Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2012 *Paper* 99

Statistical hypothesis test of factor loading in principal component analysis and its application to metabolite set enrichment analysis

Hiroyuki Yamamoto* Tamaki Fujimori† Hajime Sato‡
Gen Ishikawa** Kenjiro Kami†† Yoshiaki Ohashi‡‡

*Human Metabolome Technologies, Inc., h.yama2396@gmail.com
†Human Metabolome Technologies, Inc., fujimori@humanmetabolome.com
‡Human Metabolome Technologies, Inc., hsato@humanmetabolome.com
**Human Metabolome Technologies, Inc., gggishi@yahoo.co.jp
††Human Metabolome Technologies, Inc., kkami@humanmetabolome.com
‡‡Human Metabolome Technologies, Inc., ohashi@humanmetabolome.com

Statistical hypothesis test of factor loading in principal component analysis and its application to metabolite set enrichment analysis

Hiroyuki Yamamoto, Tamaki Fujimori, Hajime Sato, Gen Ishikawa, Kenjiro Kami, and Yoshiaki Ohashi

Abstract

Principal component analysis (PCA) has been widely used to visualize high-dimensional metabolomic data in a two- or three-dimensional subspace. In metabolomics, some metabolites (e.g. top 10 metabolites) have been subjectively selected when using factor loading in PCA, and biological inferences for these metabolites are made. However, this approach is possible to lead biased biological inferences because these metabolites are not objectively selected by statistical criterion. We proposed a statistical procedure to pick up metabolites by statistical hypothesis test of factor loading in PCA and make biological inferences by metabolite set enrichment analysis (MSEA) for these significant metabolites. This procedure depends on the fact that the eigenvector in PCA for autoscaled data is proportional to the correlation coefficient between PC score and each metabolite levels. We applied this approach for two metabolomic data of mice liver samples. 136 of 282 metabolites in first case study and 66 of 275 metabolites in second case study were statistically significant. This result suggests that to set the previously-determined number of metabolites is not appropriate because the number of significant metabolites is different in each study when using factor loading in PCA. Moreover, MSEA was performed for these significant metabolites and significant metabolic pathways can be detected. These results are acceptable when compared with previous biological knowledge. It is essential to select metabolites statistically for making unbiased biological inferences from metabolome data, when using factor loading in PCA. We proposed a statistical procedure to pick up metabolites by statistical hypothesis test of factor loading in PCA and make biological

inferences by MSEA for these significant metabolites. We developed an R package ”mseapca” to perform this approach. The “mseapca” package is publicity available on CRAN website.

Background

Metabolomics is a science based on exhaustive-profiling of metabolites. In metabolomics, various analytical technologies such as capillary electrophoresis–mass spectrometry (CE–MS), liquid chromatography–mass spectrometry (LC–MS), gas chromatography–mass spectrometry (GC–MS) and nuclear magnetic resonance (NMR) etc. have been used. Statistical analysis for these analytical data has been studied in chemometrics research area [1]. In metabolomics, chemometrics approaches commercing with multivariate analysis such as principal component analysis (PCA) have been mainly applied.

PCA [2] has been routinely used to visualize high-dimensional metabolomic data in two- or three-dimensional subspace in metabolomics as well as a heat map in transcriptmics. A scatter plot of PC score vectors (a score plot) can be used for outlier detection or discovering biologically interpretable patterns. Typically, upon finding a specific PC score is related to a phenotype of interest [3, 4], such as time course or group information, the corresponding factor loading has been evaluated to discover meaningful metabolites for making biological inferences.

In many metabolomics research articles [5, 6, 7], an eigenvector in PCA (eq. 1) has been used as factor loading. For making biological inferences, some metabolites (e.g. top 10 metabolites) has been subjectively selected by using the eigenvector. However, this approach has some problems. For example, many metabolites varies with phenotype in a study and only a few metabolites in another study. An existing approach using the eigenvector equals to set the same number of metabolites for making biological inferences among these different studies. As a result, biological interpretation might be performed by using not significant metabolites varied with phenotype irrelevantly. Additionally, the value of eigenvector itself does not have statistical sense because it is nothing but normalized that the sum of squares is 1 just for computational reason.

The eigenvectors in PCA for autoscaled data [8] are proportional to correlation coefficient between PC scores and variables. This fact is well-known in multivariate analysis literature [9] but appears not to be appreciated in metabolomics. In the present study, factor loading was defined as correlation coefficient between PC scores and variables. This definition can be used to

perform statistical hypothesis testing and can select significant metabolites objectively by statistical criteria.

Significant metabolites are selected in some way and afterward biological inferences are made for these metabolites by biologists. They often make biological inferences with respect to a biological functional unit such as metabolic pathway (e.g. “glycolysis is notably activated” or “an amino acid metabolism is significantly suppressed”). In gene expression analysis, gene set enrichment analysis (GSEA) has been applied to find significant gene sets by using gene ontology (GO) terms. In metabolomics, metabolite set enrichment analysis (MSEA) [10] can find significant metabolic pathway. MSEA has been computed by some approaches such as over representation analysis (ORA) [11], Subramanian’s GSEA [12] and global test [13]. MSEA is convenience to make biological inferences from metabolomic data, but this approach has not been applied for metabolites selected by factor loading in PCA.

In the present study, we performed statistical hypothesis test of factor loading in PCA for two metabolomic data sets of mice liver samples as case studies. This approach can select significant metabolites when using factor loading in PCA, and MSEA by ORA approach can be applied for these significant metabolites. We developed the R package ‘mseapca’ to work sequence from statistical hypothesis test of factor loading in PCA to MSEA.

Theory

Principal Component Analysis (PCA)

Consider a mean-centered data matrix $\mathbf{X}$ that has samples in each row and variables in each column. A score vector is related to the data matrix by $\mathbf{t} = \mathbf{Xw}$ where $\mathbf{w}$ is a vector of weights. PCA is formulated as the optimization problem of maximizing the variance of the score vector $\mathbf{t}$:

$$\begin{aligned} &\max \mathrm{var}(\mathbf{t}) \\ &\textit{subject to } \mathbf{w}'\mathbf{w} = 1 \end{aligned} \qquad (1\text{-}1)$$

and the weight vector $\mathbf{w}$ is often used for factor loading. After a transformation, eq. (1-1) can be rewritten as the eigenvalue problem

$$\frac{1}{n-1}\mathbf{X}'\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \qquad (1\text{-}2)$$

An eigenvector $\mathbf{w}$ and eigenvalue λ of eq. (1-2) can be computed by using numerical computation libraries for singular value decomposition. An eigenvalue λ corresponds to the variance of the PC score vector formed using the associated eigenvector as the weight vector.

A coefficient of the correlation between the PC score and the p-th variable can be defined as

$$\mathrm{corr}\,(\mathbf{t}, \mathbf{x}_\mathbf{p}) = \frac{\mathbf{t}'\mathbf{x}_\mathbf{p}/n-1}{\sqrt{\mathrm{var}(\mathbf{t})}\sqrt{\mathrm{var}(\mathbf{x}_p)}} \qquad (1\text{-}3)$$

where $\mathbf{t}'$ is the transpose of $\mathbf{t}$. Introducing $\mathbf{c}$ as the column vector in which the p-th element is 1 and the others are 0, so that $\mathbf{x_p} = \mathbf{Xc}$, we have

$$\mathrm{corr}(\mathbf{t}, \mathbf{x}_\mathbf{p}) = \frac{\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{c}/n-1}{\sqrt{\mathrm{var}(\mathbf{t})}\sqrt{\mathrm{var}(\mathbf{x}_\mathbf{p})}} \qquad (1\text{-}4)$$

Transposing eq. (1-2) gives $\mathbf{w}'\mathbf{X}'\mathbf{X}/n-1 = \lambda\mathbf{w}'$ which can be substituted in eq. (1-4), giving

$$\mathrm{corr}(\mathbf{t}, \mathbf{x}_\mathbf{p}) = \frac{\lambda\mathbf{w}'\mathbf{c}}{\sqrt{\mathrm{var}(\mathbf{t})}\sqrt{\mathrm{var}(\mathbf{x}_p)}} \qquad (1\text{-}5)$$

Then the variance of the score vector can be replaced by λ and the standard deviation of $\mathbf{x_p}$ is replaced by σ_p. Finally, the correlation between PC score and variables can be written as

$$\mathrm{corr}(\mathbf{t}, \mathbf{x}_\mathbf{p}) = \frac{\lambda\mathbf{w}'\mathbf{c}}{\sqrt{\mathrm{var}(\mathbf{t})}\sqrt{\mathrm{var}(\mathbf{x}_\mathbf{p})}} = \frac{\lambda w_p}{\sqrt{\lambda}\sigma_p} = \frac{\sqrt{\lambda}w_p}{\sigma_p} \qquad (1\text{-}6)$$

With data scaled to unit variance (autoscaling), the weight w_p is proportional to the correlation coefficient between the PC score and variable $\mathbf{x_p}$ because $\sigma_p = 1$ in eq. (1-6). Thus, the factor loading can be defined as the correlation coefficient in eq. (1.6). On the basis of this definition, we can perform a statistical test for factor loading in PCA, using the well-known fact that, for a correlation coefficient r, the

statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (1\text{-}7)$$

has a *t*-distribution with (n-2) degrees of freedom. Then, we can select variables that have a statistically significant correlation with the PC score and make biological inferences using these variables.

Materials and Methods

Sample preparation, metabolome analysis and data processing

BKS.Cg-*m*+/*m*+/Jcl (normal mice), 12-h fasting normal mice, BKS.C-+*Lepr*^{db}/+*Lepr*^{db}/Jcl (*db*/*db* mice) and *db*/*db* mice orally administered pioglitazone for 10 days were used. The mice were seven-week-old males with unlimited access to food and water except for those on the 12-h fasting. The concentration of administered pioglitazone was 100 mg/10 mL/kg. The pioglitazone was purchased from Takeda Pharmaceutical Co. ltd, and purified by NARD Institute Ltd. After sampling, the livers were excised and stored at –80 ℃. From purchase and breeding of the mice to collection of liver samples, all experiments were performed at the Kitayama Labs Co. Ltd. The sample preparation procedure to extract metabolites was that described by Ooga et al [14].

The metabolite extracts were measured by capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS). CE-TOFMS was carried out using Agilent Capillary Electrophoresis Systems equipped with an Agilent 6210 time-of-flight mass spectrometer, an Agilent 1100 isocratic HPLC pump, an Agilent G1603A CE-MS adapter kit and an Agilent G1607A CE-ESI-MS sprayer kit (Agilent Technologies, Waldbronn, Germany). The system was controlled by G2201AA ChemStation Software version B.03.01 for CE (Agilent Technologies, Waldbronn, Germany). Modified analytical methods for the measurement of cationic [15] and anionic metabolites [16] were used. Measurement data were processed by peak processing software [17]. Signal peaks corresponding to isotopomers, adduct ions and other product ions of known metabolites were

excluded. Then all signal peaks potentially corresponding to authentic compounds were extracted, and their migration time (MT) was normalized using those of the internal standards (methionine sulfone and CSA for cations and anions, respectively). Thereafter, the alignment of peaks was performed according to the *m/z* values and normalized MT values. Finally, peak areas are normalized against those of the internal standards. The resultant relative area values were further normalized by the sample weight. Annotation tables were produced from CE-TOFMS measurements of standard compounds, and were aligned with datasets according to similar *m/z* values and normalized MT values.

Statistical Analysis

In the present study, all computation were performed by R [18] and "mseapca" [19] package. A missing value was imputed to 0 for a computation of PCA. A metabolite set list was created by referring to KEGG [20], and partial modification was performed by manual curation. The xml file of metabolite set list used in this study is included in "mseapca" package.

Software

An R package "mseapca" [19] consists of three major features. The first one is to create a list of metabolic pathway. A "*csv2list*" function converts your own csv file in which first column is name of metabolic pathway and second column is metabolite IDs to list format in R. A "*pathway_class*" function converts KEGG's tar.gz files (e.g. *hsa.tar.gz* in *Homo sapience*) to list format of metabolic pathway. The KEGG's tar.gz files can be downloaded from KEGG FTP according to your own license. A "*list2xml*" function converts list format of metabolic pathway to xml format. This xml format can be saved as xml file by using "*saveXML*" function in "*XML*" package. A "*read_pathway*" function can read the created xml file and convert to list of metabolic pathway for a computation of MSEA.

The second one is a "*pca_scaled*" function to perform PCA. In this function, data matrix is automatically scaled to zero mean and unit variance (autoscaling) for each metabolites. PC scores, factor loadings and p-value and

q-value by Benjamini and Hochberg [21] as the results of statistical hypothesis test of factor loading are returned. In this function, factor loading is defined as a correlation coefficient between PC score and each metabolite levels.

The third one is to perform MSEA. A "*msea_ora*" function can perform MSEA by over representation analysis (ORA) [11]. In this function, statistical hypothesis test of cross tabulation is performed by one-sided Fisher's exact test. A "*msea_sub*" function performs MSEA implemented in the same fashion as GSEA by Subramanian et al. In this function, a permutation procedure is performed for a metabolite set rather than class label. This procedure corresponds to "gene set" of permutation type in GSEA-P software [22]. A leading-edge subset analysis is also undertaken following the GSEA procedure [21].

An R package "mseapca" can be freely available from CRAN web site [19]. For details, see the reference manual in CRAN web site of "mseapca" [19] for more information.

4. Results

4.1. Case study 1: a comparative study of control and 12-h fasting mice

We describe the utilization of statistical hypothesis test of factor loading in PCA by using metabolome data in two studies. First case study is a comparative study of normal and 12 hour fasting mouse. 5 liver samples each for control and 12 hour fasting mice were used for metabolome analysis. As a result, 282 metabolites were identified.

PCA was performed for this metabolome data preprocessed by autoscaling. The score plot of PCA (Fig. 1 (A)) showed that the PC1 score of control and fasting mice were negative and positive, respectively. This result suggests that the PC1 score was positively related to fasting effect. In this case, metabolites which have large positive factor loading in PC1 tended to increase and negative factor loading tended to decreased with 12 hour fasting.

Statistical hypothesis test for factor loading in PC1 was performed, and 136 metabolites were statistically significant under $p<0.05$ (Supplementary Table1). MSEA by ORA for factor loading was performed for positive and negatively significant metabolites independently (Table1). The purine metabolism

was significantly activated in 12 hour fasting mice under $p<0.05$. Glycolysis was significantly suppressed under $q<0.05$ and pentose phospate pathway, TCA cycle, cystein metabolism and polyamine metabolism were significantly suprressed in 12 hour fasting under $p<0.05$. MSEA by Subramanian et al. was also performed as a reference (Table1). The histidine metabolism and purine metabolism having negative normalized enrichment score (NES) were significantly activated in 12 hour fasting mice under $p<0.05$. Glycolysis having positive NES was significantly suppressed under $q<0.05$ and pentose phospate pathway, TCA cycle and polyamine metabolism were significantly suppressed in 12 hour fasting under $p<0.05$. These results suggest that the results of these two MSEA approaches are largely consistence.

The results of MSEA for factor loading in PC1 suggested that the process of energy metabolism, such as glycolysis and the TCA cycle, were decreased by 12 hour fasting. The suppression of these metabolic pathway suggests that glycogen is drained and glucose supplementation is restricted in mouse liver under the condition of fasting 12 hour. And, body weight was 22.20±0.84 (mean±SD) in normal mouse and 20.0±0.71 in 12 hour fasting mouse, and statistically significant decreased (p=0.0021) during 12 hour fasting by Welch's *t*-test. This result shows that the suppression of energy metabolism might result in a decrease of body weight.

4.2. Case study 2: a comparative study between diabetes model mice without and with administered pioglitazone

The *db*/*db* mouse is a model of obesity, diabetes and dyslipidemia in which leptin receptor activity is deficient because the mice are homozygous for a point mutation in the leptin receptor gene [23]. Pioglitazone reduces insulin resistance in the liver and decreases glucose level in the blood [24, 25]. Hence, it is used for the treatment of diabetes.

We compared the metabolome data of mouse liver samples from *db*/*db* mice with and without pioglitazone in order to examine the effect of administering pioglitazone to *db*/*db* mice. 5 liver samples each for db/db mice and db/db with

administering pioglitazone were used for metabolome analysis. As a result, 275 metabolites were identified.

We performed PCA on data preprocessed by autoscaling in a comparative study of *db*/*db* mice with and without administered pioglitazone. The score plot is shown in Fig. 5. A perfect separation between groups could be achieved on the first PC axis (Fig. 3), and thus we focused on this axis. The PC1 score of the *db*/*db* mice with and without administered pioglitazone showed positive and negative value, respectively, suggesting that PC1 score is positively related to the effect of the administration of pioglitazone.

Statistical hypothesis test for factor loading in PC1 was performed, and 66 metabolites were statistically significant under $p<0.05$. MSEA for factor loading was performed as well as the previous section (Table2). In both MSEA by ORA and Subramanian's approach, glycolysis was only statistically significant activated by administered pioglitazone under $p<0.05$. Pioglitazone is a peroxisome proliferator-activated receptor (PPAR) activating agent. Lee et al. [26] suggested that PPARδ ameliorated hyperglycemia by increasing glucose flux through the regulation of gene expression. And administering pioglitazone was known to reduce the glucose level in blood [24, 25].

In the present study, glucose blood level was 369.6±64.8 (mean±SD) in *db*/*db* mouse and 332.8±131.9 in *db*/*db* mice with administered pioglitazone. The change of glucose blood level was not significantly decreased ($p=0.596$) by administered pioglitazone by Welch's *t*-test. This result suggests that metabolome analysis could detect the subtle change caused by administering pioglitazone in glycolysis pathway.

Discussion

Metabolite selection by statistical hypothesis test of factor loading in PCA has some advantages. This approach was applied for metabolome data in two case studies of mouse liver samples. 136 of 282 metabolites were significantly correlated with PC1 score associated with groups in first case study and 66 of 275 metabolites in second study. The number of significant metabolites is twice larger in first case study than in second case study. This result suggests that to set the

previously-determined number of metabolites (e.g. top 10 metabolites) is not appropriate because the number of significant metabolites is different in each studies. Additionally, we note the relationship between contribution ratio and the number of significant metabolites for factor loading in PCA. The ratio of the number of significant metabolites to all detected metabolites is 0.482 (=136/282) in first case study and 0.24 (=66/275) in second case study. The contribution ratio in PC1 is 40.5% in first case study and 24.2% in second case study. This result suggests that an implicit relationship between contribution ratio and the number of significant metabolites under same sample size.

In both two case studies, we focused on PC1 (Fig.1 and Fig.2) which shows difference between groups. Then, we compared this approach with ordinary statistical hypothesis test such as *t*-test. 122 metabolites were significant in first case study and 56 metabolites were significant in second case study by Welch's *t*-test. As compared with metabolites picked up by statistical test of factor loading, 112 metabolites and 47 metabolites shared in each study. The fact that most significant metabolites shared between two approaches suggests that statistical test of factor loading in PCA can be readily used to pick up metabolites as a special case of two-sample test when a difference between groups appears in PC score. On the other hand, in metabolomics, complex studies such as a fermentation process by microorganism [27] having various time points or groups and drug administration having various concentration in various condition [28]. Statistical test for factor loading in PCA can be widely used not only two-sample study but also various studies when PC score associated with phenotype can be found.

MSEA was performed for significant metabolites and acceptable biological inferences were given in two case studies. In conventional approach, previously-determined number of metabolite (e.g. 10 metabolites) has been subjectively selected for making biological inferences. According to this approach, MSEA was performed for the top 10 metabolites having large negative factor loading in first case study. No significant metabolic pathway was detected under $p<0.05$ (data not shown). In this case, only 10 metabolites were too small for making acceptable biological inferences. Even if significant metabolic pathways

were detected by MSEA for not significant metabolites, it is doubtful whether these significant metabolic pathways are meaningful statistically and biologically. To make unbiased biological inferences by using statistical analysis, significant metabolites should be selected by using statistical test of factor loading when using in PCA.

In the present study, two MSEA approaches of ORA and Subramanian's approach were performed. As a way of utilizing factor loading for GSEA, Rudolf S. N. Fehrmann et al. [29] named the PC score associated with phenotype as transcriptional system regulators (TSR) score, and factor loading corresponding to TSR score was used for GSEA by Subramanian's approach. This directly uses the factor loading but does not use the result of statistical hypothesis test of factor loading. As far as we know, an approach combining GSEA or MSEA with the result of statistical hypothesis test of factor loading in PCA has not been reported.

The result of both MSEA of ORA and Subramanian's approach has almost same results in two case studies. In a comparison of computational time between two MSEA approaches, it cost 441.43 seconds by Subramanian's approach and 0.83 seconds by ORA in first case study. This result showed that MSEA of ORA has an advantage about computational cost. Conventionally, PCA and MSEA can be computed in different step or software independently. There has not been software which can compute as sequence from PCA and statistical hypothesis test of factor loading to MSEA. Therefore, we developed an R package "mseapca" to work sequence from statistical hypothesis test of factor loading in PCA to MSEA.

Conclusion

In metabolomics, targeted metabolites for making biological inferences were subjectively selected when using factor loading in PCA. We proposed a statistical procedure to pick up metabolites by statistical hypothesis test of factor loading in PCA, and these significant metabolites was used to find significant metabolic pathway by MSEA. We applied this approach to two metabolomic data of mouse liver samples, and acceptable results were given when compared with previous biological knowledge. We developed an R package "mseapca" to examine our approach readily. There are many users of PCA in metabolomics.

Our approach can improve an existing use of PCA and is expected to be widely applied.

References

[1] Lavine B, Workman J: **Chemometrics**. *Anal Chem*. 2010, **82**(12):4699-4711.

[2] Jolliffe I T: **Principal Component Analysis**. 2nd edition. New York: Springer-Verlag; 2002.

[3] Ringnér M: **What is principal component analysis?** *Nat. Biotechnol.* 2008, **26**:303-304.

[4] Landgrebe J, Wurst W, Welzl G: **Permutation-validated principal components analysis of microarray data**. *Genome Biol.* 2002, **3**(4).

[5] Dileo M V, Strahan G D, den Bakker M and Hoekenga O A; **Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome**. *PLoS One.* 2011, **6**(10).

[6] Dewar B J, Keshari K, Jeffries R, Dzeja P, Graves L M and Macdonald J M, **Metabolic assessment of a novel chronic myelogenous leukemic cell line and an imatinib resistant subline by H NMR spectroscopy.** *Metabolomics*. 2010, **6**(3):439-450.

[7] Maruyama K, Takeda M, Kidokoro S, Yamada K, Sakuma Y, Urano K, Fujita M, Yoshiwara K, Matsukura S, Morishita Y, Sasaki R, Suzuki H, Saito K, Shibata D, Shinozaki K, Yamaguchi-Shinozaki K: **Metabolic Pathways Involved in Cold Acclimation Identified by Integrated Analysis of Metabolites and Transcripts Regulated by DREB1A and DREB2A.** *Plant Physiol.* 2009, **150**(4):1972-80.

[8] Van den Berg R A, Hoefsloot H C, Westerhuis J A, Smilde A K, van der Werf M J: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics*. 2006, **7**.

[9] Afifi A, May S, Clark V A: **Practical Multivariate Analysis.** 5th Edition. London: Chapman and Hall/CRC; 2011.

[10] Xia J and Wishart D S: **Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst.** *Nature Protocols*. 2011, **6**: 743-760.

[11] Draghici S, Khatri P, Martins R P, Ostermeier G C, Krawetz S A: **Global function profiling of gene expression.** *Genomics.* 2003, **81**:98-104.

[12] Subramanian A, Tamayo P, Mootha V K, Mukherjee S, Ebert B L, Gillette M A, Paulovich A, Pomeroy S L, Golub T R, Lander E S and Mesirov J P: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *PNAS.* 2005, **102**(43):15545-15550.

[13] Goeman J J, van de Geer S A, de Kort F, van Houwelingen H C: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics.* 2004, **20**(1): 93-99.

[14] Ooga T, Sato H, Nagashima A, Sasaki K, Tomita M, Soga T, Ohashi Y: **Metabolomic anatomy of an animal model revealing homeostatic imbalances in dyslipidaemia**. *Mol. BioSys*. 2011, **7**(4):1217-23.

[15] Soga T, Heiger D N: **Amino Acid Analysis by Capillary Electrophoresis Electrospray Ionization Mass Spectrometry**. *Anal. Chem.* 2000, **72**: 1236-1241.

[16] Soga T, Ueno Y, Naraoka H, Ohashi Y, Tomita M, Nishioka T: **Analysis of nucleotides by pressure-assisted capillary electrophoresis mass spectrometry using silanol mask technique.** *J. Chromatogr. A*. 2007, **1159**:125-133.

[17] Sugimoto M, Wong D, Hirayama A, Soga T, Tomita M: **Capillary Electrophoresis Mass Spectrometry-based Saliva Metabolomics Identifies Oral, Breast and Pancreatic Cancer-Specific Profiles**. *Metabolomics* 2010, **6**:78-95.

[18] R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria; 2005.

[19] mseapca: Metabolite set enrichment analysis for factor loading in principal component analysis [http://cran.r-project.org/web/packages/mseapca/]

[20] Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res*. 2000, **28**: 27-30.

[21] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J. Roy. Statist. Soc. Ser. B* 1995, **57**(1):289–300.

[22] Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: **GSEA-P: a desktop application for Gene Set Enrichment Analysis**. *Bioinformatics* 2007, **23**(23):3251-3253.

[23] Sharma K, McCue P, Dunn SR: **Diabetic kidney disease in the db/db mouse**. *Am. J. Physiol. Renal Physiol.* 2003, **284**:F1138-1144.

[24] Kemnitz JW, Elson DF, Roecker EB, Baum, ST, Bergman RN, and Maglasson MD: **Pioglitazone increases insulin sensitivity, reduces blood glucose, insulin, and lipid levels, and lowers blood pressure, in obese, insulin-resistant rhesus monkeys**. *Diabetes* 1994, **43**, 204-211.

[25] Smith U: **Pioglitazone: mechanism of action**. *Int. J. Clin. Pract.* 2001, **121**, 13-18.

[26] Lee CH., Olson P, Hevener A, Mehl I, Chong LW, Olefsky JM., Gonzalez FJ., Ham J, Kang H, Peters JM, Evans RM: **PPARδ regulates glucose metabolism and insulin sensitivity.** *PNAS* 2006, **103**(9), 3444-3449.

[27] Yamamoto H, Yamaji H, Abe Y, Harada K, Waluyo D, Fukusaki E, Kondo A, Ohno H, Fukuda H: **Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables**. *Chemom. Intell. Lab. Syst.* 2009, **98**(2):136-142.

[28] Smilde AK, Jeroen J. Jansen, Huub C. J. Hoefsloot, Lamers RAN, van der Greef J, Timmerman ME: **ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data**. *Bioinformatics* **21**(13): 3043-3048.

[29] Fehrmann RSN, de Jonge HJM, ter Elst A, de Vries A, Crijns AGP, Weidenaar AC, Gerbens F, de Jong S, van der Zee AGJ, de Vries EGE, Kamps WA, Hofstra RMW, te Meerman GJ, de Bont ESJM: **A New Perspective on Transcriptional System Regulation (TSR): Towards TSR Profiling**. *PLoS ONE* **3**(2).

Figures

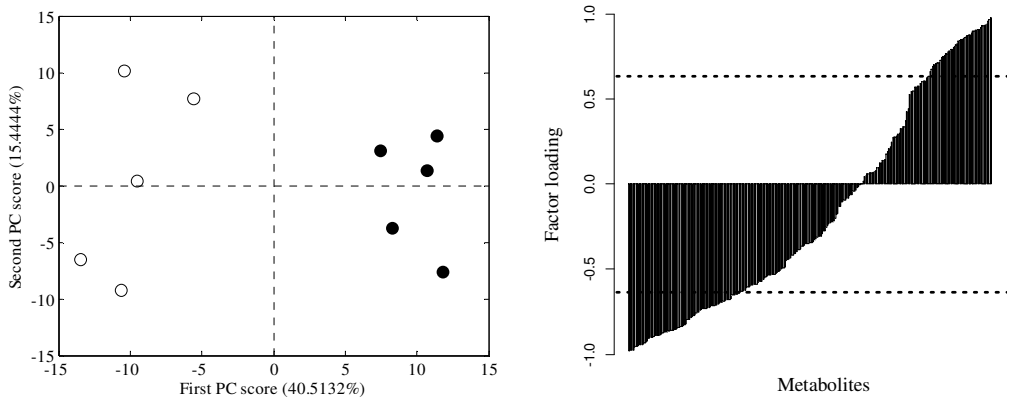

Fig. 1. Result of PCA in a comparative study of normal and 12-h fasting mice. (A) Score plot of PC1 and PC2. Symbols: (○) control mouse; (●) 12-h fasting mouse. (B) Factor loading plot in PC1. Metabolites are sorted in ascending order of the value of factor loading. The dotted line shows the significant level under $p<0.05$.

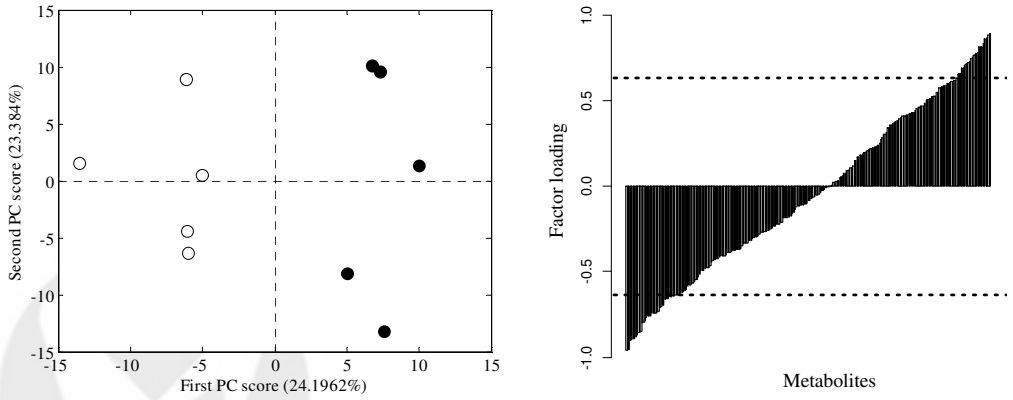

Fig. 2. Result of PCA in a comparative study of diabetes model mice with and without administered pioglitazone. (A) Score plot of PC1 and PC2. Symbols: (○) diabetes model mouse (*db*/*db* mouse) without administered pioglitazone; (●) *db*/*db* mouse with administered pioglitazone. (B) Factor loading plot in PC1. Metabolites are sorted in ascending order of the value of factor loading. The

dotted line shows the significant level under $p<0.05$. The dotted line shows significant level under $p<0.05$.

Table 1. Results of MSEA by ORA and Subramanian's approach in a comparative study of normal and 12-h fasting mice.

	ORA						Subramanian's approach				
	positive correlation with PC1			negative correlation with PC1			positive correlation with PC1				
	p -value		*q* -value	*p* -value		*q* -value		NES	*p* -value		*q* -value	
Glycolysis	1.0000		1.0000	0.0001	*	0.0036	**	-2.0048	0.0000	*	0.0074	**
Pentose phosphate pathway	1.0000		1.0000	0.0308	*	0.2000		-1.6040	0.0283	*	0.1781	
TCA cycle	1.0000		1.0000	0.0040	*	0.0519		-1.6208	0.0165	*	0.2433	
Glutamic acid and glutamine metabolism	1.0000		1.0000	0.4901		0.9801		-1.0936	0.3497		0.5193	
Alanine, aspartic acid and asparagine metabolism	0.8254		1.0000	0.2878		0.8313		-1.1897	0.2625		0.4379	
Lysine metabolism	0.8567		1.0000	0.8681		1.0000		-0.8172	0.7078		0.8274	
Valine, leucine and isoleucine metabolism	1.0000		1.0000	0.7735		1.0000		-0.9392	0.5636		0.7311	
Glycine, serine and threonine metabolism	0.7434		1.0000	0.0720		0.2445		-1.2557	0.1803		0.3704	
Cysteine metabolism	0.6489		1.0000	0.0412	*	0.2142		-1.3259	0.1611		0.3371	
Methionine metabolism	0.6178		1.0000	0.8444		1.0000		0.8697	0.6147		0.7197	
Shikimic acid metabolism	1.0000		1.0000	0.4901		0.9801		-1.2676	0.1745		0.3901	
Histidine metabolism	0.5434		1.0000	1.0000		1.0000		1.8978	0.0080	*	0.0520	
Urea cycle	0.8567		1.0000	0.0507		0.2197		-1.3352	0.1524		0.3705	
Proline metabolism	1.0000		1.0000	0.6001		1.0000		-1.1524	0.3041		0.4619	
Polyamine metabolism	1.0000		1.0000	0.0308	*	0.2000		-1.5785	0.0309	*	0.1581	
Tryptophan metabolism	0.7413		1.0000	0.9269		1.0000		-0.7141	0.8260		0.8764	
Tyrosine metabolism	1.0000		1.0000	0.0752		0.2445		-1.3601	0.1176		0.3820	
beta-alanine metabolism	0.2111		1.0000	0.8616		1.0000		0.9218	0.5531		0.7578	
Taurine metabolism	1.0000		1.0000	0.3721		0.9675		-1.4114	0.1010		0.3535	
Creatine metabolism	0.7874		1.0000	0.4705		0.9801		-0.8041	0.7093		0.7984	
Purine metabolism	0.0285	*	0.7411	0.9983		1.0000		1.6391	0.0220	*	0.1290	
Pyrimidine metabolism	0.9649		1.0000	0.9860		1.0000		0.7965	0.7355		0.7258	
Ribonucleotide metabolism	0.3473		1.0000	1.0000		1.0000		1.1184	0.2800		0.6998	
Deoxyribonucleotide	1.0000		1.0000	1.0000		1.0000		-0.6743	0.9776		0.8725	
Conjugated bile acid	0.5361		1.0000	1.0000		1.0000		1.0384	0.3671		0.6834	
Nicotinic acid metaboilsm	0.3473		1.0000	0.5357		0.9949		-0.8360	0.6536		0.8510	

*p <0.05, **q<0.05,

Table2. Results of MSEA by ORA and Subramanian's approach in a comparative study of diabetes model mice with and without administered pioglitazone.

	ORA					Subramanian's approach		
	positive correlation with PC1			negative correlation with PC1		positive correlation with PC1		
	p-value		q-value	p-value	q-value	NES	p-value	q-value
Glycolysis	0.0090	*	0.2250	0.7982	1.0000	1.6888	0.0198 *	0.3485
Pentose phosphate pathway	1.0000		1.0000	1.0000	1.0000	1.3813	0.1378	0.9520
TCA cycle	1.0000		1.0000	1.0000	1.0000	-1.4237	0.0916	0.8261
Glutamic acid and glutamine metabolism	1.0000		1.0000	0.5471	1.0000	-1.2740	0.2062	0.4203
Alanine, aspartic acid and asparagine metabolism	0.5531		1.0000	1.0000	1.0000	0.7161	0.8049	1.0000
Lysine metabolism	1.0000		1.0000	1.0000	1.0000	-0.6892	0.8152	0.9603
Valine, leucine and isoleucine metabolism	0.3294		1.0000	1.0000	1.0000	0.8275	0.6150	0.9834
Glycine, serine and threonine metabolism	0.1405		0.7024	0.8617	1.0000	1.1241	0.2699	0.9823
Cysteine metabolism	0.7041		1.0000	0.2461	1.0000	-0.9727	0.4840	0.8845
Methionine metabolism	1.0000		1.0000	1.0000	1.0000	1.0388	0.4167	0.9189
Shikimic acid metabolism	0.3294		1.0000	1.0000	1.0000	1.1493	0.3098	1.0000
Histidine metabolism	1.0000		1.0000	1.0000	1.0000	0.7463	0.7770	1.0000
Urea cycle	1.0000		1.0000	1.0000	1.0000	0.6208	0.9293	0.9238
Proline metabolism	0.5051		1.0000	1.0000	1.0000	0.6493	0.8869	1.0000
Polyamine metabolism	0.1344		0.7024	0.2701	1.0000	1.0695	0.3818	0.9813
Tryptophan metabolism	0.1018		0.7024	1.0000	1.0000	1.2380	0.2247	1.0000
Tyrosine metabolism	0.4521		1.0000	1.0000	1.0000	0.9319	0.5367	0.8548
beta-alanine metabolism	0.3893		1.0000	0.6321	1.0000	0.6450	0.8899	0.9596
Taurine metabolism	0.0577		0.7024	0.4410	1.0000	0.9952	0.4654	0.8048
Creatine metabolism	1.0000		1.0000	0.7206	1.0000	-0.7887	0.7230	0.9414
Purine metabolism	1.0000		1.0000	0.2583	1.0000	-1.3788	0.0947	0.5203
Pyrimidine metabolism	1.0000		1.0000	0.9252	1.0000	-0.7940	0.7196	1.0000
Ribonucleotide metabolism	1.0000		1.0000	0.2461	1.0000	-1.3605	0.1215	0.3792
Conjugated bile acid	1.0000		1.0000	0.4687	1.0000	-0.5472	0.9689	0.9709
Nicotinic acid metaboilsm	0.3161		1.0000	0.5431	1.0000	0.9991	0.4427	0.8973

*p <0.05, **q<0.05,