results

Results of the simulation study are listed in Table 2. Compared with the limma method, PLS-based analysis achieved much higher sensitivity under all choices of η. The proportion among genes declared significant that were not differentially expressed was slightly increased, which is an inevitable price to pay to achieve a better performance in terms of detection power. Overall, the PLS-based method achieved higher sensitivity while maintaining reasonable specificity and impressively small FDR and FNR.

Table 2 Performance analyses of limma and PLS under the setting of independently expressed genes in two groups

NSR	Method	Sensitivity	Specificity	FDR	FNR
0.1	Limma	0.376 7	0.999 1	0.034 2	0.038 3
	PLS	0.993 3	0.995 3	0.068 8	0.000 4
0.5	Limma	0.173 3	1.000 0	0.000 0	0.050 1
	PLS	0.826 7	0.997 2	0.049 8	0.011 0
1.0	Limma	0.126 7	1.000 0	0.000 0	0.052 8
	PLS	0.473 3	0.998 7	0.040 5	0.032 6

NSR: noise/signal ratio; Sensitivity: proportion among differentially expressed genes that were declared significant; Specificity: proportion among non-differentially expressed genes that were declared non-significant; FDR: false discovery rate, proportion among genes declared significant that were not differentially expressed; FNR: false non-discovery rate, proportion among genes declared non-significant that were differentially expressed

As shown in Table 3, three latent variables with $P<0.05$ were selected. Sample classification according to the three latent variables is illustrated in Fig. 1. Compared with the normal controls, a total of 1 529 probe sets (representing 1 254 genes) were differentially expressed. Among them, 949 probe sets were upregulated while 580 ones were downregulated. Detailed information for these genes is listed in Table S1. Thirty-nine differentially expressed genes were found with the absolute value of fold change (|FC|)>1.5 (Fig. 2).

For the 1 254 differentially expressed genes, 620 genes could be mapped on KEGG pathways. Meanwhile, for all the well-characterized human genes in the Affymetrix Human Genome U133A Array, 5 157 genes could be mapped on KEGG pathways. As shown in Table 4, aberrantly regulated genes were enriched in 58 pathways. Among them, the phagosome pathway was identified to be with the most significant enrichment. According to the classification in the KEGG database, 25 of these pathways are involved in the inflammatory/immune response, such

Table 3 Statistics for the first 10 PLS latent variables estimated with logistic regression

Latent variable	Estimate	Std. Error	*P*-value
V1	23.928	7.868	0.002 36**
V2	10.783	3.528	0.002 24**
V3	5.743	2.714	0.027 3*
V4	3.350	2.274	0.140 7
V5	2.771	2.226	0.213 0
V6	2.591	2.247	0.248 9
V7	0.837	2.177	0.700 7
V8	0.332	2.102	0.874 5
V9	0.176	2.109	0.933 4
V10	0.066	2.111	0.974 9

Only the first three latent variables with $P<0.05$ were selected. * $P<0.05$; ** $P<0.01$

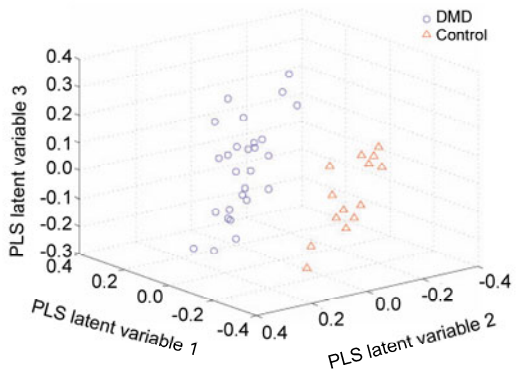

Fig. 1 Sample classification using the selected three partial least squares (PLS) latent variables

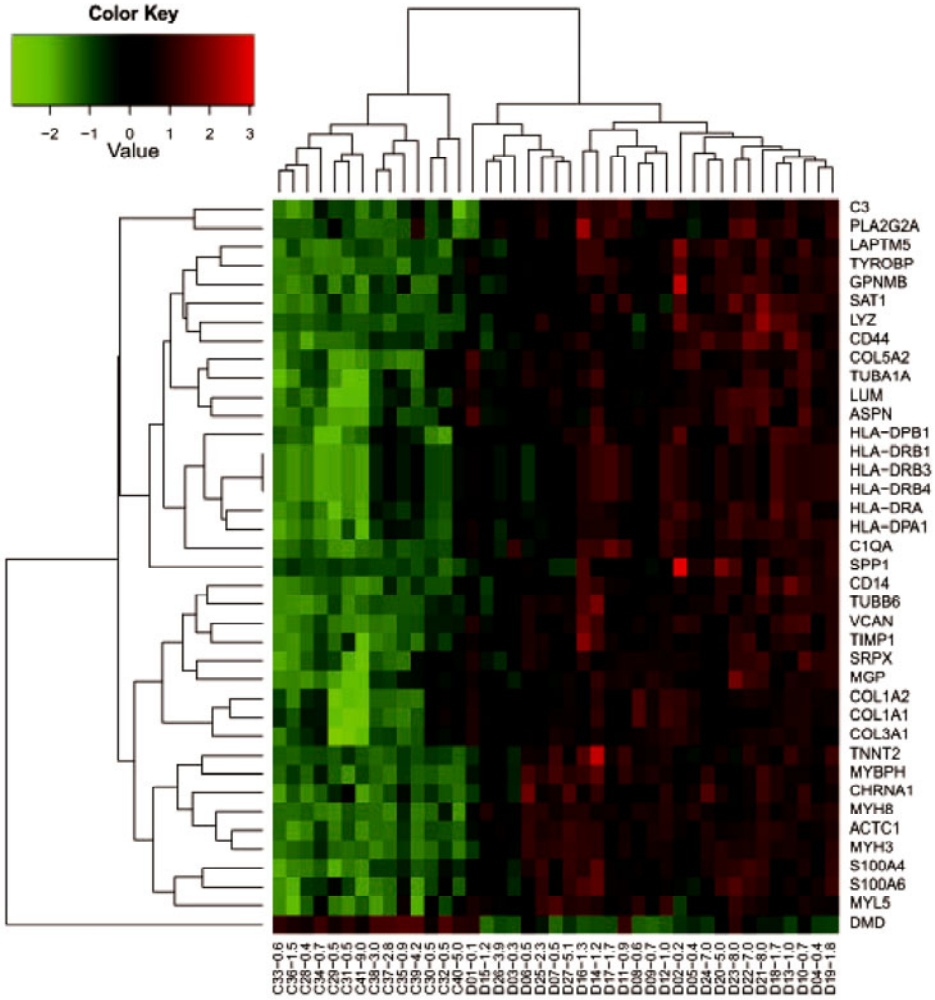

Fig. 2 Expression levels across all samples of the 39 differentially expressed genes with the absolute value of fold change (|FC|)>1.5

Information for each sample corresponding to Table 1 is shown in the *X* axis (format: ID-age). The 39 genes were combined and hierarchically clustered to represent the expression patterns using average linkage and Euclidean distance as a measurement of similarity. Red represents upregulation and green represents downregulation. Precise color scheme is illustrated in the color key (Note: for interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

Table 4 Pathways enriched with differentially expressed genes

ID	Pathway description	Pathway class	False discovery rate
04145	Phagosome	Transport and catabolism	1.63×10^{-12}
05150	Staphylococcus aureus infection	Infectious diseases	2.74×10^{-12}
04612	Antigen processing and presentation	Immune system	4.82×10^{-12}
05416	Viral myocarditis	Cardiovascular diseases	5.92×10^{-10}
05152	Tuberculosis	Infectious diseases	9.58×10^{-10}
05330	Allograft rejection	Immune diseases	1.18×10^{-8}
05140	Leishmaniasis	Infectious diseases	3.07×10^{-8}
05332	Graft-versus-host disease	Immune diseases	5.83×10^{-8}
04514	Cell adhesion molecules (CAMs)	Signaling molecules and interaction	1.35×10^{-7}
04940	Type I diabetes mellitus	Endocrine and metabolic diseases	1.47×10^{-7}
05145	Toxoplasmosis	Infectious diseases	2.00×10^{-7}
04510	Focal adhesion	Cell communication	4.50×10^{-7}
05130	Pathogenic Escherichia coli infection	Infectious diseases	1.03×10^{-6}
05133	Pertussis	Infectious diseases	2.16×10^{-6}
04672	Intestinal immune network for IgA production	Immune system	2.87×10^{-6}
04670	Leukocyte transendothelial migration	Immune system	5.04×10^{-6}
05310	Asthma	Immune diseases	5.80×10^{-6}
05320	Autoimmune thyroid disease	Immune diseases	6.88×10^{-6}
04512	ECM-receptor interaction	Signaling molecules and interaction	8.31×10^{-6}
05166	Human T-cell leukemia virus (HTLV)-I infection	Infectious diseases	9.21×10^{-6}
05169	Epstein-Barr virus infection	Infectious diseases	9.06×10^{-5}
04610	Complement and coagulation cascades	Immune system	1.02×10^{-4}
04142	Lysosome	Transport and catabolism	1.95×10^{-4}
04971	Gastric acid secretion	Digestive system	2.18×10^{-4}
00072	Synthesis and degradation of ketone bodies	Lipid metabolism	4.60×10^{-4}
05323	Rheumatoid arthritis	Immune diseases	5.11×10^{-4}
05146	Amoebiasis	Infectious diseases	6.60×10^{-4}
00020	Citrate cycle (TCA cycle)	Carbohydrate metabolism	7.71×10^{-4}
00071	Fatty acid metabolism	Lipid metabolism	1.28×10^{-3}
05012	Parkinson's disease	Neurodegenerative diseases	1.36×10^{-3}
04976	Bile secretion	Digestive dystem	1.38×10^{-3}
00640	Propanoate metabolism	Carbohydrate metabolism	1.86×10^{-3}
05322	Systemic lupus erythematosus	Immune diseases	2.21×10^{-3}
04380	Osteoclast differentiation	Development	2.70×10^{-3}
04810	Regulation of actin cytoskeleton	Cell motility	3.43×10^{-3}
05414	Dilated cardiomyopathy (DCM)	Cardiovascular diseases	5.22×10^{-3}
05168	Herpes simplex infection	Infectious diseases	5.79×10^{-3}
04974	Protein digestion and absorption	Digestive system	8.45×10^{-3}
05164	Influenza A	Infectious diseases	1.03×10^{-2}
05010	Alzheimer's disease	Neurodegenerative diseases	1.06×10^{-2}
04540	Gap junction	Cell communication	1.34×10^{-2}
05134	Legionellosis	Infectious diseases	1.50×10^{-2}
00062	Fatty acid elongation	Lipid metabolism	1.50×10^{-2}
00280	Valine, leucine and isoleucine degradation	Amino acid metabolism	1.56×10^{-2}
04650	Natural killer cell mediated cytotoxicity	Immune system	1.89×10^{-2}
05144	Malaria	Infectious diseases	1.89×10^{-2}
00630	Glyoxylate and dicarboxylate metabolism	Carbohydrate metabolism	1.97×10^{-2}
05222	Small cell lung cancer	Cancers	2.34×10^{-2}
00410	β-Alanine metabolism	Metabolism of other amino acids	2.51×10^{-2}
00010	Glycolysis/gluconeogenesis	Carbohydrate metabolism	2.83×10^{-2}
05220	Chronic myeloid leukemia	Cancers	2.92×10^{-2}
00650	Butanoate metabolism	Carbohydrate metabolism	3.02×10^{-2}
04725	Cholinergic synapse	Nervous system	3.59×10^{-2}
04141	Protein processing in endoplasmic reticulum	Folding, sorting and degradation	3.67×10^{-2}
00360	Phenylalanine metabolism	Amino acid metabolism	3.84×10^{-2}
05219	Bladder cancer	Cancers	4.39×10^{-2}
04640	Hematopoietic cell lineage	Immune system	4.44×10^{-2}
05214	Glioma	Cancers	4.46×10^{-2}

as the immune system, immune disease, and infectious disease. In addition, two DMD-related pathways, the dilated cardiomyopathy pathway and the viral myocarditis pathway, were included. Other pathways were metabolism, cell motility, signaling molecules and interaction, transport and catabolism, and so on.

4 Discussion

Progressive pathophysiology of DMD is highly complex, involving many secondary changes. Genome-wide expression profiling is a powerful procedure for investigating the downstream pathophysiological cascades in DMD patients. For expression profiling analysis, it is a challenge to create an effective mathematical model to deal with small sample sizes with a large number of genes (Golub *et al.*, 1999). Previous gene expression studies mainly used variance/regression analysis to identify the differentially expressed genes from their respective arrays for the two types of samples. However, with this procedure, the true picture of differential expression may be blurred by hidden biological effects that cannot be removed by a routine normalizing method. Chakraborty and Datta (2012) demonstrated better performance of the PLS-based method compared with variance/regression analysis. Our simulation study also revealed that the PLS-based method achieved higher sensitivity while maintaining reasonable specificity and impressively small FDR and FNR compared with the commonly used limma method (Table 2).

Here, with two combined datasets downloaded from the GEO database, we used a PLS-based multivariate linear model to describe the relationship between the genes expression and DMD disease status and further detected differentially expressed genes related to DMD pathogenesis. The age factor was also included in the model since DMD is a progressive disease. According to the logistical regression analysis, three latent variables were finally selected (Fig. 1), and they performed well in the classification of the samples. Total of 1254 genes were identified as significantly differentially expressed.

As consistent with previous studies (Chen *et al.*, 2000), upregulation of genes which encode proteins belonging to the major histocompatibility complex was found. In addition, most of the pathways in which aberrantly regulated enriched genes were involved resulted in an inflammatory/immune response. This may result from the infiltration of immune cells into the muscles (Mcdouall *et al.*, 1990; Spencer *et al.*, 1997; Cai *et al.*, 2000) and also in elevated levels of various inflammatory cytokines. For muscle regeneration, two genes which encode embryonic and perinatal myosin heavy chains (MYH3 and MYH8) were overexpressed, consistent with Haslett *et al.* (2002). *MYH8* was the most significantly differentially expressed gene. Their expression is considered as a hallmark of muscle regeneration after birth and muscular dystrophies. Other muscle structure and regeneration genes encoding tubulin (*TUBA1A* and *TUBB6*) or myosin (*MYBPH* and *MYL5*) were also overexpressed with |FC|>1.5, implicating the ongoing muscle regeneration process in DMD patients. In addition, increased expressions of genes encoding ECM components, such as fibril-forming collagens (types I and III), were also detected. This is also consistent with previous studies (Haslett *et al.*, 2002; Pescatori *et al.*, 2007), and increased ECM synthesis may contribute to the progressive fibrosis of muscle in DMD patients. Moreover, ECM-receptor interaction (hsa4512) is one of the pathways that are enriched with deregulated genes and all differentially expressed genes in this pathway are upregulated (Fig. 3).

Several deregulated genes with |FC|>1.5 have not been proposed in previous expression studies, such as *SRPX*, *GPNMB*, *SAT1*, and *LYZ*. Their relationships with DMD were still unknown. However, it is worth further investigation. Take the *SRPX* gene for example, it was upregulated with FC=1.73. SRPX is a cytoskeleton associated protein which interacts with Pelota (PELO). Their interaction may facilitate PELO to detect and degrade aberrant mRNAs (Burnicka-Turek *et al.*, 2010). Since PELO is associated with actin microfilaments of mammalian cells (Burnicka-Turek *et al.*, 2010), overexpression of *SRPX* may be related to the muscle regeneration process in DMD patients. For the pathways enriched with differentially expressed genes, the fatty acid metabolism pathway (has00071) was downregulated. As shown in Fig. 4, 12 genes in this pathway were differentially expressed and all of them were downregulated. Nishio *et al.* (1990) demonstrated that the free fatty

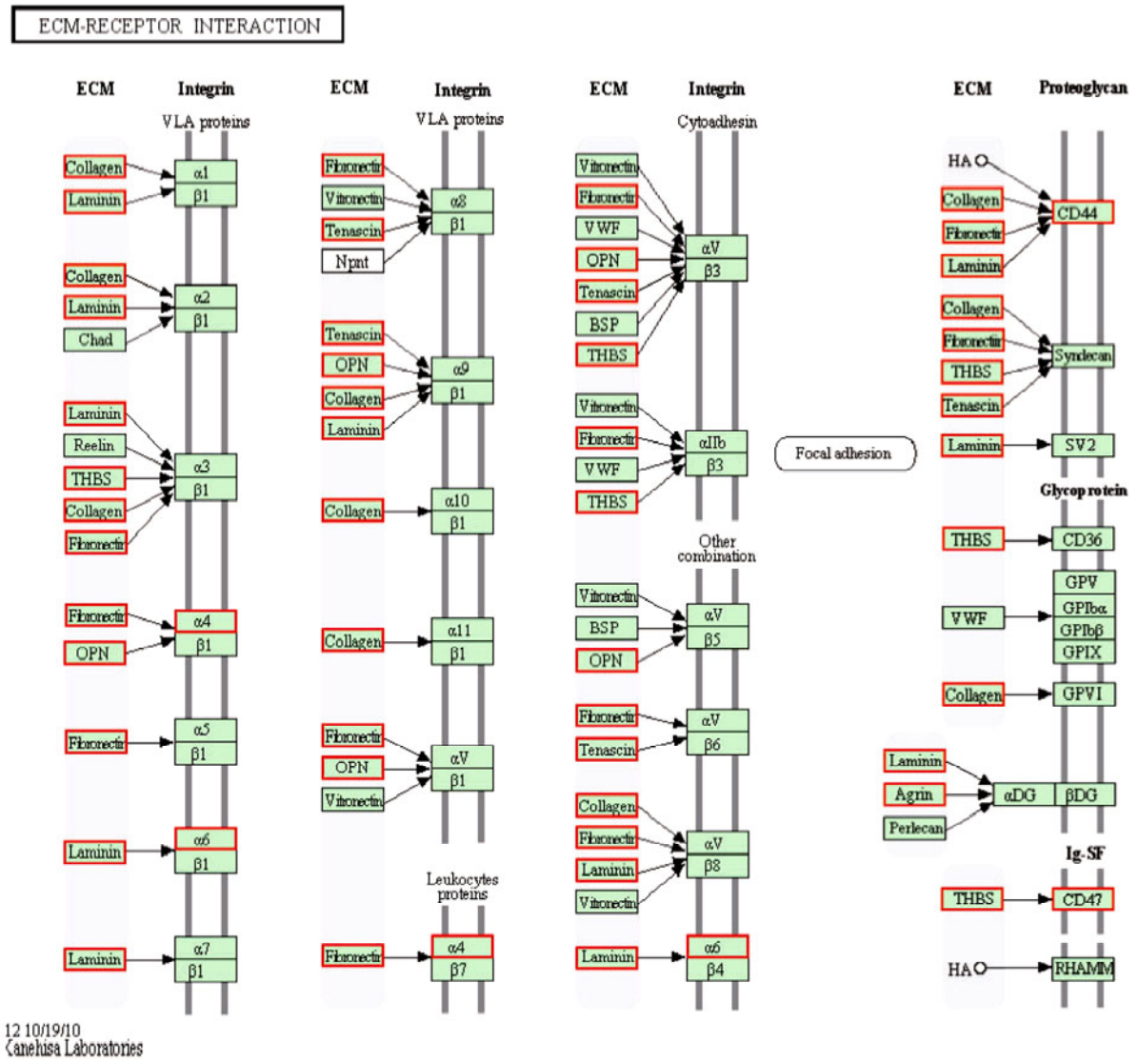

Fig. 3 Modified "ECM-receptor interaction" (hsa4512) pathway from KEGG

Protein symbols were marked according to gene expression pattern to reflect gene-centric data. Those with red frames are overexpressed (Note: for interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

acid level in DMD patients was higher than that in the control group no matter whether the serum creatine kinase level was high or low. Downregulation of the fatty acid metabolism pathway may be the underlying cause of the increased level of fatty acids. Increased concentration of fatty acid may serve as energy sources or substrates, sparing muscle protein (Nishio *et al.*, 1990). However, this fatty tissue infiltration process may also contribute to the progressive muscle wasting of DMD patients.

In summary, using two datasets downloaded from the GEO database, we carried out a PLS-based analysis to identify differentially expressed genes that may contribute to the pathology of DMD. Except for the genes related with inflammation, muscle regeneration, and ECM modeling, we found some genes with high FC, which have not been proposed by previous expression profile studies. In addition, we also found that all differentially expressed genes in the fatty acid metabolism pathway were downregulated,

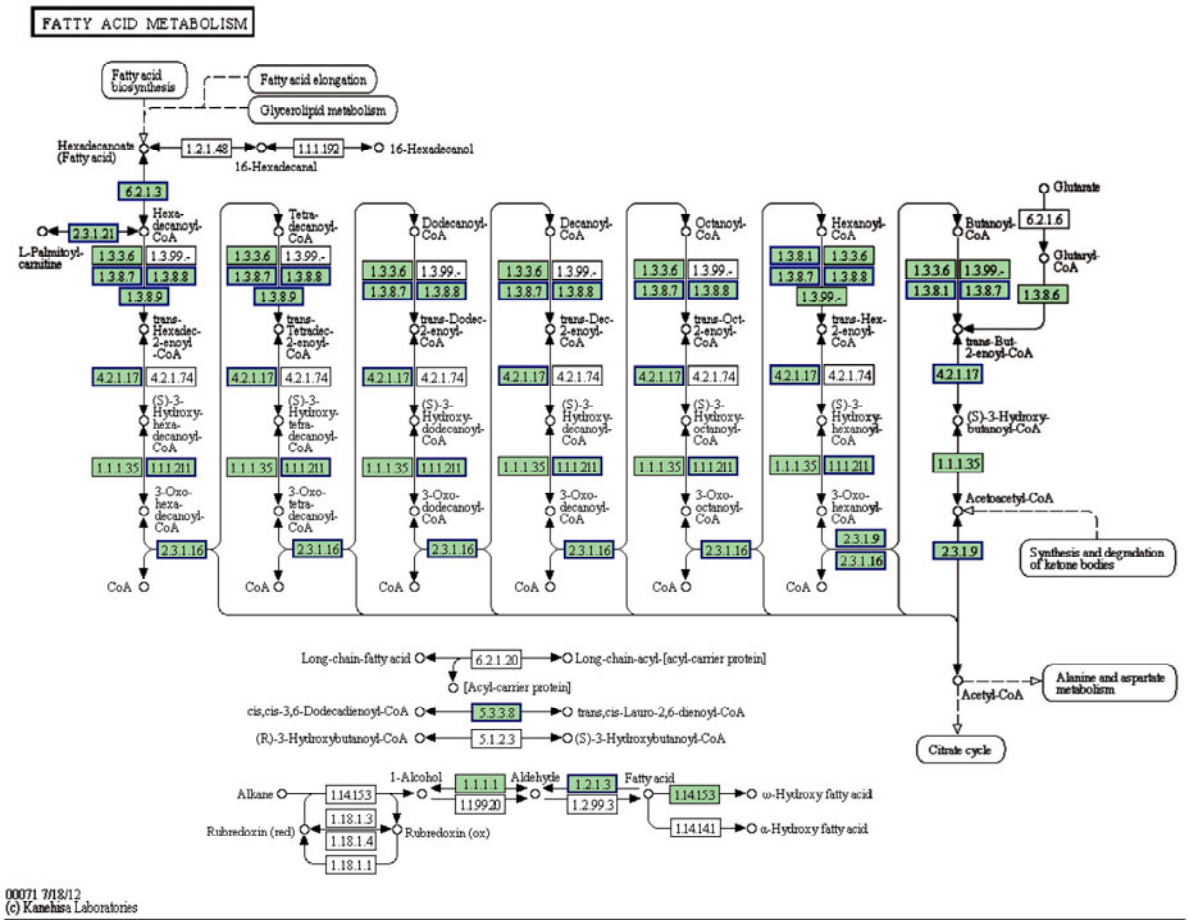

Fig. 4 Modified "fatty acid metabolism pathway" (hsa00071) from KEGG

Protein symbols were marked according to gene expression pattern to reflect gene-centric data. Those with green frames are downregulated (Note: for interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

which may be the cause of the high concentration of fatty acid in DMD patients and which is also related to the progressive muscle wasting process. Our results provide a better understanding for the downstream mechanisms of DMD and will further offer help for producing new adjuvant treatments.

Compliance with ethics guidelines

Hui-bo AN, Hua-cheng ZHENG, Li ZHANG, Lin MA, and Zheng-yan LIU declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Altamirano, F., López, J.R., Henriquez, C., Molinski, T., Allen, P.D., Jaimovich, E., 2012. Increased resting intracellular calcium modulates NF-κB-dependent inducible nitric-oxide synthase gene expression in dystrophic *mdx* skeletal myotubes. *J. Biol. Chem.*, **287**(25):20876-20887. [doi:10.1074/jbc.M112.344929]

Burnicka-Turek, O., Kata, A., Buyandelger, B., Ebermann, L., Kramann, N., Burfeind, P., Hoyer-Fender, S., Engel, W., Adham, I.M., 2010. Pelota interacts with HAX1, EIF3G and SRPX and the resulting protein complexes are associated with the actin cytoskeleton. *BMC Cell Biol.*, **11**(1):28. [doi:10.1186/1471-2121-11-28]

Bushby, K., Finkel, R., Birnkrant, D.J., Case, L.E., Clemens, P.R., Cripe, L., Kaul, A., Kinnett, K., Mcdonald, C., Pandya, S., *et al.*, 2010. Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and pharmacological and psychosocial management. *Lancet Neurol.*, **9**(1):77-93. [doi:10.1016/S1474-4422(09)70271-6]

Cai, B., Spencer, M.J., Nakamura, G., Tseng-Ong, L., Tidball, J.G., 2000. Eosinophilia of dystrophin-deficient muscle is promoted by perforin-mediated cytotoxicity by T cell effectors. *Am. J. Pathol*., **156**(5):1789-1796. [doi:10.1016/S0002-9440(10)65050-X]

Chakraborty, S., Datta, S., 2012. Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics*, **28**(6):799-806. [doi:10.1093/bioinformatics/bts022]

Chen, Y.W., Zhao, P., Borup, R., Hoffman, E.P., 2000. Expression profiling in the muscular dystrophies: identification of novel aspects of molecular pathophysiology. *J. Cell Biol*., **151**(6):1321-1336. [doi:10.1083/jcb.151.6.1321]

Chen, Y.W., Nagaraju, K., Bakay, M., Mcintyre, O., Rawat, R., Shi, R., Hoffman, E.P., 2005. Early onset of inflammation and later involvement of TGFβ in Duchenne muscular dystrophy. *Neurology*, **65**(6):826-834. [doi:10.1212/01.wnl.0000173836.09176.c4]

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., *et al.*, 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439):531-537. [doi:10.1126/science.286.5439.531]

Gorospe, J.R., Tharp, M.D., Hinckley, J., Kornegay, J.N., Hoffman, E.P., 1994. A role for mast cells in the progression of Duchenne muscular dystrophy? Correlations in dystrophin-deficient humans, dogs, and mice. *J. Neurol. Sci*., **122**(1):44-56. [doi:10.1016/0022-510X(94)90050-7]

Gosselin, R., Rodrigue, D., Duchesne, C., 2010. A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemometr. Intell. Lab. Syst.*, **100**(1):12-21. [doi:10.1016/j.chemolab.2009.09.005]

Haslett, J.N., Sanoudou, D., Kho, A.T., Bennett, R.R., Greenberg, S.A., Kohane, I.S., Beggs, A.H., Kunkel, L.M., 2002. Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle. *PNAS*, **99**(23):15000-15005. [doi:10.1073/pnas.192571199]

Head, S.I., 2010. Branched fibres in old dystrophic *mdx* muscle are associated with mechanical weakening of the sarcolemma, abnormal Ca^{2+} transients and a breakdown of Ca^{2+} homeostasis during fatigue. *Exp. Physiol*., **95**(5):641-656. [doi:10.1113/expphysiol.2009.052019]

Helland, I.S., 1988. On the structure of partial least squares regression. *Commun. Stat. Simul. Comput.*, **17**(2):581-607. [doi:10.1080/03610918808812681]

Helland, I.S., 1990. Partial least squares regression and statistical model. *Scand. J. Stat*., **17**:97-144.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2):249-264. [doi:10.1093/biostatistics/4.2.249]

Koenig, M., Hoffman, E.P., Bertelson, C.J., Monaco, A.P., Feener, C., Kunkel, L.M., 1987. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the *DMD* gene in normal and affected individuals. *Cell*, **50**(3):509-517. [doi:10.1016/0092-8674(87)90504-6]

Kunkel, L.M., Monaco, A.P., Hoffman, E., Koenig, M., Feener, C., Bertelson, C., 1987. Molecular studies of progressive muscular dystrophy (Duchenne). *Enzyme*, **38**(1-4):72-75.

Martins, J.P.A., Teofilo, R.F., Ferreira, M.M.C., 2010. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. *J. Chemometr.*, **24**(5-6):320-332. [doi:10.1002/cem.1309]

Mcdouall, R.M., Dunn, M.J., Dubowitz, V., 1990. Nature of the mononuclear infiltrate and the mechanism of muscle damage in juvenile dermatomyositis and Duchenne muscular dystrophy. *J. Neurol. Sci*., **99**(2-3):199-217. [doi:10.1016/0022-510X(90)90156-H]

Monici, M.C., Aguennouz, M., Mazzeo, A., Messina, C., Vita, G., 2003. Activation of nuclear factor-κB in inflammatory myopathies and Duchenne muscular dystrophy. *Neurology*, **60**(6):993-997. [doi:10.1212/01.WNL.0000049913.27181.51]

Nishio, H., Wada, H., Matsuo, T., Horikawa, H., Takahashi, K., Nakajima, T., Matsuo, M., Nakamura, H., 1990. Glucose, free fatty acid and ketone body metabolism in Duchenne muscular dystrophy. *Brain Dev*., **12**(4):390-402. [doi:10.1016/S0387-7604(12)80071-4]

Pescatori, M., Broccolini, A., Minetti, C., Bertini, E., Bruno, C., D′Amico, A., Bernardini, C., Mirabella, M., Silvestri, G., Giglio, V., *et al.*, 2007. Gene expression profiling in the early phases of DMD: a constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression. *FASEB J*., **21**(4):1210-1226. [doi:10.1096/fj.06-7285com]

Spencer, M.J., Walsh, C.M., Dorshkind, K.A., Rodriguez, E.M., Tidball, J.G., 1997. Myonuclear apoptosis in dystrophic mdx muscle occurs by perforin-mediated cytotoxicity. *J. Clin. Invest*., **99**(11):2745-2751. [doi:10.1172/JCI119464]

Straub, V., Campbell, K.P., 1997. Muscular dystrophies and the dystrophin-glycoprotein complex. *Curr. Opin. Neurol*., **10**(2):168-175.

Wong, B., Gilbert, D.L., Walker, W.L., Liao, I.H., Lit, L., Stamova, B., Jickling, G., Apperson, M., Sharp, F.R., 2009. Gene expression in blood of subjects with Duchenne muscular dystrophy. *Neurogenetics*, **10**(2):117-125. [doi:10.1007/s10048-008-0167-8]

List of electronic supplementary materials

Table S1 Information of differentially expressed genes in DMD samples