Different-batch metabolome analysis of *Saccharomyces cerevisiae* based on gas chromatography/mass spectrometry

Naoki KAWASE¹, Hiroshi TSUGAWA^{1,2}, Takeshi BAMBA¹, Eiichiro FUKUSAKI¹

¹ Department of Biotechnology, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

² RIKEN Center for Sustainable Resource Science, Tsurumi-ku, Yokohama 230-0045, Japan

Key words : GC/MS; metabolomics; *Saccharomyces cerevisiae*; different batches; time segmented SIM

Running title: Reproducible yeast metabolomics protocol

Correspondent footnote

E-mail address: fukusaki@bio.eng.osaka-u.ac.jp

Telephone and fax number: +81-(0)6-6879-7424

Abstract

Each experimental step in metabolomics based on mass spectrometry for microorganisms, such as cultivation, sampling, extraction of metabolites, analysis, and data processing includes different systematic errors. Even if the same protocol is used, it is difficult to compare the data from different cultivation days or different analysis days. To obtain reliable quantitative data, it is necessary to develop an analytical workflow that can reduce errors from different batch of cultivation and analysis days. We compared metabolomics methods for *Saccharomyces cerevisiae* in terms of reproducibility to optimize the analytical workflow, particularly quenching and data processing. Our data also showed that reproducible data could be obtained with high signal to noise ratio. Therefore, we optimized a time segmented SIM (Selective Ion Monitoring) method for high sensitive analysis with low- risk of false positives. The optimized workflow was applied to metabolome analysis of single transcription factor deletion mutants. As a result, we obtained clusters that were independent of cultivation day and analysis day but were strain-dependent. This study can help to implement large-scale or long-term studies, in which samples are divided among several laboratories because of the high number of samples.

1 Introduction

2 *Saccharomyces cerevisiae* is a model organism and an essential microorganism
3 for brewing and fermented foods. There are many reports about the metabolomics of
4 *Saccharomyces cerevisiae*. (e.g., function analysis of genes based on metabolome
5 information or screening of aging-related mutants)[1–3]. In these studies, projection of
6 latent structure is frequently used to predict biological phenotypes. When quantitative
7 metabolome data are applied to explanatory variables, the data including systematic
8 errors can cause over-fit problems and make the model difficult for practical use[3][4].
9 Moreover, there have been some efforts to establish a metabolome database, including
10 quantitative information for each organism[5][6]. Batch cultivations are generally
11 applied for the construction of metabolome databases for microorganisms because
12 continuous cultivations or feeding cultures that can precisely control cultivation
13 conditions are not practical, and they are cost-prohibitive to apply to many mutants.
14 However, data errors are more frequent in batch cultivation than in continuous or
15 feeding cultivations where cultivation conditions can be well controlled. This problem
16 makes it difficult to compare large amounts of data from different batch cultures.

17 Figure 1 shows the results when different cultivation days were applied to
18 principle component analysis (PCA). The samples formed clusters depending on their
19 cultivation day. Then, we tried to evaluate the reproducibility and reliability of each
20 procedure of yeast metabolomics based on batch culture and select repeatable methods.

21 In the process of yeast metabolomics, cultivation, quenching (which stops
22 metabolism), and analysis considerably contribute to data reproducibility. If the culture

1 temperature and cell density at sampling are kept constant, small differences in media
2 components, atmosphere temperature, and human error affect the data on different
3 experimental days. Therefore, the sensitivity of mass spectrometry changes daily, and
4 data from different analytical days might include different systematic errors. We
5 checked the reproducibility of some methods in each step over different experimental
6 days. For quenching, there have been reports about leakage of metabolites at the
7 quenching step with different intensities for several quenching methods[7][8]. However,
8 there have been no reports about the reproducibility of quenching methods over
9 different experimental days; therefore, we compared cold methanol quenching, fast
10 filtration, and quenching by liquid nitrogen. We also compared the data processing
11 methods for comparison of data from different batches. In this research, the following
12 three established methods were compared: internal standard normalization, external
13 standard normalization, and total intensity normalization.[9][10]

14 In this study, we found that sensitivity and peak intensity were important to
15 obtain a highly reproducible data profile. Therefore, to obtain more reproducible data,
16 we also optimized time-segmented single-ion monitoring (SIM) to obtain intensities
17 with a higher signal-to-noise (S/N) ratio[11]. In fact, the SIM method could decrease the
18 reliability of peak identification and increase the risk of false positive rates because SIM
19 methods limit the number of reference ions for peak assessment [12]. Although to detect
20 three diagnostic ions for each metabolite should be used according to the previous study,
21 the selection method or its reliability assessment with retention time information have
22 not been reported for metabolomics study in detail. According to our knowledge and the
23 previous report, it is important for accurate peak assessment to consider the mass
24 spectra's base peak and its retention time information. Therefore, we attempted to

construct a new method for selecting diagnostic SIM with bounds of base peak and retention time.

Lastly, we applied our protocol of yeast metabolomics to transcription factor deletion mutants to validate the reliability and robustness of our protocol. We chose eight different transcription factor single deletion mutants. In previous research, these strains were reported to have different ethanol tolerance[13], and we confirmed they had different metabolome features. Using our method, those features could be detected in merged data from different cultivation and analysis days and different GC/MS systems. Our protocol was shown to be reproducible and practical for yeast, and it is expected to be suitable for application to large-scale metabolomics research.

Materials and Methods

Strain and cultivation

Saccharomyces cerevisiae X2180 was used to compare the methods of yeast metabolomics. 8 *Saccharomyces cerevisiae* knockout mutant strains (*ace2Δ*, *bas1Δ*, *mat3Δ*, *rlm1Δ*, *rtg1Δ*, *skn7Δ*, *srb2Δ*, and *stp2Δ*) from the genome deletion project (EUROSCARF collection Frankfurt <http://web.uni-frankfurt.de/fb15/mikro/euroscarf/index.html>) were used to validate the protocol. The host strain was BY4772 (MAT α ; his3 Δ 1; leu2 Δ 0; lys2 Δ 0; ura3 Δ 0). X2180 was cultivated on YPD agar plates (30°C) for 2 days. Single colonies were inoculated into 5 mL liquid YPD (10 g L⁻¹ yeast extract, 20 g L⁻¹ polypeptone, 20 g L⁻¹ glucose) and 5 mL liquid SD (6.7 g L⁻¹ yeast nitrogen base without amino acids, 20 g L⁻¹ glucose) and incubated with shaking (180 rpm) at 30°C for 13 h (YPD) or 18 h (SD). Cultures were transferred to 50 mL of each liquid medium in 200 mL flasks with baffles and incubated from an OD₆₀₀ 0.01 to an OD₆₀₀ 2. The pre-culture and main culture were incubated with shaking (180 rpm) at 30°C. The main culture started at an OD₆₀₀ 0.1. Cells were harvested at an OD₆₀₀ 2 by all of the 3 quenching methods. The sampling volume was 2 mL and 4 replicates were prepared from the pre-preculture. BY4742 and its mutants were cultivated as described for X2180 in liquid YPD. BY4742 and its mutants were harvested when they reached an OD₆₀₀ 1, and the sampling volume was 5 mL. Filtration with a wash method was conducted.

Quenching

We compared the reproducibility of the following three quenching methods and determined the effect of washing with water at the quenching step[3, 7, 14]. For fast

filtration, cells were collected by vacuum filtration using nylon membrane filters (Merck Millipore diameter: 25 mm; pore size: 0.45 μm). After aspiration, filters were immediately placed into 2-mL tubes and freeze dried. Wash procedures used the same volume of Milli-Q water (Millipore, Tokyo, Japan) as the sample broth and were conducted at 4°C. For liquid nitrogen quenching, culture solutions (2 mL) were centrifuged at 2380 $\times g$ at -9°C for 3 min, and then supernatants were removed. Milli-Q water (2 mL) at 4°C was used to wash the sample. The suspension was centrifuged using the same conditions and freeze dried after removing the supernatant. For cold methanol quenching, culture media (2 mL) was added to 10 mL methanol that had been pre-cooled to -60°C. We prepared two different quenching solvents, 100% methanol and 80% methanol (20% H₂O), so that the final concentrations of the solvents after the culture injection were 83% and 67%, respectively. We used the final concentrations to describe the solvent in the Result and Discussion sections. The suspension was centrifuged at described above. To wash samples, 2 mL of Milli-Q water at 4°C was added, and the sample was resuspended. After centrifugation, supernatants were removed, and then the sample was freeze dried. In liquid nitrogen quenching and fast filtration, metabolic reactions are quenched when the cells are immersed in liquid nitrogen (-30 sec fast filtration, 3.5 min liquid nitrogen quenching). On the other hand, in cold methanol quenching, metabolite reactions are stopped when the culture is added to the quenching solvent (-1 sec).

Extraction and derivatization

Extraction and derivatization was performed as described in previous research[3]. The filtered samples were crushed using a ball mill for 5 min at 20 Hz

before extraction to increase the extraction efficiency. Samples in 1 mL extraction solvent, which consisted of methanol/water/chloroform (2.5:1:1) were incubated at 37°C with vigorous shaking (1200 rpm). As an internal standard, 60 µL ribitol (0.2 mg mL⁻¹) was added before incubation. After centrifugation at 16000 ×g for 3 min, 800 µL supernatant was transferred to a 1.5 mL micro tube and mixed with 400 µL Milli-Q water. Centrifugation was performed to separate the nonpolar phase. A 700 µL aliquot from the polar phase was dried using a centrifugal concentrator (VC-36S, TAITEC Co., Tokyo, Japan) for 2 h and subsequently freeze dried. Hydrophilic metabolites were derivatized by methoxyamine hydrochloride (Sigma Aldrich, St. Louis, MO, USA) in pyridine (50 µL, 10 mg mL⁻¹) at 30°C for 90 min, and then, 25 µL of N-methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) (GL Sciences, Tokyo, Japan) was added at 37°C for 30 min.

GC/MS analysis

Pegasus GC-TOF/MS, autosampler 7683B series injector (Agilent Co., Palo Alto, CA, USA), the gas chromatograph 6890N (Agilent), and the mass spectrometer Pegasus III TOF (LECO) were used for X2180. GC-Q/MS, the auto sampler AOC-20is series injector (Shimadzu Inc., Kyoto, Japan), the gas chromatograph GC-2010 Plus (Shimadzu), and the mass spectrometer GCMS-QP2010 Ultra (Shimadzu) were used for BY 4742 and its mutants. CP-SIL 8 CB low bleed/MS fused silica capillary columns (5% diphenyl and 95% dimethylpolysiloxane), 30 m × 0.25 mm i.d. × 0.25 µm film thickness (Varian, Palo Alto, CA, USA) were installed in both platforms. For GC-TOF/MS, the front inlet temperature was 230°C. The helium gas flow rate through the column was 1 mL min⁻¹ for GC-TOF/MS and 1.12 mL min⁻¹ for GC-Q/MS. The

column temperature was held at 80°C for 2 min, raised to 330°C at a rate of 15°C min⁻¹, and maintained for 6 min. The temperatures of the transfer line and ion source were 250°C and 200°C, respectively. 20 scans per second were recorded in the mass range m/z 85-500. For GC-Q/MS, the SIM mode was also carried out. The injection temperature was 230°C. The column temperature was held at 80°C for 2 min, raised at a rate of 30°C min⁻¹ to 120°C, and then raised to 136°C at a rate of 20°C min⁻¹; the temperature was then raised to 300°C at a rate of 30°C min⁻¹. The column temperature was held at 300°C for 3 min. The interface temperature was 250°C. The temperature of the ion source was 200°C. The SIM program is shown in Table 1. We used GC-TOF/MS to assess the reproducibility of metabolomics procedures using wild-type yeast. In transcription factor-deletion mutant experiments, we used 2 different GC-Q/MS systems. The same kind of column had been installed in both systems and the same methods were applied. However, each instrument was operated using the respective auto tuning files to create a virtual environment of a large-scale experiment performed in different laboratories.

Data processing

MS data were exported in the net CDF format. Peak selection and alignment were conducted using MetAlign software (Wageningen UR, The Netherlands, available for free at <http://www.pri.wur.nl/UK/products/MetAlign/>)[15]. The CSV format results were obtained. In this study, we improved the previously reported software AIoutput which can perform peak identification, prediction, and data integration from results exported from MetAlign. The improvement provides a targeted approach to hunt objective metabolite peaks and quantify them by user-defined mass ion. This

development allows the same data matrix to be generated repeatedly; thus, the validation of biological hypotheses and real application of discriminant or regression analyses is now possible. Our program is freely available at <http://prime.psc.riken.jp/>[16]. We focused on 64 compounds that are often detected in yeast (Table S1). SIMCA-P+ 12.0 (Umetrics, Sweden) was used for multivariate analysis. We also examined three data processing methods, i.e., internal standard normalization, external standard normalization, and total intensity normalization, in order to identify the best way to compare data from different cultivation and analysis days. These three methods were applied to four different data batches to assess the ability to correct for variance between different batches. In general, in metabolomics, internal standard normalization is used to correct errors during sample preparation and variance in MS sensitivity. For our analysis, we used ribitol as internal standard. The intensity values of all detected metabolites were divided by the intensity of ribitol. In the external standard normalization method, we generated a calibration curve by analyzing a standard mixture (1.25 pmol to 312.5 pmol, 8 points). Standard mixtures were analyzed before and after yeast samples. Calibration curves were made in the range of $R^2 > 0.95$. This method has been proposed to correct the error from different batch data by estimating the concentration of each metabolite with a calibration curve. In the total intensity normalization method, the intensity of each compound was divided by the sum of the intensity of all target peaks. This method can reduce the variance of MS sensitivity and compare the ratio between all peaks in each sample.

Results and discussion

Reproducibility of procedures of yeast metabolomics

To compare the reproducibility of data from different cultivation and analysis days, we cultivated the wild type (X2180) four times and used three different quenching methods on each day. The effect of the wash step was also estimated. In this paper, we defined “different cultivation” as the cultivation from different agar plate incubations. We also refer to the word “different analysis days” as different derivatization batches and under different auto tuning conditions. Reproducibility was evaluated using RSD values of 29 compounds that had sufficient intensity (intensity of m/z for quantification had more than 1000 arbitrary units in the filter method sample).

The samples cultured in SD medium showed only a few peaks that were high enough to evaluate reproducibility, and the reproducibility results were poor. We observed that compounds detected with a low S/N showed low reproducibility. On the other hand, the YPD sample had high intensity and high reproducibility. The YPD sample showed higher intensity because the yeast extract includes many compounds (like amino acids) that could spike the intensity. The SD medium consists of minimum minerals for growth; therefore, samples in SD medium showed low intensity and reproducibility. Low S/N peaks are easily affected by noise signals. Based on these data, we selected YPD medium for subsequent experiments to assess quenching methods and data processing methods.

We compared three different quenching methods and investigated the effect on reproducibility if the cells were washed. 16 samples (4 cultivation batches; $n = 4$ for each batch) were used to calculate RSD values. RSD values of 29 metabolites showing sufficient intensity were used to assess reproducibility by calculating the ratio of

1 compounds with RSD values less than 30% (Figure 2).

2 As shown Figure 2(a), fast filtration samples showed the highest
3 reproducibility across four different cultivations and analysis of replicates. The fast
4 filtration method had the simple and easiest procedure. In this method, cells are
5 separated from the medium during filtration. Then, the metabolic reaction is quenched
6 when the cells are immersed in liquid nitrogen. On the other hand, cold methanol
7 quenching, which is often used because of its high speed quenching, showed low
8 reproducibility. This method has a limitation in that parts of the cell pellet are
9 sometimes lost when the supernatant is removed. This might be because cells were not
10 strongly retained to the tubes. In addition, we found that the shape of the cell pellets had
11 low reproducibility in each experimental trial. This would be a reason that contributes to
12 the variance at the extraction step. In general, the cold methanol quenching method is
13 suitable to take metabolic snapshots instantaneously (-1 sec). On the other hand, with
14 respect to the reproducibility, fast filtration (-30 sec) would have advantages. In
15 previous research, cold methanol quenching was conducted using a system that enables
16 samples to be taken from fermenters and quenched immediately[17]. However, when
17 the special equipment is not available, the reproducibility of the manual cold methanol
18 quenching method could not be attained. Furthermore, in the fast filtration method, the
19 loss of reproducibility associated with the wash step was low. In this experiment, we
20 used 2 mL of water for the wash sample and it might not be sufficient to wash out the
21 residual medium completely. In a previous report, however, the effect of the washing
22 volume was estimated using stable isotopes and the authors demonstrated that a
23 washing volume of 2 mL could reduce the residual amount within the experimental
24 error range [18]. In this report, we validated the hypothesis that important compounds

1 could be real metabolites from cells using ^{13}C glucose as described below. We observed
2 a decrease in the intensity associated with the washing step but did not find a significant
3 difference in the intensity among the quenching methods (data not shown). Therefore,
4 we assumed that the extraction efficiencies would not be different among quenching
5 methods.

6 For data normalization, the total intensity normalization method, in which the
7 intensity of each peak is divided by total intensity of all the detected peaks in each
8 sample, had the best reproducibility for data from different cultivation and analysis days
9 (Figure 2(c)). We hypothesized this normalization method reduced variance from
10 different sensitivity between different analysis days by comparing the ratio of each peak
11 to the whole chromatogram. As shown in Figure 2(c), total intensity normalization
12 could improve the reproducibility of cold methanol data that had significance variance
13 in cell volume, as described above. In this study, external standard normalization
14 showed low reproducibility because differences in the sensitivity that depend on the
15 analysis day would cause different dynamic ranges in the calibration curve (Figure
16 2(b)).

17 Taken together, these data indicate that the best combination of quenching and
18 data normalization methods is filter quenching and washing combined with total
19 intensity normalization. These data showed 17.2% of compounds with RSD values less
20 than 10% and 69.0% of compounds with RSD values from 10% to 20%. Indeed, we
21 achieved high reproducibility considering these results were from samples isolated on
22 different cultivation and analysis days. This method would improve the reproducibility
23 of samples cultivated in SD medium, and its investigation is ongoing.

Time-segmented SIM

Since compounds with low intensities gave low reproducibility, we optimized time-segmented SIM for *S. cerevisiae* to detect peaks with high intensities. In SIM mode, limiting detected ions can cause decreased reliability for compound identification. To avoid misidentification, we chose detected ions depending on the following threshold [19]:

$$S_{spc} = (M_{SIM} \cdot M_{Scan}) / (|M_{SIM}| + |M_{Scan}|) \geq 0.9$$

Ion selection was based on the spectral similarity score (S_{spc}) between the mass spectrum in the in-house library (M_{Scan} , obtained scan analysis m/z 85–500) and the selected mass spectrum (M_{SIM}). Time-segmented SIM was set based on the retention index of the alkane mixture because it can apply to different temperature gradients. We applied this method for *S. cerevisiae*; the peak intensities were twice as high compared to those in scan mode. We also optimized temperature gradient conditions; the analytical time was half as long, and we accurately detected 64 compounds.

Validation of the high reproducibility protocol

According to the results from the comparison of metabolomics procedures, we generated a protocol to include cultivation in YPD liquid media, quenching by fast filtration with a washing step, high sensitivity analysis in time-segment SIM, and normalization by total intensities of targeted compounds. We applied this protocol to eight mutant strains to validate its high reproducibility. These strains were single knockout mutants of different transcription factors, and we confirmed they could be separated by PCA if they are cultivated and analyzed at the same time (data not shown). A previous study showed that these eight strains have different ethanol tolerance[13],

and are therefore expected to have different metabolic features. With these factors, our method can be proved to have high reproducibility when the eight strains are separated, including replicates from different cultivation and analysis days. Cultivations were conducted on two different days. In each sampling time, two filters were obtained from the same flasks; these two samples were analyzed on different days and using two different GCMS-QP2010 Ultra (Shimadzu). Different GC/MS systems have different sensitivities and can produce separate systematic errors. Table 2 shows the average of similarity for each strain, including the four replicates calculated by using all targeted metabolite data.

The average of similarity was over 0.95 in all conditions, indicating that our protocol can obtain metabolome data with high similarity from different cultivation days, analysis days, or different GC/MS systems.

For samples cultivated in YPD medium, it was possible that quantitative values of metabolites were affected by residual medium even if they were washed at the quenching step. Therefore, we investigated differences in strain detection by applying multivariate analysis. We conducted multivariate analysis using SIMCA-P+ Ver12 (Version 12.0; Umetrics, Sweden). We prepared the following four data sets: (1) samples cultivated and analyzed on day 1, (2) samples cultivated on day 1 and analyzed on another day, (3) samples cultivated another day, (4) the same sample as in (3) but analyzed by GCMS-QP2010 Ultra with different part numbers. First, PCA was conducted to identify differences between data sets. Mean center was chosen as a pre-processing method. Figure 3(a) shows the data normalized by the internal standard normalization method, which is often applied in metabolomics studies.

Samples were color-coded depending on the cultivation day and symbols

1 describe the analysis days or instruments. As shown in Figure 3(a), samples were
2 separated by cultivation days on first components. These data indicate internal standard
3 normalization could not isolate the variance from different cultivation days. The
4 normalization method using a single internal standard would not correct
5 compound-dependent systematic errors of extraction efficiency and derivatization
6 efficiency and its errors might become non-negligible when data from different batches
7 are compared. In some previous researches, the usage of multiple internal standards was
8 recommended. In a large scale study, however, this might be a problem with regard to
9 the errors between several internal standards [20]. To achieve data fusion from different
10 batches, this variance has to be small and the data unsplit by batches. Then, we applied
11 total intensity normalization to these data. Figure 3(b) shows color-coding by cultivation
12 day, and Figure 3(c) illustrates the results colored-coded by strain.

13 In Figure 3(b), samples cultivated on different days were mixed and did not
14 cluster based on cultivation days. These data indicate we can reduce the effect of
15 cultivation days. Therefore, we colored data by strain, and a separation trend for each
16 strain by first component was observed (Figure 3(c)). These results indicate our
17 normalization method reduces variance depending on cultivation day, analysis day, and
18 analytical platform, and it can reveal features of metabolites in each strain. When data
19 from different batches are compared or fused, internal standard normalization is not
20 appropriate to eliminate day-dependent variance and is difficult to compare metabolome
21 data. However, total intensity normalization could isolate metabolome features of each
22 strain. Recently, there have been reports discussing data fusion of different batch data
23 using QC samples[21][22]. This method does not require QC samples, and it is
24 advantageous for large-scale throughput studies, which is an objective of data fusion

1 between batches.

2 Furthermore, we used Orthogonal Projection to Latent Structure Discriminant
3 Analysis (OPLS-DA) to more clearly separate strains and to identify metabolites that
4 have significant strain differences[23]. We made eight models separating each strain.
5 The pre-processing method was Parato.

6 We generated an OPLS-DA model that separated each strain one by one. From
7 these models, we investigated the metabolites that affected separation from other strains.
8 We used the VIP value, which is an index of contribution, to calculate thresholds by
9 bootstrap methods in all models at a 5% significance level and 10,000 times
10 re-sampling[24]. Metabolites that had higher VIP values than thresholds were
11 designated as large contributing compounds (Table 3).

12 Table 3 shows that 20 metabolites contributed to strain separation. In this
13 experiment, we used YPD medium to obtain high intensity, but those compounds might
14 not have been real metabolites from cells, they could be medium components because
15 YPD medium is a rich medium. Therefore, fast filtration applied for cell collection
16 would not remove medium components from the samples. At that point, it should have
17 been proved that the selected metabolites were produced during cell metabolism. To
18 identify yeast metabolites, we cultivated yeast in minimum medium (SD) using ^{13}C
19 glucose and analyzed its metabolites. If key metabolites were from cells, their mass
20 spectrums would shift to higher m/z values compared to the ^{12}C standards. Our data
21 showed only 23 compounds with different mass spectrums of the 64 target compounds.
22 However, almost all metabolites in Table 3 had shifted mass spectrums except for
23 phosphate, leucine, and isoleucine. Phosphate does not contain carbon, so we did not
24 expect to observe a shift. From these data, contributed metabolites for separation were

1 identified as real metabolites. Although about 50 to 70 metabolites are generally
2 identified by GC/MS from yeast, these OPLS-DA results indicate the information is
3 accurate for around 20 metabolites.

4 Our PCA-score plot (Figure 3(c)) shows the specific metabolome without
5 systematic errors in each mutant. Moreover, we could find a clear separation in this
6 PCA-score plot. In particular, *ace2Δ*, *mot3Δ* and *rtg1Δ* showed the clear separation
7 because these transcription factors would perform relatively different functions. *srb2Δ*
8 and *skn7Δ* were clustered closely because these transcription factors would perform
9 functions that are not considered to affect the metabolome directly according to the
10 previous report [25][26]. The *stp2Δ* mutant was also clustered near *srb2Δ* and *skn7Δ* in
11 Figure 3(c) and this mutant should have shown clear separation from other mutants
12 because its gene encodes branched-amino acid permease [27]. We confirmed *stp2Δ*
13 mutant was separated from *srb2Δ* and *skn7Δ* on first and sixth principle components,
14 and branched-amino acids contributed to its separation (data not shown). In our data, the
15 *rtg1Δ* mutant showed significantly lower glutamic acid intensity (Figure S1). Previously,
16 Liao and Butow reported that cells containing null alleles of *RTG1* and *RTG2* were
17 auxotrophic for glutamic acid [28]. They also mentioned that *RTG1* and *RTG2* genes
18 affect the expression of *CIT2*, a gene that encodes citrate synthase. In our data, the
19 *rtg1Δ* mutant showed lower intensity of the integrated peak of citrate and isocitrate than
20 other transcription factor deletion mutants did. The results of the previous report could
21 be validated over the different batches.

22 In this research, we compared the data from different cultivation days and
23 different analytical days. We compared the reproducibility of each experimental step
24 using wild type yeast and optimized the protocol. We also generated metabolic profiles

1 from different cultivation days, different analytical days, and different GC/MS systems
2 for transcription factor deletion mutants. Normalization using total intensity can be
3 applied to other samples, such as clinical samples that are usually analyzed on a large
4 scale or in a long-term study. We also propose that this normalization method might be
5 applied to machine learning as a useful data processing method.

6

1 **Acknowledgment**

- 2 This study was partially supported by JST, Strategic International Collaborative
3 Research Program, SICORP for JP-US Metabolomics. HT was also supported by
4 Grant-in-Aid for Young Scientists B 25871136.

Reference

1. **Raamsdonk LM., Teusink B., Broadhurst D., Zhang N., Hayes a., Walsh MC., Berden J a., Brindle KM., Kell DB., Rowland JJ., Westerhoff H V., Van Dam K., Oliver SG.:** A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol*, **19**:45–50 (2001).
2. **Ding M-Z., Zhou X., Yuan Y-J.:** Metabolome profiling reveals adaptive evolution of *Saccharomyces cerevisiae* during repeated vacuum fermentations. *Metabolomics*, **6**:42–55 (2009).
3. **Yoshida R., Tamura T., Takaoka C., Harada K., Kobayashi A., Mukai Y., Fukusaki E.:** Metabolomics-based systematic prediction of yeast lifespan and its application for semi-rational screening of ageing-related mutants. *Aging cell*, **9**:616–25 (2010).
4. **Cevallos-Cevallos JM., Danyluk MD., Reyes-De-Corcuera JI.:** GC-MS Based Metabolomics for Rapid Simultaneous Detection of *Escherichia coli* O157:H7, *Salmonella Typhimurium*, *Salmonella Muenchen*, and *Salmonella Hartford* in Ground Beef and Chicken. *J Food Sci*, **76**:M238–46 (2011).
5. **Jenkins H., Hardy N., Beckmann M., Draper J., Smith AR., Taylor J., Fiehn O., Goodacre R., Bino RJ., Hall R., and other 17 authors:** A proposed framework for the description of plant metabolomics experiments and their results. *Nat Biotechnol*, **22**:1601–6 (2004).
6. **Jewison T., Knox C., Neveu V., Djoumbou Y., Guo AC., Lee J., Liu P., Mandal R., Krishnamurthy R., Sinelnikov I., Wilson M., Wishart DS.:** YMDB: the Yeast Metabolome Database. *Nucleic Acids Res*, **40**:D815–20 (2012).
7. **Canelas AB., Ras C., Pierick A., Dam JC., Heijnen JJ., Gulik WM.:** Leakage-free rapid quenching technique for yeast metabolomics. *Metabolomics*, **4**:226–239 (2008).
8. **Bolten CJ., Kiefer P., Letisse F., Portais J-C., Wittmann C.:** Sampling for metabolome analysis of microorganisms. *Anal Chem*, **79**:3843–9 (2007).

- 1 9. **Tsugawa H., Bamba T., Shinohara M., Nishiumi S., Yoshida M., Fukusaki E.:**
2 Practical non-targeted gas chromatography/mass spectrometry-based metabolomics
3 platform for metabolic phenotype analysis. *J Biosci Bioeng*, **112**:292–298 (2011).
- 4 10. **Quackenbush J.:** Microarray data normalization and transformation. *Nat Genet*,
5 **32**:496–501 (2002).
- 6 11. **Koutsouba V., Heberer T., Fuhrmann B., Schmidt-Baumler K., Tsipi D.,**
7 **Hiskia a.:** Determination of polar pharmaceuticals in sewage water of Greece by gas
8 chromatography-mass spectrometry. *Chemosphere*, **51**:69–75 (2003).
- 9 12. **Stein SE, Heller DN.:** On the risk of false positive identification using multiple ion
10 monitoring in qualitative mass spectrometry: large-scale intercomparisons with a
11 comprehensive mass spectral library. *J Am Soc Mass Spectrom*, **17**:823–35 (2006).
- 12 13. **Yoshikawa K., Tanaka T., Furusawa C., Nagahisa K., Hirasawa T., Shimizu**
13 **H.:** Comprehensive phenotypic analysis for identification of genes affecting growth
14 under ethanol stress in *Saccharomyces cerevisiae*. *FEMS Yeast Res*, **9**:32–44 (2009).
- 15 14. **Kato H., Izumi Y., Hasunuma T., Matsuda F., Kondo A.:** Widely targeted
16 metabolic profiling analysis of yeast central metabolites. *J Biosci Bioeng*, **113**:665–73
17 (2012).
- 18 15. **Lommen A.:** MetAlign: interface-driven, versatile metabolomics tool for
19 hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem*, **81**:3079–86
20 (2009).
- 21 16. **Sakurai T., Yamada Y., Sawada Y., Matsuda F., Akiyama K., Shinozaki K.,**
22 **Hirai MY., Saito K.:** PRIME Update: Innovative Content for Plant Metabolomics and
23 Integration of Gene Expression and Metabolite Accumulation. *Plant Cell Physiol*, **54**:e5
24 (2013).
- 25 17. **Lange HC., Eman M., Van Zuijlen G., Visser D., Van Dam JC., Frank J., De**
26 **Mattos MJ., Heijnen JJ.:** Improved rapid sampling for in vivo kinetics of intracellular
27 metabolites in *Saccharomyces cerevisiae*. *Biotechnol and Bioeng*, **75**:406–15 (2001).

18. **Kim S., Lee DY., Wohlgemuth G., Park HS., Fiehn O., Kim KH.:** Evaluation and Optimization of Metabolome Sample Preparation Methods for *Saccharomyces cerevisiae*. *Anal Chem*, **85**:2169-2176 (2013).
19. **Tsugawa H., Tsujimoto Y., Arita M., Bamba T., Fukusaki E.:** GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). *BMC bioinformatics* 2011, **12**:131.
20. **Sysi-Aho M., Katajamaa M., Yetukuri L., Oresic M.:** Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* 2007, **8**:93.
21. **Johnson WE., Li C., Rabinovic A.:** Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**:118–27 (2007).
22. **Zelena E., Dunn WB., Broadhurst D., Francis-McIntyre S., Carroll KM., Begley P., O’Hagan S., Knowles JD., Halsall A., Wilson ID., Kell DB.:** Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Anal Chem*, **81**:1357–64 (2009).
23. **Trygg J., Wold S.:** Orthogonal projections to latent structures (O-PLS). *J Chemom*, **16**:119–128 (2002).
24. **Taylor P., Efron B.:** Beiter Bootstrap Confidence Intervals. *J Am Stat Assoc* **82**:37–41 (2012).
25. **Brown JL., North S., Bussey H.:** SKN7, a yeast multicopy suppressor of a mutation affecting cell wall beta-glucan assembly, encodes a product with domains homologous to prokaryotic two-component regulators and to heat shock transcription factors. *J Bacteriol*, **175**:6908–15(1993).
26. **Kim, Y. J., Björklund, S., Li, Y., Sayre, M. H., Kornberg, R. D.:** A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell*, 77(4), 599–608(1994).
27. **De Boer M., Nielsen PS., Bebelman JP., Heerikhuizen H., Andersen H a., Planta RJ.:** Stp1p, Stp2p and Abf1p are involved in regulation of expression of the

- 1 amino acid transporter gene BAP3 of *Saccharomyces cerevisiae*. *Nucleic Acids Res*,
2 **28**:974–81 (2000).
- 3 28. **Liao X., Butow RA.:** RTG1 and RTG2 : Two Yeast Genes Requbd for a Novel
4 Path of Communication from Mitochondria to the Nucleus. *Cell*, **72**:61–71(1993).
- 5

Figures

Figure 1: Score plot from principle component analysis (PCA) of data from different cultivation days and different analysis days. This plot includes data of four different batches.

Figure 2: Ratio of RSD value in each quenching method. The Y-axis describes the ratio of compounds with low RSD value in 29-targeted metabolites. The dots range is the ratio of compounds with an RSD of less than 10%, the grid range is the ratio of compounds with an RSD between 10–20%, and the black-color range is the ratio of compounds with an RSD between 20–30%.

(a) Internal standard normalization

(b) External standard normalization

(c) Total intensity normalization

Figure 3: Score plot from PCA of eight mutants from two different cultivation days and two different analysis days. (a) Internal standard normalization was applied as the normalization method, and data were color-coded by the cultivation day. (b) Total intensity normalization was used to remove systematic errors, and data were color-coded by the cultivation day. (c) Total intensity normalization was used to remove systematic errors, and data were color-coded by mutants.

Figure 1

Figure 2

Figure 3

Table 1: SIM methods. An analysis was segmented into 15 parts and each part included less than 21 SIM channels.

Segment number	Range	SIM Num.	Targeted compounds	Compounds number	Detect <i>m/z</i>			
1	C10-C12	11	n-Propylamine	9	86	89	114	116
			2-Hydroxypyridine		117	128	130	131
			Pyruvate+Oxaloacetic acid		147	152	174	
			Lactic acid					
			Glycolic acid					
			Alanine					
			n-Butylamine					
			2-Hydroxybutyrate					
			2-Aminobutyric acid					
2	C12-C13	21	Valine	11	86	100	103	105
			Urea		106	117	133	135
			Benzoic acid		136	142	144	147
			2-Aminoethanol		158	171	174	179
			Glycerol		180	189	205	299
			Phosphate		300			
			Leucine					
			Isoleucine					
			Proline					
			Nicotinic acid					
3	C13-C14	15	Glycine	8				
			Succinic acid		99	100	101	113
			Uracil		117	119	147	149
			Fumaric acid		193	204	218	219
			Serine		241	245	255	
			4-Methyl Benzoic acid					
			Threonine					
			Allothreonine					
			Thymine					
4	C14-C15	8	β-Alanine	3	86	100	133	136
			Malic acid		147	174	179	248
			Nicotinamide					
5	C15-C16	17	Aspartic acid	8	86	89	98	100

			Methionine		128	140	147	156
			trans-4-Hydroxy-L-proline		170	174	176	179
			Pyroglutamic acid		198	230	232	240
			Cytosine		254			
			4-Aminobutyric acid					
			α -Ketoglutaric acid					
			4-Hydroxyphenethyl alcohol					
6	C16-C17	15	Glutamic acid	5	91	100	103	116
			Phenylalanine		117	128	129	132
			Lauric acid		141	147	156	192
			Ribose		217	218	246	
			Asparagine					
7	C17-1830(RI)	17	Ribitol(Internal standard)	8	86	89	100	103
			Aconitate		129	142	147	155
			5-Amino-Levulinate		156	157	172	174
			Glutamine		204	217	229	256
			Shikimic acid		273			
			Citric acid + Isocitric acid					
			Ornithine					
			Citrulline					
8	1830(RI)-C19	4	Adenine	2	103	129	264	265
			Glucose					
9	C19-C20	14	Lysine	5	86	100	103	128
			Sorbitol		147	154	156	174
			Histidine		205	217	218	281
			Tyrosine		282	319		
			Gallic acid					
10	C20-C21	4	Inositol	1	147	191	217	305
11	C21-C22	3	Cystathionine	1	100	128	218	
12	C22-C27	17	Tryptophan	3	89	100	101	103
			Cysteine+Cystine		129	133	146	147
			Fructose-6-phosphate		157	160	202	217
					218	299	315	387
					388			
13	C27-C28	6	Trehalose	1	103	129	147	191
					217	361		

Range: time separated by retention indices of *n*-alkane mixtures

SIM number: the number of detected ions in each segment

Compound number: the number of detected compounds in each segment

Table 2: Correlation coefficient between each instrument, cultivation day, and analysis day.

	<i>ace2Δ</i>	<i>bas1Δ</i>	<i>mot3Δ</i>	<i>rlm1Δ</i>	<i>rtg1Δ</i>	<i>skn7Δ</i>	<i>srb2Δ</i>	<i>stp2Δ</i>	Ave.
Instruments	0.981	0.976	0.987	0.986	0.983	0.987	0.987	0.981	0.984
Cultivation days	0.985	0.950	0.979	0.986	0.964	0.984	0.969	0.984	0.975
Analysis days	0.974	0.968	0.964	0.983	0.974	0.982	0.979	0.986	0.976

Table 3: Evaluation of discrimination models of OPLS-DA and list of metabolites with higher VIP values in each OPLS-DA model.

Strain name	<i>ace2Δ</i>	<i>bas1Δ</i>	<i>motΔ3</i>	<i>rlm1Δ</i>	<i>rtg1Δ</i>	<i>srb2Δ</i>	<i>skn7Δ</i>	<i>stp2Δ</i>
R2	0.669	0.874	0.802	0.493	0.983	0.748	0.911	0.815
Q2	0.507	0.708	0.606	0.239	0.951	0.487	0.824	0.636
bootstrap score	0.904	0.955	0.918	0.923	0.829	0.916	0.923	0.973
Compound	Glycerol	Glycerol	Glycerol	Glucose	Glutamic	Citric acid	Glycerol	Glutamic
name(VIP value)	(4.980)	(3.003)	(3.957)	(4.168)	acid(4.649)	(3.288)	(3.797)	acid(3.409)
	Leucine	Alanine	Lysine	Aspartic	Alanine	Proline	Alanine	Serine
	(2.069)	(2.577)	(2.718)	acid(3.823)	(3.589)	(2.867)	(2.435)	(3.101)
	Aspartic acid	Phosphate	Phosphate	Glutamic	Ornithine	Phosphate	Lysine	Glycine
	(1.946)	(2.492)	(2.340)	acid(2.131)	(2.568)	(2.772)	(2.397)	(2.240)
	Alanine	Aspartic	Glutamic	Serine	Glutamine	Aspartic	Phosphate	Ornithine
	(1.619)	acid(2.434)	acid(2.004)	(1.391)	(1.472)	acid(2.049)	(2.359)	(1.964)
	Pyroglutamic	Glutamic	Glucose	Glycerol	Glycine	Serine	Glucose	Leucine
	acid(1.565)	acid(1.940)	(1.898)	(1.244)	(1.365)	(1.956)	(2.354)	(1.919)
	Glucose	Lysine	Glutamine	Threonine	Citric acid	Glucose	Aspartic	Pyroglutamic
	(1.420)	(1.779)	(1.690)	(1.094)	(1.285)	(1.775)	acid(1.438)	acid(1.600)
	Isoleucine	Glucose	Leucine	Lysine	Glycerol	Threonine	Glutamic	Aspartic acid
	(1.302)	(1.685)	(1.457)	(1.078)	(0.964)	(1.548)	acid (1.225)	(1.264)
	Valine	Glycine	Pyroglutamic	Leucine		Glycine	Leucine	Citric acid
	(1.206)	(1.644)	acid(1.253)	(0.999)		(1.348)	(1.153)	(1.092)
	Glycine	Threonine	Citric acid			Alanine	Citric acid	Glutamine
	(1.033)	(1.273)	(1.146)			(1.281)	(1.144)	(1.092)
	Proline	Leucine	Alanine			Glutamic	Ornithine	Glycerol
	(0.982)	(1.243)	(0.992)			acid(1.199)	(1.058)	(1.091)
	Lysine	Histidine				Inositol	Asparagine	Phosphate
	(0.981)	(1.060)				(1.191)	(1.026)	(1.051)
						Histidine	Malic acid	Isoleucine
						(1.172)	(0.966)	(0.990)
								Threonine
								(0.967)

VIP values were determined to elucidate features of each strain.

Boot strap score: threshold by boot strap

Strain name describes the strain separated from other strains.