Metabolomics (2010) 6:137–145
DOI 10.1007/s11306-009-0182-2

ORIGINAL ARTICLE

Rice-*Arabidopsis* FOX line screening with FT-NIR-based fingerprinting for GC-TOF/MS-based metabolite profiling

Makoto Suzuki · Miyako Kusano · Hideki Takahashi · Yumiko Nakamura · Naomi Hayashi · Makoto Kobayashi · Takanari Ichikawa · Minami Matsui · Hirohiko Hirochika · Kazuki Saito

Received: 30 June 2009 / Accepted: 5 October 2009 / Published online: 17 October 2009

Abstract The full-length cDNA over-expressing (FOX) gene hunting system is useful for genome-wide gain-of-function analysis. The screening of FOX lines requires a high-throughput metabolomic method that can detect a wide range of metabolites. Fourier transform-near-infrared (FT-NIR) spectroscopy in combination with the chemometric approach has been used to analyze metabolite fingerprints. Since FT-NIR spectroscopy can be used to analyze a solid sample without destructive extraction, this technique enables untargeted analysis and high-throughput screening focusing on the alteration of metabolite composition. We performed non-destructive FT-NIR-based fingerprinting to screen seed samples of 3000 rice-*Arabidopsis* FOX lines; the samples were obtained from transgenic *Arabidopsis thaliana* lines that overexpressed rice full-length cDNA. Subsequently, the candidate lines exhibiting alteration in their metabolite fingerprints were analyzed by gas chromatography-time-of-flight/mass spectrometry (GC-TOF/MS) in order to assess their metabolite profiles. Finally, multivariate regression using orthogonal projections to latent structures (O2PLS) was used to elucidate the predictive metabolites obtained in FT-NIR analysis by integration of the datasets obtained from FT-NIR and GC-TOF/MS analyses. FT-NIR-based fingerprinting is a technically efficient method in that it facilitates non-destructive analysis in a high-throughput manner. Furthermore, with the integrated analysis used here, we were able to discover unique metabotypes in rice-*Arabidopsis* FOX lines; thus, this approach is beneficial for investigating the function of rice genes related to metabolism.

Keywords FT-NIR · GC-TOF/MS · FOX hunting system · Rice · *Arabidopsis* · Metabolomics · O2PLS

1 Introduction

Rice (*Oryza sativa*) is one of the most important crops in the world. The rice genome has been sequenced and found to comprise more than 32,000 genes (Gojobori 2007). The future challenge in this regard lies in identifying the biological functions of these genes. Rice full-length cDNAs have been collected for comprehensive analysis of the functions of rice genes (Kikuchi et al. 2003). The full-length cDNA over-expressing (FOX) gene hunting system, which employs a gain-of-function approach, has been used for investigating gene function (Ichikawa et al. 2006; Kondou et al. 2009). When the FOX hunting system is applied to individual transgenic plants, the dominant phenotype of a mutation is caused by the overexpression of the full-length cDNA introduced. This technique can be applied to the identification and characterization of rice

Electronic supplementary material The online version of this article (doi:10.1007/s11306-009-0182-2) contains supplementary material, which is available to authorized users.

M. Suzuki · M. Kusano · H. Takahashi · Y. Nakamura · N. Hayashi · M. Kobayashi · T. Ichikawa · M. Matsui · K. Saito (✉)
RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan
e-mail: ksaito@psc.riken.jp

M. Suzuki · K. Saito
Graduate School of Pharmaceutical Sciences, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan

H. Hirochika
National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba 305-8602, Japan

genes using *Arabidopsis thaliana* as the heterologous host of transgenes. The overexpression of rice genes results in not only visible phenotypes but also invisible phenotypes, including remarkable alterations in the metabolite composition (metabotype) (Hall 2006; Raamsdonk et al. 2001).

Many analytical technologies based on gas chromatography-mass spectrometry (GC-MS), liquid chromatography (LC)-MS, capillary electrophoresis (CE)-MS, nuclear magnetic resonance (NMR) spectroscopy, or fourier transform-infrared (FT-IR) spectroscopy have been developed for evaluating the metabotypes (Allwood et al. 2006; Bauer et al. 2006; Fiehn et al. 2000; Grata et al. 2008; Sato et al. 2004; Ward et al. 2003; Yang and Yen 2002). Since MS-based techniques have high selectivity and sensitivity for the identification and quantification of metabolites, they have been extensively used for metabolite profiling. On the other hand, NMR and FT-IR spectroscopy have low selectivity but can be used to discriminate between biological samples on the basis of differences in their metabolite composition. Therefore, these techniques are often used for metabolite fingerprinting (Ellis et al. 2007; Fiehn 2002). However, these techniques have limited dynamic range, and they are time consuming to prepare samples for analysis.

Recently, Fourier transform-near-infrared (FT-NIR) spectroscopy has been widely used for quality assessment of industrial materials and natural products, owing to its simplicity and rapidness (Hall and Pollard 1992; Ikeda et al. 2007; Rodriguez Otero et al. 1997). Unlike a typical NMR and FT-IR techniques, FT-NIR spectroscopic analysis does not entail destructive preparation steps such as homogenization or extraction using organic solvent, thereby enabling the use of large sets of seed samples and a greater number of assays with the same sample. Absorption in the NIR spectral region (4000–10000 cm^{-1}) allows the detection of overtones and combinations of the fundamental vibrations derived from the stretching and bending of NH, OH, and CH groups (Weyer and Lo 2002; Workman 2000), while absorption in the mid-IR spectral region (400–4000 cm^{-1}) can detect the fundamental vibrations of organic molecules, which enables to classify the differences in macromolecule composition such as cell wall structures (Mouille et al. 2003). FT-NIR can be used to determine metabolite levels such as the content of amino acid and fatty acid composition in seed sample (Kovalenko et al. 2006; Sato et al. 1998). Therefore, FT-NIR spectra indicate the metabolite composition that is available for metabolite fingerprinting (Munck et al. 2001).

The metabolite fingerprint dataset was statistically analyzed by chemometric approaches, including principal component analysis (PCA) and orthogonal projection to latent structures-discriminant analysis (OPLS-DA), for evaluating biological alteration (Pohjanen et al. 2007). In addition, a recent study employed multivariate regression using orthogonal projections to latent structures (O2PLS) to a combination of "omics" datasets (Bylesjo et al. 2007; Bylesjo et al. 2009; Rantalainen et al. 2006). This methodology can be used to extract the joint variation from different analytical platforms for the interpretation of complex biological process. In this context, O2PLS can be applied to the integration of FT-NIR datasets and other omics datasets such as those obtained from gas chromatography-time-of-flight/mass spectrometry (GC-TOF/MS) to elucidate the association between FT-NIR spectral data and the metabolite levels observed in GC-TOF/MS analysis.

In this study, we developed a non-destructive analytical method using FT-NIR spectroscopy for screening seed samples of rice-*Arabidopsis* FOX lines; these samples were obtained from transgenic *A. thaliana* lines that overexpressed rice full-length cDNA. Next, re-transformants of candidate genes were analyzed using OPLS-DA for clearly assessing alteration in metabolite fingerprints. Moreover, GC-TOF/MS was used to confirm the change in metabolite profiles. Finally, the predictive metabolites obtained in FT-NIR analysis were studied in greater detail by applying the O2PLS method.

2 Materials and methods

2.1 Plant material

For the evaluation of discrimination abilities, seed samples of five various *Arabidopsis* ecotypes (Col-0, Ws, Ler, Nossen, and C24) were used. *Arabidopsis* transgenic lines expressing rice full-length cDNA (rice FOX lines) under the control of the *CaMV 35S* promoter were constructed using the *Agrobacterium tumefaciens in planta* transformation method (Clough and Bent 1998). We screened seed samples of the T2 generation that did not exhibit any visible phenotypes. A total of 3003 lines (74 batches) were analyzed by FT-NIR spectroscopy. FOX lines in each batch comprised samples that were maximum 51 lines harvested during the same growth period with a cultivation container system (Arasystem, Gent, Belgium). To diminish the environmental effect, we analyzed FOX seeds of each batch separately. Subsequently, genomic DNA of each candidate line exhibiting alteration of metabolite fingerprints was isolated from the corresponding seed sample. Then, rice cDNA was sequenced for the annotation of gene function. For annotation of the inserted rice cDNA, the sequences of the candidate genes were analyzed on the basis of information available in the Knowledge-based

Oryza Molecular biological Encyclopedia (KOME) and the Rice Annotation Project (RAP) databases. The re-transformants (five individual transgenic plants for each candidate) were constructed to confirm the reproducibility of the metabotype for each line. The details of the method have been described by Kondou et al (2009). *Arabidopsis* seeds (Col-0), which were harvested at the same time as the re-transformants, were used as the control. A total of 26 re-transformants (127 samples) were analyzed by FT-NIR spectroscopy. Among them, seven lines (34 samples) exhibited changes in their metabolite fingerprints; these were analyzed by GC-TOF/MS.

2.2 FT-NIR analysis

For the screening of FOX lines, 200 seeds were placed in a 0.3 ml glass tube for each line, and the FT-NIR spectra of the seeds of each line were directly measured six times. FT-NIR spectra were measured with a Nicolet 6700 FT-IR equipped with a Smart Near-IR UpDRIFT, CaF_2 beamsplitter, and cooled InGaAs detector (Thermo Electron Corporation, Madison, USA). The mirror velocity at a resolution of 4 cm^{-1} was 1.2659 cm/s. The total number of data points in the range of 4500–7500 cm^{-1} was 1556 for each spectrum. The diffuse reflectance spectrum was obtained by ratioing the single beam spectrum against the background spectrum using spectralon (LabSphere, Inc.). Each spectrum was recorded as an average of 32 scans using OMNIC 7.2a (Thermo Electron Corporation, Madison, USA). Sample information and the raw spectral dataset of rice-*Arabidopsis* FOX seeds are available through Rice FOX database (http://ricefox.psc.riken.jp/). The obtained FT-NIR spectra were transformed with multiplicative signal correction (MSC) (Geladi et al. 1985) to minimize the variations in sample path length that are caused by the light scatter effect resulting from the differences in individual seed shape. Then, overlapping absorption peaks were clarified with the 25-point polynomial-fit Savitzky-Golay second derivation (Savitzky and Golay 1964).

2.3 GC-TOF/MS analysis

To assess the metabolite profiles of the candidate lines that showed altered their metabolite fingerprints in FT-NIR analysis, 200 seeds of each of the candidate lines were extracted at a concentration of 10 mg/ml, derivatized, and then analyzed by GC-TOF/MS as described in Kusano et al (2007). A total of 266 metabolite peaks were extracted for each seed sample. Of them, 67 peaks were identified or annotated as known metabolites, 186 peaks were of unknown metabolites, and 13 peaks were annotated as mass spectral tags (MSTs) (Schauer et al. 2005).

2.4 Statistical analysis

Before multivariate analysis, the corrected FT-NIR spectral datasets were mean-centered and the GC-TOF/MS dataset were scaled to unit variance following $\log_{10}$-transformation. The multivariate models were calculated using PCA, OPLS-DA, and O2PLS implemented in SIMCA-P + version 12 (Umetrics AB, Umeå, Sweden). The ellipse in the PC score plot represents the confidence region of the model based on Hotteling's T^2 statistic (Hotelling 1931; Mason et al. 2001). The significance level of the confidence region was defined at 0.05, and the data that fell outside the ellipse were determined to belong to candidate lines. These models were validated using 7-fold cross-validation or analysis of variance of cross-validated predictive residuals (CV-ANOVA) (Eriksson et al. 2008). Cross-validation is an internal predictive validation method for determining the number of significant components by calculating the total amount of explained X-variance (R^2X), Y-variance (R^2Y), and cross-validated predictive ability (Q^2Y). A component is significant when Q^2Y is positive value. Additionally, the variance related to class separation (R^2_PX) was calculated by OPLS-DA. CV-ANOVA is based on an ANOVA assessment of the cross-validatory predictive residuals of the models. The statistical Welch's *t* test was performed and false discovery rate (FDR), which have been proven to be reliable for determining the significance of multiple testing (Storey 2002), were calculated using Microsoft Office Excel 2003 software. *Q*-value for FDR less than 0.05 was regarded as significant.

3 Results and discussion

3.1 Metabolite fingerprinting of seeds of various *Arabidopsis* ecotypes by FT-NIR spectroscopy

To evaluate discrimination abilities of metabolite fingerprints of *Arabidopsis* seeds by FT-NIR spectroscopy, five various *Arabidopsis* ecotypes were analyzed. The FT-NIR spectral data of the five ecotype seeds showed broad peaks and extensive overlapping of NIR absorption bands derived from complex chemical components in the sample (Fig. 1a). For elimination of baseline shift and enhancement of shoulder peaks, MSC and second derivation were applied to the spectral dataset (Fig. 1b). After spectral correction, multivariate analysis was performed to evaluate the corrected spectra. PCA was performed to visualize the strongest varying components of the spectra obtained for various *Arabidopsis* ecotypes whose metabolite fingerprints were altered. In the case of *Arabidopsis* seeds, each sample was distributed according to the ecotypes showing different metabolite fingerprints in the PC score scatter plot

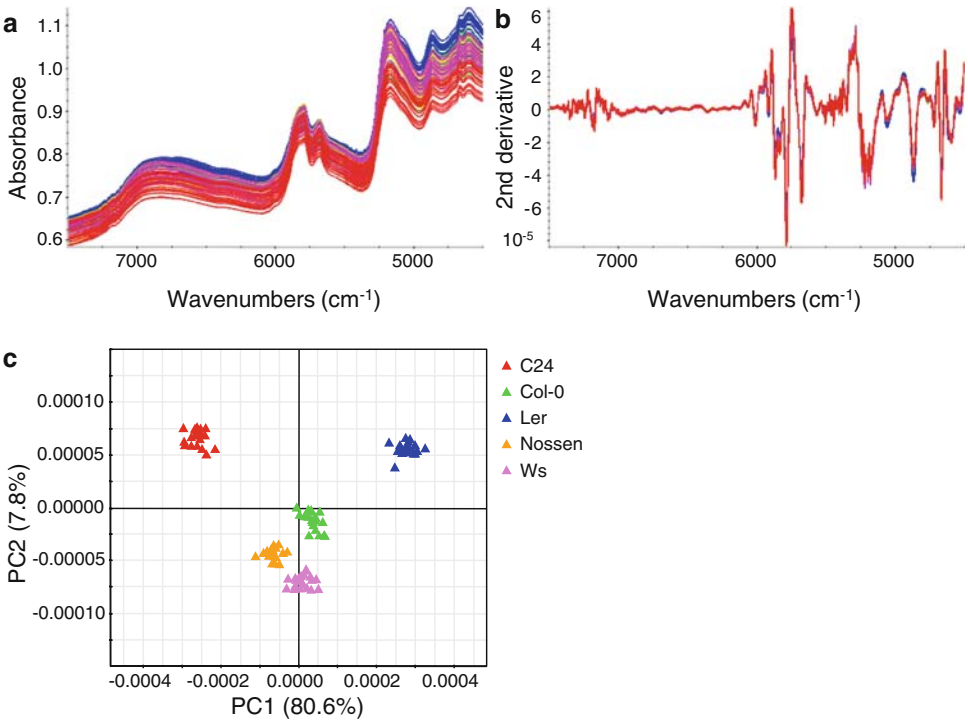

Fig. 1 Typical FT-NIR spectra of seed samples of five *Arabidopsis* ecotypes. Two hundred seeds were placed in a glass tube, and the FT-NIR spectra were measured 20 times. **a** Raw spectral data of the seed samples. **b** The corrected spectra by MSC and 2nd derivation. **c** Discrimination abilities of metabolite fingerprints of *Arabidopsis* seeds. The plot of the principal component 1 (PC1) versus principal component 2 (PC2) is presented. Each *colored symbol* represents an ecotype

(Fig. 1c). Furthermore, the acquisition time per analysis (30 s) in FT-NIR spectroscopy is short; therefore, this method is beneficial for large-scale screening.

3.2 Screening of seeds of rice-*Arabidopsis* FOX lines by FT-NIR spectroscopy

In order to screen rice-*Arabidopsis* FOX lines that show specific metabolite fingerprints (i.e., metabotype) in seeds, FOX seeds were analyzed using FT-NIR spectroscopy. In this study, FOX seeds in the T2 generation were used. The corrected spectra of these lines were applied for PCA to filter candidate lines.

The candidate lines showed different distribution patterns when compared with the major distribution patterns of the other lines in the PC score scatter plot (Fig. 2). Using FT-NIR-based metabolite fingerprinting, 3,003 FOX lines were analyzed. From the result of the analysis, 30 lines that showed altered metabolite fingerprints were selected as the candidate lines. Among them, 26 lines—in which the rice full-length cDNA was correctly inserted—were used for further analysis.

3.3 Assessment of the metabolite fingerprints of re-transformants

We assessed the changes in the metabolite fingerprints of re-transformants harboring rice full-length cDNA; 26 candidate lines fell in this category. Their metabolite fingerprints were compared with those of the wild type by FT-NIR. For clearly assessing the effect of candidate genes, the differences between metabolite fingerprints were confirmed by OPLS-DA. OPLS-DA uses information on categorical response Y (wild type or the re-transformants) to decompose spectral data into a predictive matrix related to biological alteration and the Y-orthogonal matrix (Pohjanen et al. 2007). This strategy allows for a more realistic interpretation of metabolite fingerprints than well-known method such as liner discriminant analysis.

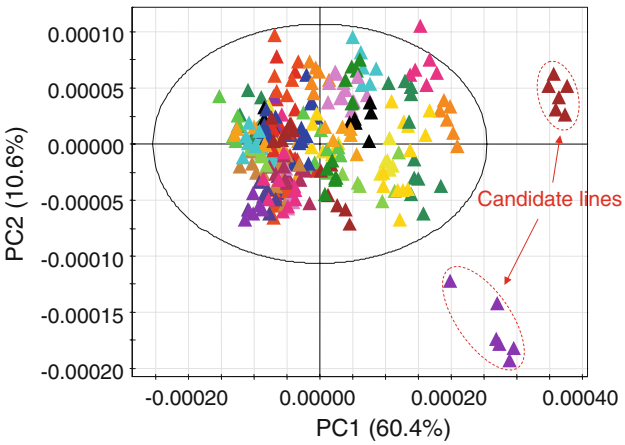

Fig. 2 PC score plot of the FT-NIR spectral dataset for the first batch containing 51 lines. The plot of the principal component 1 (PC1) versus principal component 2 (PC2) is presented. Each *colored symbol* represents an individual line. The *ellipse* represents the confidence region of the model based on Hotteling's T^2 statistic ($\alpha = 0.05$)

Among the 26 candidate lines, the seven lines listed in Table 1 showed differences in their metabolite fingerprints compared with the wild type without overfitting in each OPLS-DA model (Fig. 3). The other 19 lines showed no significant difference with regard to discrimination from the wild type. The loading spectra of the predictive component shown in Fig. 4 indicate the importance of absorption bands for the discrimination of the re-transformants from the wild type. The absorption band from 6950 to 7400 cm^{-1} has been attributed to a combination of the first overtones of the C–H stretch. The first overtone of the C–H stretch was located from 5600 to 6150 cm^{-1}. The absorption band at around 4850 cm^{-1} corresponded to combinations of O–H or N–H. The band near 5200 cm^{-1} has been assigned to the second overtone of the C=O stretch. The combination of the C–H stretch and C–C stretch derived from the benzene moiety was located at around 4675 cm^{-1}. Other minor absorption bands were found to overlap in each spectral region (Weyer and Lo 2002; Workman 2000). The shape of the loading spectra showed a specific pattern for each candidate line. This result suggests that the overexpression of the seven rice genes that were introduced into *A. thaliana* have various effects on their metabolite fingerprints, respectively.

3.4 Assessment of metabolite profiles by GC-TOF/MS

For the assessment of rice gene functions, it is necessary to further analyze the metabolite profiles obtained. To identify the metabolites whose levels differed among the re-transformant lines, we analyzed seven candidate lines by GC-TOF/MS. OPLS-DA was carried out to classify and interpret the metabolite profiles. Samples that were misclassified in the OPLS-DA model were excluded for clear assessment of metabolite profiles. Candidates 1–6 could be clearly discriminated from the wild type, but Candidate 7 was overfitted in the OPLS-DA model (supplementary Fig. 1). The loading weights of the OPLS-DA model, fold change, and *p*-value of the *t* test for significantly altered metabolites in each candidate line are listed in Table 2. For Candidate 6, the known metabolites did not show any alteration, but 13 unknown compounds were found to have significantly altered (data not shown).

The changes in the metabolite profiles were unique for each candidate line. In relation to Candidate 1, it has been reported that T-DNA insertional *Arabidopsis* mutants of the *ETFβ* gene showed significant accumulation of several amino acids, isovaleryl CoA, and phytanoyl CoA during dark-induced carbohydrate deprivation (Ishizaki et al. 2006). It is expected that the changes in the metabolite profile of Candidate 1 are influenced by the function of ETF. In Candidate 3, resistance to *Pseudomonas syringae DC3000* was confirmed by Mori et al in other study (Kondou et al. 2009). Further metabolomic analyses of different tissues at different stages of growth and under different stress conditions would enable advanced investigation of rice gene functions.

3.5 Relationships between absorption in the FT-NIR spectra and the metabolite levels determined by GC-TOF/MS

The changes in metabolite fingerprints were specific for each candidate line, and information about individual metabolites could have been obtained using the FT-NIR spectra. Here, however, we used the O2PLS multivariate regression method to identify the predictive metabolites in FT-NIR analysis. Spectroscopic and chromatographic techniques entail systematic variations such as baseline shift and background noise. With O2PLS, such irrelevant variations can be removed, and joint variation related to biological alteration can be extracted from metabolomics

Table 1 Candidate lines exhibiting alteration of metabolite fingerprints

Line	cDNA annotation	Accession number in KOME	Accession number in RAP	KOGs
Candidate 1	Electron transfer flavoprotein alpha-subunit, mitochondrial precursor (Alpha-ETF)	J013093P06	Os03g0835400	Energy production and conversion
Candidate 2	Sodium-dicarboxylate cotransporter-like	J013123J03	Os09g0484900	Inorganic ion transport and metabolism
Candidate 3	SufBD family protein	J013162I04	Os01g0127300	No similarity sequence
Candidate 4	Conserved hypothetical protein	J033068G21	Os02g0224900	No similarity sequence
Candidate 5	VHS domain-containing protein	J013105N04	Os08g0109000	Intracellular trafficking, secretion, and vesicular transport
Candidate 6	Aldo/keto reductase family protein	J033132F18	Os04g0339400	Energy production and conversion
Candidate 7	Conserved hypothetical protein	J023079A07	Os03g0197000	Lipid transport and metabolism

The sequence of the inserted full-length cDNA and other information on it was obtained from the Knowledge-based Oryza Molecular biological Encyclopedia (KOME) and Rice Annotation Project (RAP) databases. The functions of the candidate genes were assigned to eukaryotic clusters of eukaryotic orthologous groups (KOGs)

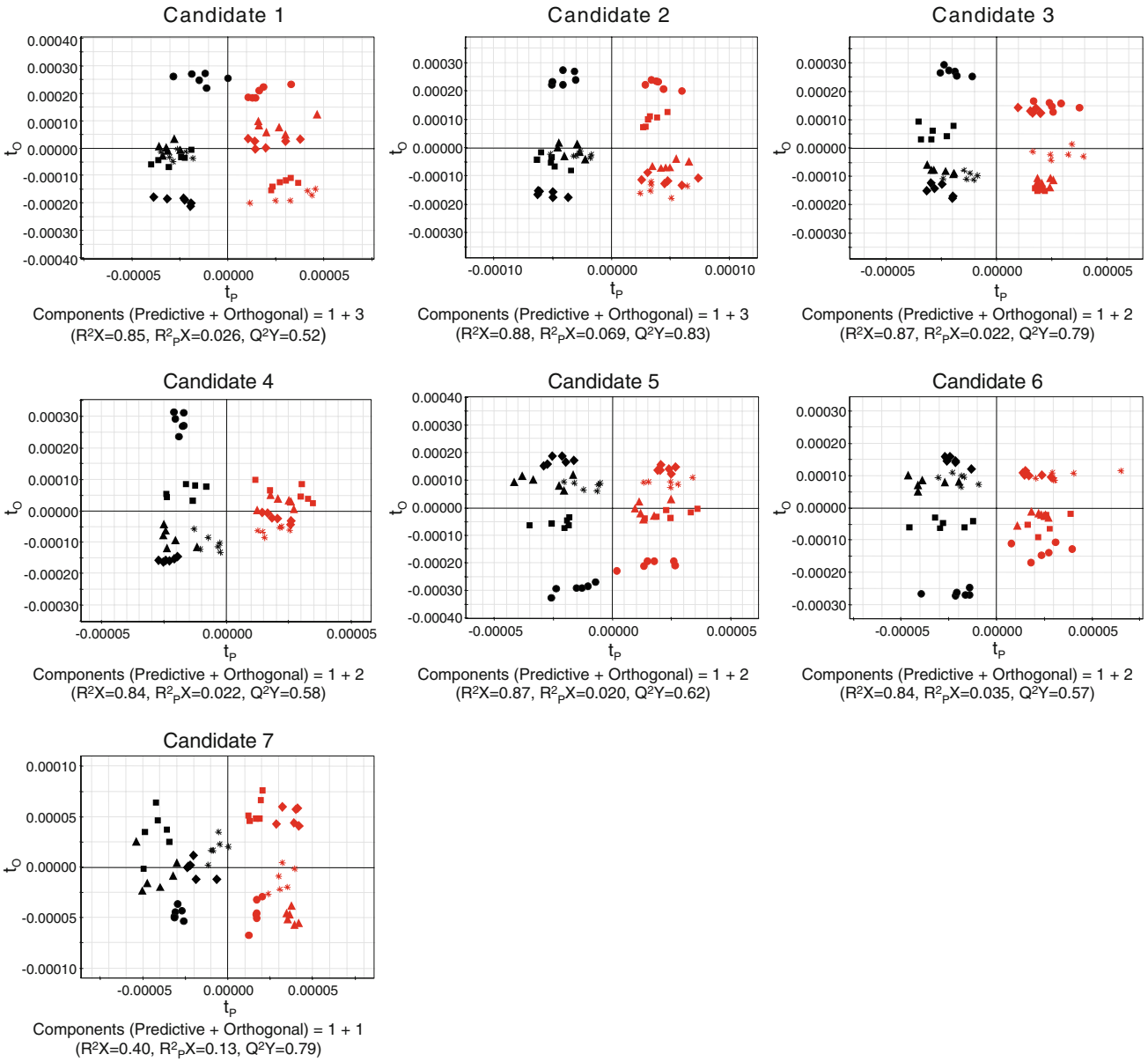

Fig. 3 Discrimination of re-transformants using OPLS-DA based on FT-NIR analysis. OPLS-DA was performed using the FT-NIR spectral dataset for each re-transformant. The plot of predictive component (t_P) versus orthogonal component 1 (t_O) is presented. The *black symbols* represent the wild type, while the *red symbols* represent re-transformants. Individual transgenic lines are represented by different symbols. Seven re-transformant lines could be clearly discriminated from the wild type

datasets (Bylesjo et al. 2007; Bylesjo et al. 2009; Rantalainen et al. 2006). Joint variation obtained from FT-NIR spectra and GC-TOF/MS datasets of known metabolites is useful for understanding the relationships between absorption in the FT-NIR spectra and metabolite levels. Here, the dataset of Candidates 1–7 and the wild type were used for constructing the model.

The O2PLS model was constructed with three predictive components that account for 81% of the total variation in the FT-NIR dataset and 21.1% of the variation in the GC-TOF/MS dataset. Moreover, we identified one orthogonal component (10.9%) in the FT-NIR dataset that was not present in the GC-TOF/MS dataset and one unrelated component (12.4%) in the GC-TOF/MS dataset that was not available in the FT-NIR dataset. To filter predictive metabolites in the joint variation of the O2PLS model, CV-ANOVA was used as the significance test. In addition, the *q*-values for FDR were calculated using the *p*-values of CV-ANOVA. The threshold of *q*-value for significantly predictive metabolites was defined at 0.05. We found 21 metabolites to be predictive in the O2PLS model (supplementary Table 1). The joint variation of predictive

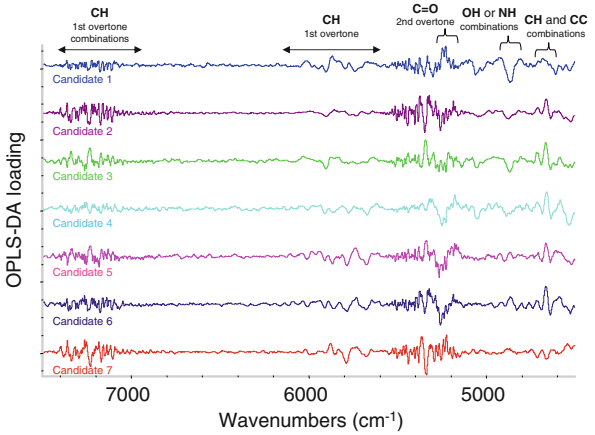

Fig. 4 Importance of absorption bands for discriminating the re-transformants from the wild type in OPLS-DA. Each absorption band indicates overtones and combinations of the fundamental vibrations of organic molecules. Loading of the predictive component in OPLS-DA showed that the absorption bands were important for discriminating the re-transformants from the wild type. The shape of the loading spectra indicated the various effects of the introduced rice genes, which brought about the change in metabolite fingerprints

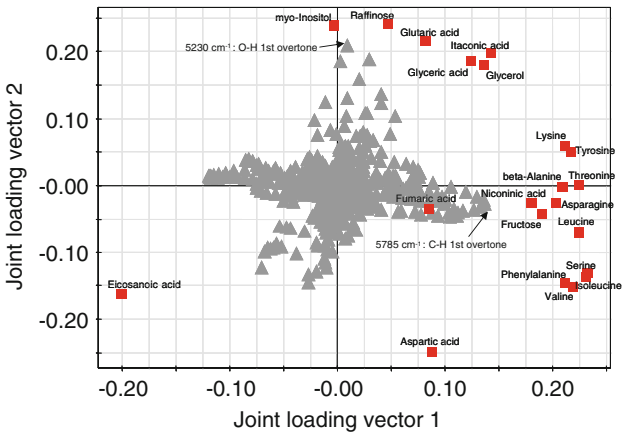

Fig. 5 Overview of the relationships between the absorption bands of FT-NIR spectra and predictive metabolites. The O2PLS loading plot obtained by integration of the FT-NIR and GC-TOF/MS datasets for Candidates 1–7 and the wild type is shown here. Loading of the FT-NIR and GC-TOF/MS datasets was concatenated to one vector (joint loading vector). The *red symbols* represent predictive metabolites, and the *gray symbols*, wave numbers in the FT-NIR spectra. Variables that are near each other are positively correlated, and those situated opposite are negatively correlated

Table 2 Significant changes in the levels of known metabolites, as determined by GC-TOF/MS analysis

	Loading weight	*X*-fold (/Col-0)[a]	*P*-value (/Col-0)[*]
Candidate 1			
Hydroxylamine	–0.116	0.781	0.004
n-Butylamine	–0.110	0.947	0.037
3-Amino-piperidin-2-one	0.101	1.332	0.045
Candidate 2			
3-Amino-piperidin-2-one	0.094	1.292	0.032
Phytol	0.082	1.295	0.040
Candidate 3			
Glycine	–0.096	0.683	0.009
Tryptamine	0.104	1.806	0.009
Galactinol	0.089	1.313	0.023
Phytol	–0.083	0.855	0.038
Malic acid	–0.100	0.480	0.038
Glutamine	–0.115	0.746	0.043
Fumaric acid	–0.102	0.228	0.047
Candidate 4			
Glutamine	–0.105	0.727	0.034
Glycine	–0.118	0.633	0.036
Candidate 5			
Glycine	–0.127	0.587	0.011
Lysine	–0.121	0.783	0.028
Threitol	–0.121	0.666	0.038
Candidate 6			
No significant metabolite			

[*] *P*-value was calculated by Welch's *t* test ($\alpha = 0.05$)

[a] Fold changes in the metabolite levels compared with the wild type

metabolites explained the chemical relationship between FT-NIR spectroscopy and GC-TOF/MS. The O2PLS loading plot shown in Fig. 5 indicates how the absorption bands of the FT-NIR spectra relate to the predictive metabolites. The relative loading weights indicate the strength of relatedness. Predictive metabolites were clustered on the basis of similarities in their chemical structure; the clustering revealed that compounds with similar structure were predicted by similar absorption patterns in their FT-NIR spectra.

The variation related to specific metabolite changes which can consider to be caused by rice gene function was not extracted in the joint variation; however, it can be explained by unique variation in the GC-TOF/MS dataset (supplementary Fig. 1). On the other hand, it is also expected that the variation in the orthogonal component of the FT-NIR spectra can explain the other aspects of the metabotypes in each candidate line. These features enable the application of FT-NIR spectroscopy to the screening of a variety of metabolites.

4 Concluding remarks

We have developed a non-destructive screening method using FT-NIR spectroscopy for the analysis of seeds of rice-*Arabidopsis* FOX lines. This method is timesaving in that it can be used to detect the metabolite fingerprints of seed material without pretreatment. A simple and rapid method is required for the screening of rice-*Arabidopsis* FOX lines; thus, FT-NIR spectroscopy was suitable for this

research. Moreover, OPLS-DA based on the GC-TOF/MS dataset revealed the changes in metabolite profiles in greater detail. In addition, the O2PLS methodology provided additional information about predictive metabolites in the FT-NIR analysis. The advantage of FT-NIR spectroscopy is that it can be used to detect the composition of a variety of metabolites. FT-NIR spectroscopy combined with chemometrics can be used for large-scale screening of gain-of-function mutant resources. Moreover, this technology will be effective in the analysis of useful plant gene functions using loss-of-function mutants such as knock-out (TILLING or insertion mutation) or knock-down (RNAi) resources.

Acknowledgments We are grateful to Dr. Masaki Mori (National Institute of Agrobiological Sciences) for providing information of pathogen resistance study. We thank Dr. Henning Redestig for discussing the manuscript and helping us improve it; and Mr. Tetsuya Sakurai (RIKEN Plant Science Center) for the management of the FOX screening dataset. We also thank Ms. Makiko Takamune, Ms. Kazue Nakabayashi, and Ms. Yoko Suzuki (RIKEN Plant Science Center) for providing technical assistance. This work was supported by the Special Coordination Fund for Promoting Science and Technology (Japan Science and Technology Agency).

References

Allwood, J. W., Ellis, D. I., Heald, J. K., Goodacre, R., & Mur, L. A. (2006). Metabolomic approaches reveal that phosphatidic and phosphatidyl glycerol phospholipids are major discriminatory non-polar metabolites in responses by *Brachypodium distachyon* to challenge by *Magnaporthe grisea*. *Plant Journal, 46*, 351–368.

Bauer, S., Vasu, P., Persson, S., Mort, A. J., & Somerville, C. R. (2006). Development and application of a suite of polysaccharide-degrading enzymes for analyzing plant cell walls. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 11417–11422.

Bylesjo, M., Eriksson, D., Kusano, M., Moritz, T., & Trygg, J. (2007). Data integration in plant biology: The O2PLS method for combined modeling of transcript and metabolite data. *Plant Journal, 52*, 1181–1191.

Bylesjo, M., Nilsson, R., Srivastava, V., et al. (2009). Integrated analysis of transcript, protein and metabolite data to study lignin biosynthesis in hybrid aspen. *Journal of Proteome Research, 8*, 199–210.

Clough, S. J., & Bent, A. F. (1998). Floral dip: A simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant Journal, 16*, 735–743.

Ellis, D. I., Dunn, W. B., Griffin, J. L., Allwood, J. W., & Goodacre, R. (2007). Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics, 8*, 1243–1266.

Eriksson, L., Trygg, J., & Wold, S. (2008). CV-ANOVA for significance testing of PLS and OPLS® models. *Journal of Chemometrics, 22*, 594–600.

Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology, 48*, 155–171.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N., & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nature Biotechnology, 18*, 1157–1161.

Geladi, P., Macdougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy, 39*, 491–500.

Gojobori, T. (2007). Curated genome annotation of *Oryza sativa* ssp japonica and comparative genome analysis with *Arabidopsis thaliana*—The Rice Annotation Project. *Genome Research, 17*, 175–183.

Grata, E., Boccard, J., Guillarme, D., et al. (2008). UPLC-TOF-MS for plant metabolomics: A sequential approach for wound marker analysis in *Arabidopsis thaliana*. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences, 871*, 261–270.

Hall, R. D. (2006). Plant metabolomics: From holistic hope, to hype, to hot topic. *The New Phytologist, 169*, 453–468.

Hall, J. W., & Pollard, A. (1992). Near-infrared spectrophotometry: A new dimension in clinical chemistry. *Clinical Chemistry, 38*, 1623–1631.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics, 2*, 360–378.

Ichikawa, T., Nakazawa, M., Kawashima, M., et al. (2006). The FOX hunting system: An alternative gain-of-function gene hunting technique. *Plant Journal, 48*, 974–985.

Ikeda, T., Kanaya, S., Yonetani, T., Kobayashi, A., & Fukusaki, E. (2007). Prediction of Japanese green tea ranking by Fourier transform near-infrared reflectance spectroscopy. *Journal of Agricultural and Food Chemistry, 55*, 9908–9912.

Ishizaki, K., Schauer, N., Larson, T. R., Graham, I. A., Fernie, A. R., & Leaver, C. J. (2006). The mitochondrial electron transfer flavoprotein complex is essential for survival of *Arabidopsis* in extended darkness. *Plant Journal, 47*, 751–760.

Kikuchi, S., Satoh, K., Nagata, T., et al. (2003). Collection, mapping, and annotation of over 28, 000 cDNA clones from Japonica rice. *Science, 301*, 376–379.

Kondou, Y., Higuchi, M., Takahashi, S., et al. (2009). Systematic approaches to using the FOX hunting system to identify useful rice genes. *Plant Journal, 57*, 883–894.

Kovalenko, I. V., Rippke, G. R., & Hurburgh, C. R. (2006). Determination of amino acid composition of soybeans (Glycine max) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry, 54*, 3485–3491.

Kusano, M., Fukushima, A., Kobayashi, M., et al. (2007). Application of a metabolomic method combining one-dimensional and two-dimensional gas chromatography-time-of-flight/mass spectrometry to metabolic phenotyping of natural variants in rice. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences, 855*, 71–79.

Mason, R. L., Chou, Y. M., & Young, J. C. (2001). Applying Hotelling's T-2 statistic to batch processes. *Journal of Quality Technology, 33*, 466–479.

Mouille, G., Robin, S., Lecomte, M., Pagant, S., & Hofte, H. (2003). Classification and identification of *Arabidopsis* cell wall mutants using Fourier-Transform InfraRed (FT-IR) microspectroscopy. *Plant Journal, 35*, 393–404.

Munck, L., Nielsen, J. P., Moller, B., et al. (2001). Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Analytica Chimica Acta, 446*, 171–186.

Pohjanen, E., Thysell, E., Jonsson, P., et al. (2007). A multivariate screening strategy for investigating metabolic effects of strenuous physical exercise in human serum. *Journal of Proteome Research, 6*, 2113–2120.

Raamsdonk, L. M., Teusink, B., Broadhurst, D., et al. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology, 19*, 45–50.

Rantalainen, M., Cloarec, O., Beckonert, O., et al. (2006). Statistically integrated metabonomic-proteomic studies on a human prostate

cancer xenograft model in mice. *Journal of Proteome Research, 5*, 2642–2655.

Rodriguez Otero, J. L., Hermida, M., & Centeno, J. (1997). Analysis of dairy products by near-infrared spectroscopy: A review. *Journal of Agricultural and Food Chemistry, 45*, 2815–2819.

Sato, S., Soga, T., Nishioka, T., & Tomita, M. (2004). Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant Journal, 40*, 151–163.

Sato, T., Uezono, I., Morishita, T., & Tetsuka, T. (1998). Nondestructive estimation of fatty acid composition in seeds of *Brassica napus* L. by near-infrared spectroscopy. *Journal of American Oil Chemists' Society, 75*, 1877–1881.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*, 1627–1639.

Schauer, N., Steinhauser, D., Strelkov, S., et al. (2005). GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Letters, 579*, 1332–1337.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 64*, 479–498.

Ward, J. L., Harris, C., Lewis, J., & Beale, M. H. (2003). Assessment of H-1 NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry, 62*, 949–957.

Weyer, L. G., & Lo, S.-C. (2002). Spectra-structure correlations in the near-infrared. In J. Chalmers & P. Griffiths (Eds.), *Handbook of vibrational spectroscopy* (pp. 1817–1837). Chichester: Wiley.

Workman, J. (2000). NIR, IR, Raman, and UV-vis spectra featuring polymers and surfactants. In J. Workman (Ed.), *Handbook of organic compounds* (pp. 77–197). San Diego: Academic Press.

Yang, J., & Yen, H. E. (2002). Early salt stress effects on the changes in chemical composition in leaves of ice plant and *Arabidopsis*. A Fourier transform infrared spectroscopy study. *Plant Physiology, 130*, 1032–1042.