1 **Random Sample Consensus combined with Partial Least Squares regression**
2 **(RANSAC-PLS) for microbial metabolomics data mining and phenotype**
3 **improvement**

4
5 **(Running title: “RANSAC-PLS FOR METABOLOMICS DATA MINING”)**

6
7 Shao Thing Teoh¹

8 Email: teoh_shao_thing@bio.eng.osaka-u.ac.jp

9
10 Miki Kitamura²

11 Email: m-kitamura@towayakuhin.co.jp

12
13 Yasumune Nakayama³

14 Email: ynaka@bio.sojo-u.ac.jp

15
16 Sastia Putri¹

17 Email: sastia_putri@bio.eng.osaka-u.ac.jp

18
19 Yukio Mukai⁴

20 Email: y_mukai@nagahama-i-bio.ac.jp

21
22 Eiichiro Fukusaki^{1*}

23 *Corresponding author

24 Email: fukusaki@bio.eng.osaka-u.ac.jp

25 Phone/Fax: +81-06-6879-7424

1

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

4 ²Towa Pharmaceutical Co., Ltd., 26-7 Ichiban-Cho, Kadoma, Osaka 571-0033, Japan

5 ³Department of Applied Microbial Technology, Sojo University, Sojo University, 4-22-1 Ikeda,
6 Kumamoto 860-0082, Japan.

7 ⁴Department of Bioscience, Nagahama Institute of Bio-Science and Technology, 1266 Tamura,
8 Nagahama, Shiga 526-0829, Japan.

9

10 **Keywords**

11 Metabolomics; Phenotype improvement; Semi-rational strain engineering; Regression
12 model; Data mining; Random Sample Consensus (RANSAC); Partial Least Squares
13 (PLS); 1-Butanol tolerance; *Saccharomyces cerevisiae*; Gas chromatography/mass
14 spectrometry

15

1 **Abstract**

2

3 In recent years, the advent of high-throughput omics technology has made possible a new
4 class of strain engineering approaches, based on identification of possible gene targets for
5 phenotype improvement from omic-level comparison of different strains or growth
6 conditions. Metabolomics, with its focus on the omic level closest to the phenotype, lends
7 itself naturally to this semi-rational methodology. When a quantitative phenotype such as
8 growth rate under stress is considered, regression modeling using multivariate techniques
9 such as Partial Least Squares (PLS) is often used to identify metabolites correlated with
10 the target phenotype. However, linear modeling techniques such as PLS require a
11 consistent metabolite-phenotype trend across the samples, which may not be the case
12 when outliers or multiple conflicting trends are present in the data. To address this, we
13 proposed a data-mining strategy that utilizes Random Sample Consensus (RANSAC) to
14 select subsets of samples with consistent trends for construction of better regression
15 models. By applying a combination of RANSAC and PLS (RANSAC-PLS) to a dataset
16 from a previous study (GC/MS metabolomics data and 1-butanol tolerance of 19 yeast
17 mutant strains), new metabolites were indicated to be correlated with tolerance within
18 certain subsets of the samples. The relevance of these metabolites to 1-butanol tolerance
19 were then validated from single-deletion strains of corresponding metabolic genes. The
20 results showed that RANSAC-PLS is a promising strategy to identify unique metabolites
21 that provide additional hints for phenotype improvement, which could not be detected by
22 traditional PLS modeling using the entire dataset.

23

24

25

1 **Introduction**

2

3 The development of high-throughput, comprehensive analytical technologies ('omics')
4 has enabled a new strategy of strain engineering for phenotype improvement, in which
5 potential gene targets relevant to the objective phenotype are inferred from omic-level
6 comparison of different sample types, for example different strains or growth conditions
7 (1,2). Termed either the semi-rational approach (3) or inverse metabolic engineering (4,5),
8 this kind of approach bypasses the need for elucidation of the underlying genetic or
9 molecular mechanism governing the phenotype, a requirement which has traditionally
10 limited rational strain engineering approaches (6). Metabolomics, with its focus on the
11 simultaneous detection, identification and quantification of the small molecules making
12 up the omic layer closest to the phenotype, i.e. the metabolome (7,8), is particularly suited
13 for application to this semi-rational methodology (9). Metabolomics has been applied to
14 dissect various microbial phenotypes (10–13), and the results support the utility of
15 metabolomics in strain engineering for phenotype improvement.

16

17 While comparison of only two sample types (e.g. strains) is relatively straightforward,
18 false positives may occur because not all differences in metabolite levels are necessarily
19 related to the phenotype under study (14,15). Comparison of a larger number of strains
20 and identifying metabolites consistently correlating with the phenotype in multiple strains
21 would reduce the false positives (16); however, pairwise comparison of all strains
22 becomes much more complicated when many strains are involved. One strategy is to take
23 advantage of the quantitative nature of metabolomics data and construct a regression
24 model with metabolite levels as predictor variables and phenotype as the response
25 variable, for example in (17) or (3). Metabolites with large coefficients in the model may

1 be inferred to be associated with the phenotype, and targeted for modification via genetic
2 approaches. In metabolomics, Partial Least Squares, also known as Projection to Latent
3 Structures (PLS) (18), or some derivative thereof, is often used as it addresses the
4 potential problem of multicollinearity while keeping model interpretation fairly
5 straightforward (19,20).

6
7 In microbial metabolomics, particularly when many strains comprising different genetic
8 backgrounds or classes of perturbed genes are involved, we may expect that some
9 metabolite-phenotype relationships would hold true only for certain subgroups of the
10 samples. This has been suggested in the work of (21), where a PLS model for ethanol
11 tolerance constructed from metabolite data of yeast transcription factor mutants failed to
12 adequately describe the results from yeast metabolic enzyme mutants, and vice versa. One
13 limitation of linear modeling techniques such as PLS is the requirement for a consistent
14 trend across all samples. When outliers not conforming to the overall trend exist, or
15 different groups of samples follow different, conflicting trends, the model may perform
16 poorly, for example showing high prediction error and low robustness under cross
17 validation (Fig. 1). This limits the utility of PLS in semi-rational phenotype improvement,
18 especially for larger datasets where multiple trends are more likely to exist and affect the
19 model performance (3,21).

20
21 As a real-world example, a dataset consisting of GC/MS metabolomics data and tolerance
22 measurements (i.e. growth rates under stress) for 75 samples, representing 19 different
23 strains from a previous study (3) was considered. It could be shown that a ‘better’ model,
24 with lower model error (RMSEE) and higher robustness (Q^2) compared to the previously-
25 reported model (which was constructed from the entire dataset), could be obtained by

1 excluding a number of ill-fitting samples (Fig. S1A). In addition, the new model with
2 trimmed dataset indicated a new important metabolite, cadaverine with a large, negative
3 PLS coefficient, which was less significant in the coefficients plot of the original model
4 (Fig. S1B). This shows that the selection of a subset of samples with consistent
5 metabolite-phenotype trend may potentially reveal additional metabolites associated with
6 the phenotype, at least in the context of the strains included in that subset.

7

8 The need for an appropriate model construction strategy that circumvents the above
9 problem has been raised in previous studies on metabolomics-based phenotype
10 improvement (3,21). It may be advantageous to select a subset of samples that show a
11 consistent correlation trend, then study that trend, i.e. identify phenotype-correlated
12 metabolites within that subset. The question that remains is how to do this selection in an
13 unbiased and objective manner. In this study, we propose the use of Random Sample
14 Consensus (RANSAC) (22), an algorithm for sample subset selection developed and
15 commonly used in the field of image processing, in conjunction with PLS regression
16 modeling (RANSAC-PLS) to address this issue. Hence, our objective is the development
17 and validation of an improved data-mining strategy for maximizing information obtained
18 from microbial metabolomics dataset using RANSAC-PLS. As proof-of-concept, we
19 applied this data-mining strategy to a dataset from an earlier study (3), and validated the
20 newly-discovered important metabolites by examining additional mutant strains.

21

22

23

1 **Materials and Methods**

2

3 **Dataset**

4 A dataset consisting of GC/MS metabolomics data and growth rate under 1.5%(v/v) 1-
5 butanol stress (used as a measure of tolerance) for 19 yeast deletion strains was taken
6 from a previously-published study (3). The dataset comprised 75 samples, 65 predictors
7 or *x*-variables (metabolite peak intensities) and 1 response or *y*-variable (tolerance).
8 Metabolite intensities were normalized to the internal standard (ribitol). The full dataset
9 is provided as can be found in Table S1 in the Supplementary Material.

10

11 **Random Sample Consensus-Partial Least Squares algorithm (RANSAC-PLS)**

12 All programming procedures were implemented in R 3.2.1 (<https://www.r-project.org/>).
13 Autoscaling (column-wise centering and scaling by standard deviation) was first
14 performed on the dataset. Orthogonal Signal Correction (OSC) (23) was implemented
15 based on sample code obtained online (<https://gist.github.com/dgrapov/5166570>) and
16 applied to the full dataset prior to RANSAC-PLS. The RANSAC algorithm was
17 implemented as originally described (22) with a small modification: submodels were
18 evaluated by a ‘score’ parameter (score = inlier num/total samples * Q^2) instead of simply
19 the number of inliers. Construction of PLS submodels within the RANSAC-PLS
20 procedure was performed using the *pls* function contained in the R extension package
21 ‘pls’ (24).

22

23 **Hierarchical Cluster Analysis (HCA)**

24 Hierarchical Cluster Analysis was performed on the matrix of model coefficients (1001
25 models × 65 coefficients) using the built-in *hclust* function of R. Each model was treated

as a ‘sample’ and the models were clustered according to similarity of coefficient profiles. Pearson correlation was chosen as the distance measure, and the Ward linkage method was used. Suitable clusters were determined by visual inspection of the heatmap and dendrogram.

Additional multivariate analysis

PLS model constructions outside of RANSAC-PLS were performed using SIMCA-P+ v12 (Umetrics, Umeå, Sweden).

New deletion strains

New single-deletion strains were proposed based on RANSAC-PLS results and obtained from the readily available EUROSCARF collection produced under the *Saccharomyces* Genome Deletion Project (33), with the exception of *bat1Δ* which was not available in the collection. Additionally, *his3Δ* was selected as a control strain since all EUROSCARF deletions were performed via *KanMX* marker insertion (*gene::KanMX*), and the parental type BY4742 (*MATα leu2Δ0 lys2Δ0 ura3Δ0 his3Δ1*) was already non-functional for *HIS3*, hence further ‘deletion’ by insertion of *KanMX* into the *HIS3* locus (*his3Δ*) should not introduce phenotypic effects beyond those caused by the marker.

Tolerance measurement and GC/MS metabolomics of new strains

All experimental procedures were performed as previously described (3). In brief, all cultivation steps were performed in 50 mL centrifuge tubes (BD, NJ, USA) capped with gas-permeable silicon plugs (Shin-Etsu Polymer, Tokyo, Japan) at 30 °C with 200 rpm rotational shaking. Strains were precultured overnight in Synthetic Complete (SC) medium, then diluted into 6 mL SC medium (spiked with 1.5% (v/v) 1-butanol for the

1 stress condition) such that the starting $OD_{600} = 0.15$. OD_{600} was further measured at 1 h
2 (non-stress condition) or 2 h intervals (stress condition), and specific growth rates were
3 calculated by regression of blank-subtracted OD_{600} measurements against time using a
4 minimum of 4 time-points. The specific growth rate under stress, μ_{stress} was taken as the
5 measure of tolerance. Samples for metabolomics were collected at $OD_{600} = 2$ under the
6 non-stress condition by filtering 5 mL of culture broth through a 25 mm diameter, 0.45
7 μm pore size nylon membrane filter (Millipore, MA, USA) via vacuum suction followed
8 by transfer of the filter-bound cells to 2 mL collection tubes and quenching in liquid
9 nitrogen. Metabolites were extracted in 1 mL mixed solvent (methanol:water:chloroform
10 in 5:2:2 ratio) at 4°C with vigorous shaking (1200 rpm), and water was added to separate
11 the polar and nonpolar phases. Hydrophilic metabolites in the polar phase were retrieved,
12 concentrated and freeze dried, then derivatized by oximation and silylation before
13 instrumental analysis on a GCMS-QP2010 Ultra (Shimadzu, Kyoto, Japan) gas
14 chromatograph coupled with quadrupole mass spectrometer using an InertCap 5MS/NP
15 column (GL Sciences, Tokyo, Japan).

16

17

Results and Discussion

Development of the RANSAC-PLS procedure

The overall flow of our RANSAC-PLS procedure is illustrated in Fig. 2. The RANSAC algorithm iteratively constructs submodels on small subsets of samples, fits the entire sample set to each submodel, and retains the submodels satisfying some criteria. Normally, some error threshold is established, and, for each submodel, the y-value prediction errors for each sample is calculated; samples with prediction error smaller than the threshold are considered ‘inliers’, and the submodels with a sufficient number of inliers (exceeding some pre-determined minimum number) are retained as ‘valid’ models (22). In our implementation, since the submodels are constructed as PLS models, an additional factor of model robustness (as determined by the Q^2 value after including the inliers for each submodel) was also considered. Hence, a ‘score’ criteria was used for determining model retention, where

$$\text{score} = \text{inlier num}/\text{total samples} * Q^2_{\text{submodel with inliers included}}$$

to simultaneously consider the number of inliers (or more accurately, the fraction of total samples that fit the model) and the robustness of the PLS model.

Orthogonal variation is widely recognized to arise in chemometrics and metabolomics data (25,26). Since orthogonal variation (for example arising from instrumental drift or experimental error) may strongly influence the construction of RANSAC submodels, as well as the affecting the fit of data points to the submodels, we used Orthogonal Signal Correction (OSC) (23) as a pre-processing method to systematically remove ‘background’ y-uncorrelated variation prior to RANSAC. The first two OSC components partitioned off 17.31% and 12.69% of total variance in the dataset, respectively, while third

1 component only accounted for 5.36% (Table S2). Based on this, we decided that first two
2 components were significant and used the dataset pre-processed with two OSC
3 components. By using OSC to remove y -uncorrelated variation from the dataset, we also
4 addressed the problem of choosing a suitable number of PLS components in a ‘soft’
5 manner. Since y -uncorrelated variation has been removed, we can now simply set the
6 number of PLS components to 1, because this component is expected to be almost
7 completely y -correlated (25).

8

9 The other parameters for RANSAC-PLS were determined as follows. We used 1/10 of
10 total samples (75 total samples, $75/10=7.5$ samples, rounded up to 8) for initial model
11 construction in each iteration. A dynamic ‘stringency’ parameter was introduced, such
12 that the y -error threshold for inliers as well as the minimum score for retaining a submodel
13 were controlled by the stringency as follows:

14
$$\text{inlier } y\text{-error threshold} = (0.5 - 0.25 * \text{stringency}) * \text{s.d.}_y$$

15
$$\text{min. score} = 1 * \text{stringency}$$

16 Note that y -error is expressed as a multiple of the total standard deviation of y (s.d._y) to
17 facilitate interpretation of its magnitude. The stringency parameter is initially set to 1, but
18 progressively decreased if no submodel was retained in the last 10000 RANSAC
19 iterations. From the equations above, the decreasing stringency parameter causes ‘inlier
20 y -error threshold’ to trend from 0.25 towards 0.5 (larger threshold means more samples
21 are classified as inliers) and ‘min. score’ to trend towards 0 (smaller threshold means more
22 submodels are retained). In this way a set of parameters that are as stringent as possible,
23 while still allowing a small number of models to be found, can be arrived at, ensuring that
24 relatively high-quality models are retained. The parameters and settings above are
25 summarized in Table S3.

Identification of new important metabolites based on RANSAC-PLS

The RANSAC-PLS procedure was set up such that it terminates after 1000 models are obtained. After settling on a stringency parameter of 0.349, RANSAC-PLS yielded a total of 1001 models, with an average of 43.57 inliers per model. The relative importance of each metabolite in a PLS model is indicated by its coefficients; to find important metabolites common to a large number of models, Hierarchical Cluster Analysis (HCA) was performed on the matrix composed of PLS coefficients from all retained models (1001 models $\times$ 65 coefficients; Table S4). The result is visualized in the heatmap shown in Fig. 3. From a visual inspection of the heatmap and dendrogram structure, we defined 3 clusters, with 264, 628 and 109 models respectively. A representative metabolite coefficient profile was then constructed for each cluster, by averaging model coefficients across the cluster (Fig. 4).

In addition to threonine (Threonine_2TMS, Threonine_3TMS) and isocitric/citric acid, which have already been investigated in (3), each cluster showed new, unique metabolites with large coefficients, for example Trehalose in Cluster 1, and Lysine_4TMS and Pyroglutamic_acid_2TMS in Cluster 2 (Fig. 4). We can consider that models in each HCA cluster were constructed from a different subset of samples (the subsets need not be mutually exclusive). Each of the above metabolites were significantly correlated with the y -variable (1-butanol tolerance) in one of the subsets, but not the entire dataset; we can observe that, for example, Pyroglutamic_acid_2TMS has a positive coefficient in Cluster 2 but has a slightly negative overall coefficient in the other clusters. In addition, Cadaverine and Valine_2TMS_major consistently showed large, negative coefficients in the clusters. This suggested these metabolites previously did not show good correlation

1 with 1-butanol tolerance due to the presence of a few ill-fitting samples, which were
2 successfully excluded by RANSAC-PLS. The important metabolites selected based on
3 RANSAC-PLS are listed in Table 1.

5 **Selection of new strains for validating RANSAC-PLS results**

6 Although the large coefficients of the above metabolites only showed the correlation of
7 these metabolites with the target phenotype (1-butanol tolerance), and did not necessarily
8 point to a causal relationship between metabolites and tolerance, previous studies have
9 shown that genetic manipulations e.g. deletions that modify the metabolite levels may in
10 fact also result in improved phenotypes (3,17,21). This implies that correlated metabolites
11 are good starting points for engineering phenotype improvement. To validate the
12 RANSAC-PLS results, we again followed this semi-rational principle of identifying gene
13 targets for phenotype improvement based on predicted changes in levels of the important
14 metabolites. The known metabolic pathways related to each metabolite were obtained
15 from the Yeast Pathways Database (<http://pathway.yeastgenome.org>). Subsequently, gene
16 deletion targets were selected to perturb metabolite levels based on the RANSAC-PLS
17 results above (Table 1). The choice of gene deletion targets were based on the hypothesis
18 that positively correlated metabolites may have a positive influence on tolerance, while
19 negatively correlated metabolites would be the opposite. Hence, we sought to find gene
20 deletions that either increase lysine, pyroglutamic acid or trehalose, or decrease
21 cadaverine or valine.

22
23 The pathways targeted for deletion are as shown in Fig. 5. Pyroglutamic acid is normally
24 converted to glutamic acid by the action of oxoprolinase (*OXPI*), and pyroglutamic acid
25 level is expected to be increased in the *OXPI* deletion strain as reported previously (27),

1 which may lead to an increase in tolerance since pyroglutamic acid was positively
2 correlated with tolerance (Fig. 4B). Meanwhile, trehalose (positive coefficient; Fig. 4A)
3 cleaved to form two units of glucose by the enzyme trehalase. Two different trehalases,
4 one active under acidic condition (encoded by *ATH1* (28)) and the other under neutral pH
5 (encoded by *NTH1* (29)) are known to exist in *S. cerevisiae*, and both genes were selected
6 as deletion targets to increase trehalose accumulation and potentially increase tolerance.
7 On the other hand, valine (negative coefficient; Fig. 4A, B) is produced from α -keto-
8 isovaleric acid in a reversible reaction catalyzed by branched-chain amino acid
9 transaminases, of which two paralogs are known (*BAT1*, *BAT2* (30)). Although the
10 reaction is reversible, deletion of either *BAT* gene may potentially cause a shift in valine
11 level, and we decided to study whether such a shift would be accompanied by a
12 corresponding change in tolerance. Finally, lysine (positive coefficient) and cadaverine
13 (negative coefficient) are substrate and product, respectively, of lysine decarboxylase,
14 therefore we expected that abolishment of this enzyme activity would simultaneously
15 increase lysine and decrease cadaverine, which may be accompanied by an increase in
16 tolerance. However, no lysine decarboxylase gene is known in *Saccharomyces cerevisiae*.
17 Instead, previous studies have indicated that the gene product of *SPE1* coding for
18 ornithine decarboxylase may also exhibit lysine decarboxylase activity (31,32); therefore,
19 we also selected *SPE1* for deletion. Single-deletion strains corresponding to the target
20 genes above are summarized in Table 1.

21

22 **Experimental validation of RANSAC-PLS results using the new strains**

23 To validate the RANSAC-PLS results, the single-deletion strains proposed above were
24 acquired from the EUROSCARF collection (see Materials and Methods). The tolerance
25 (growth rate under 1-butanol stress) of these new strains were then measured (Fig. 6A;

1 see also Table S5). We discovered that *spe1*Δ could not grow in SC medium and was
2 therefore incompatible with the cultivation procedure used so far (3), and hence excluded
3 this strain from further consideration. Of the remaining strains, we found that *oxp1*Δ,
4 *bat2*Δ and *nth1*Δ (but not *ath1*Δ) indeed exhibited higher tolerance compared to *his3*Δ,
5 and we determined that the differences were statistically significant according to
6 Student's *t*-test (*p*-value < 0.05) (Table S5).

7
8 Finally, metabolomics was performed to verify that the targeted metabolite levels did
9 change in accordance to the tolerance increases (Fig. 6B-D; see also Table S6). We found
10 that higher tolerance of *oxp1*Δ was accompanied by increased pyroglutamic acid
11 accumulation (Fig. 6B), verifying the causal link between this metabolite and 1-butanol
12 tolerance. As well, among the two trehalase gene deletants, *nth1*Δ showed much higher
13 trehalose accumulation compared to *ath1*Δ, although both strains had higher trehalose
14 than the control strain *his3*Δ (Fig. 6C), consistent with the higher tolerance of *nth1*Δ
15 compared to the other two strains. The difference in trehalose accumulation of the two
16 trehalase deletants may be explained by the fact that *ATH1* mainly functions under acidic
17 conditions (28) and hence its knockout would have a lesser effect on trehalose levels. On
18 the other hand, although the *bat2*Δ strain showed significantly higher tolerance, its valine
19 level did not noticeably decrease (Fig. 6D), possibly because *BAT2* is preferentially
20 involved in branched-chain amino acid catabolism rather than biosynthesis (34). Overall,
21 although there were one discrepancy between expected metabolite levels and tolerance,
22 three out of four proposed gene deletion targets improved tolerance, indicating the
23 usefulness of our strategy. In addition, it should be noted that while trehalose is known to
24 be involved in various stress responses (35–37), pyroglutamic acid has not been
25 previously linked to either stress response or tolerance. Hence, the metabolomics-based

1 semi-rational strategy, in conjunction with improved data mining by RANSAC-PLS,
2 could also potentially reveal new metabolites related to a phenotype under study and
3 provide greater insight into the mechanisms governing the phenotype.

4 **Conclusion**

6 In conclusion, we have developed a computational procedure based on Random Sample
7 Consensus (RANSAC) combined with Partial Least Squares (PLS) modeling to extract
8 sample subsets showing consistent metabolite-phenotype correlation trends. We applied
9 the procedure to a dataset containing GC/MS metabolomics data as predictor variables
10 and 1-butanol tolerance as the response. By further analyzing the output with Hierarchical
11 Cluster Analysis (HCA), unique metabolites correlated with tolerance different subsets of
12 the samples could be identified. New deletion strains were proposed on the basis of these
13 new metabolites and tolerance improvement was indeed observed in some of these strains.
14 Our results showed that RANSAC-PLS could successfully indicate important metabolites
15 which cannot be found using PLS modeling alone. Further, by obtaining new tolerant
16 strains based on these metabolites, the usefulness of our strategy in phenotype
17 improvement was validated.

1 **Acknowledgements**

2

3 This work was partially supported by the Japan Science and Technology Agency (JST)
4 through a grant awarded under the NSF-JST Strategic International Collaborative
5 Research Program (SICORP) for JP-US Metabolomics.

6

7 This study will be included in a dissertation submitted by Shao Thing Teoh to Osaka
8 University in partial fulfillment of the requirement for his doctoral degree.

9

10

References

1. **Park, S.J., Lee, S.Y., Cho, J., Kim, T.Y., Lee, J.W., Park, J.H., and Han, M.-J.:** Global physiological understanding and metabolic engineering of microorganisms based on omics studies, *Appl. Microbiol. Biotechnol.*, **68**, 567–579 (2005).
2. **Yoshikawa, K., Furusawa, C., Hirasawa, T., and Shimizu, H.:** Design of Superior Cell Factories Based on Systems Wide Omics Analysis, pp. 57–81, in: C. Wittmann, S.Y. Lee (Ed.), *Syst. Metab. Eng.*, Springer Netherlands, (2012). http://link.springer.com/chapter/10.1007/978-94-007-4534-6_3 (accessed March 10, 2013).
3. **Teoh, S.T., Putri, S., Mukai, Y., Bamba, T., and Fukusaki, E.:** A metabolomics-based strategy for identification of gene targets for phenotype improvement and its application to 1-butanol tolerance in *Saccharomyces cerevisiae*, *Biotechnol. Biofuels*, **8**, 144 (2015).
4. **Hong, M.-E., Lee, K.-S., Yu, B.J., Sung, Y.-J., Park, S.M., Koo, H.M., Kweon, D.-H., Park, J.C., and Jin, Y.-S.:** Identification of gene targets eliciting improved alcohol tolerance in *Saccharomyces cerevisiae* through inverse metabolic engineering, *J. Biotechnol.*, **149**, 52–59 (2010).
5. **Santos, C.N.S., and Stephanopoulos, G.:** Combinatorial engineering of microbes for optimizing cellular phenotype, *Curr. Opin. Chem. Biol.*, **12**, 168–176 (2008).
6. **Lee, S.Y., Lee, D.-Y., and Kim, T.Y.:** Systems biotechnology for strain improvement, *Trends Biotechnol.*, **23**, 349–358 (2005).
7. **Fukusaki, E., and Kobayashi, A.:** Plant metabolomics: potential for practical operation, *J. Biosci. Bioeng.*, **100**, 347–354 (2005).
8. **Fiehn, O.:** Metabolomics — the link between genotypes and phenotypes, pp. 155–171, in: C. Town (Ed.), *Funct. Genomics*, Springer Netherlands, (2002). http://link.springer.com/chapter/10.1007/978-94-010-0448-0_11 (accessed December 23, 2012).
9. **Putri, S.P., Nakayama, Y., Matsuda, F., Uchikata, T., Kobayashi, S., Matsubara, A., and Fukusaki, E.:** Current metabolomics: Practical applications, *J. Biosci. Bioeng.*, **115**, 579–589 (2013).
10. **Braaksma, M., Bijlsma, S., Coulier, L., Punt, P.J., and van der Werf, M.J.:** Metabolomics as a tool for target identification in strain improvement: the influence of phenotype definition, *Microbiology*, **157**, 147–159 (2011).
11. **Nugroho, R.H., Yoshikawa, K., and Shimizu, H.:** Metabolomic analysis of acid stress response in *Saccharomyces cerevisiae*, *J. Biosci. Bioeng.*, **120**, 396–404

- 1 (2015).
- 2 12. **Wang, J., Chen, L., Tian, X., Gao, L., Niu, X., Shi, M., and Zhang, W.:** Global
3 Metabolomic and Network analysis of *Escherichia coli* Responses to Exogenous
4 Biofuels, *J. Proteome Res.*, **12**, 5302–5312 (2013).
- 5 13. **Wang, Y., Shi, M., Niu, X., Zhang, X., Gao, L., Chen, L., Wang, J., and Zhang,
6 W.:** Metabolomic basis of laboratory evolution of butanol tolerance in
7 photosynthetic *Synechocystis* sp. PCC 6803, *Microb. Cell Factories*, **13**, 151 (2014).
- 8 14. **Beckmann, M., Enot, D.P., Overy, D.P., and Draper, J.:** Representation,
9 Comparison, and Interpretation of Metabolome Fingerprint Data for Total
10 Composition Analysis and Quality Trait Investigation in Potato Cultivars, *J. Agric.
11 Food Chem.*, **55**, 3444–3451 (2007).
- 12 15. **Böttcher, C., Westphal, L., Schmotz, C., Prade, E., Scheel, D., and Glawischnig,
13 E.:** The Multifunctional Enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3)
14 Converts Cysteine-Indole-3-Acetonitrile to Camalexin in the Indole-3-Acetonitrile
15 Metabolic Network of *Arabidopsis thaliana*, *Plant Cell*, **21**, 1830–1845 (2009).
- 16 16. **Patti, G.J., Tautenhahn, R., and Siuzdak, G.:** Meta-analysis of untargeted
17 metabolomic data from multiple profiling experiments, *Nat. Protoc.*, **7**, 508–516
18 (2012).
- 19 17. **Yoshida, R., Tamura, T., Takaoka, C., Harada, K., Kobayashi, A., Mukai, Y.,
20 and Fukusaki, E.:** Metabolomics-based systematic prediction of yeast lifespan and
21 its application for semi-rational screening of ageing-related mutants, *Aging Cell*, **9**,
22 616–625 (2010).
- 23 18. **Wold, S., Sjöström, M., and Eriksson, L.:** PLS-regression: a basic tool of
24 chemometrics, *Chemom. Intell. Lab. Syst.*, **58**, 109–130 (2001).
- 25 19. **Boulesteix, A.-L., and Strimmer, K.:** Partial least squares: a versatile tool for the
26 analysis of high-dimensional genomic data, *Brief. Bioinform.*, **8**, 32–44 (2007).
- 27 20. **Fonville, J.M., Richards, S.E., Barton, R.H., Boulange, C.L., Ebbels, T.M.D.,
28 Nicholson, J.K., Holmes, E., and Dumas, M.-E.:** The evolution of partial least
29 squares models and related chemometric approaches in metabonomics and
30 metabolic phenotyping, *J. Chemom.*, **24**, 636–649 (2010).
- 31 21. **Ohta, E., Nakayama, Y., Mukai, Y., Bamba, T., and Fukusaki, E.:** Metabolomic
32 approach for improving ethanol stress tolerance in *Saccharomyces cerevisiae*, *J.
33 Biosci. Bioeng.*, doi:10.1016/j.jbiosc.2015.08.006 (available online September 3,
34 2015).
- 35 22. **Fischler, M.A., and Bolles, R.C.:** Random sample consensus: a paradigm for model
36 fitting with applications to image analysis and automated cartography, *Commun.
37 ACM*, **24**, 381–395 (1981).

- 1 23. **Blanco, M., Coello, J., Montoliu, I., and Romero, M.A.:** Orthogonal signal
2 correction in near infrared calibration, *Anal. Chim. Acta*, **434**, 125–132 (2001).
- 3 24. **Mevik, B.-H., and Wehrens, R.:** The pls package: principal component and partial
4 least squares regression in R, *J. Stat. Softw.*, **18**, 1–24 (2007).
- 5 25. **Trygg, J., and Wold, S.:** Orthogonal projections to latent structures (O-PLS), *J.*
6 *Chemom.*, **16**, 119–128 (2002).
- 7 26. **Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E.J., Edlund, U.,**
8 **Shockcor, J.P., Gottfries, J., Moritz, T., and Trygg, J.:** Visualization of GC/TOF-
9 MS-Based Metabolomics Data for Identification of Biochemically Interesting
10 Compounds Using OPLS Class Models, *Anal. Chem.*, **80**, 115–122 (2008).
- 11 27. **Lu, W., Clasquin, M.F., Melamud, E., Amador-Noguez, D., Caudy, A.A., and**
12 **Rabinowitz, J.D.:** Metabolomic Analysis via Reversed-Phase Ion-Pairing Liquid
13 Chromatography Coupled to a Stand Alone Orbitrap Mass Spectrometer, *Anal.*
14 *Chem.*, **82**, 3212–3221 (2010).
- 15 28. **Alizadeh, P., and Klionsky, D.J.:** Purification and biochemical characterization of
16 the ATH1 gene product, vacuolar acid trehalase, from *Saccharomyces cerevisiae*,
17 *FEBS Lett.*, **391**, 273–278 (1996).
- 18 29. **Kopp, M., Müller, H., and Holzer, H.:** Molecular analysis of the neutral trehalase
19 gene from *Saccharomyces cerevisiae*, *J. Biol. Chem.*, **268**, 4766–4774 (1993).
- 20 30. **Kispal, G., Steiner, H., Court, D.A., Rolinski, B., and Lill, R.:** Mitochondrial and
21 Cytosolic Branched-chain Amino Acid Transaminases from Yeast, Homologs of the
22 myc Oncogene-regulated Eca39 Protein, *J. Biol. Chem.*, **271**, 24458–24464 (1996).
- 23 31. **Walters, D., and Cowley, T.:** Polyamine metabolism in *Saccharomyces cerevisiae*
24 exposed to ethanol, *Microbiol. Res.*, **153**, 179–184 (1998).
- 25 32. **Walters, D.R., and Cowley, T.:** Formation of cadaverine derivatives in
26 *Saccharomyces cerevisiae*, *FEMS Microbiol. Lett.*, **145**, 255–259 (1996).
- 27 33. **Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S.,**
28 **Lucau-Danila, A., Anderson, K., André, B., and other 63 authors:** Functional
29 profiling of the *Saccharomyces cerevisiae* genome, *Nature*, **418**, 387–391 (2002).
- 30 34. **Colón, M., Hernández, F., López, K., Quezada, H., González, J., López, G.,**
31 **Aranda, C., and González, A.:** *Saccharomyces cerevisiae* Bat1 and Bat2
32 Aminotransferases Have Functionally Diverged from the Ancestral-Like
33 *Kluyveromyces lactis* Orthologous Enzyme, *PLoS ONE*, **6**, e16099 (2011).
- 34 35. **Iordachescu, M., and Imai, R.:** Trehalose Biosynthesis in Response to Abiotic
35 Stresses, *J. Integr. Plant Biol.*, **50**, 1223–1229 (2008).
- 36 36. **Jia, K., Zhang, Y., and Li, Y.:** Systematic engineering of microorganisms to
37 improve alcohol tolerance, *Eng. Life Sci.*, **10**, 422–429 (2010).

- 1 37. **Reyes, L.H., Abdelaal, A.S., and Kao, K.C.:** Genetic Determinants for n-Butanol
2 Tolerance in Evolved *Escherichia coli* Mutants: Cross Adaptation and Antagonistic
3 Pleiotropy between n-Butanol and Other Stressors, *Appl. Environ. Microbiol.*, **79**,
4 5313–5320 (2013).

5

6

7

8

Figure Legends

Fig. 1. Illustration of the effects of multiple sample subsets showing different correlation trends on the regression

When not all samples conform to the same metabolite-phenotype correlation trend, the results of linear regression modeling may be drastically affected. In the example shown here, A (black circles) and B (white squares) are two groups of sample showing different correlation trends: in group A, x (metabolite) and y (phenotype) are positively correlated ($r_A = 0.9$), while in group B, x and y are almost uncorrelated ($r_B = 0.1$). If samples from both A and B are considered, the correlation takes a value somewhere in the middle ($r_{A+B} = 0.5$), but does not describe either A or B accurately. Further, if an outlier sample (gray triangle) with extreme values is included in the regression, the correlation becomes negative ($r_{all} = -0.5$) which is not true for the majority of the dataset. In microbial metabolomics, multiple metabolite-phenotype correlation trends may arise due to different mechanisms governing the phenotype in different strain groups, while outliers may occur due to gross measurement error.

Fig. 2. The RANSAC-PLS algorithm and overall workflow of the procedure.

First, y -uncorrelated variation is removed by Orthogonal Signal Correction (OSC). Then, the RANSAC-PLS algorithm (indicated by steps enclosed in the large box) iteratively constructs PLS submodels on small subsets of samples, fits the entire sample set to each submodel, calculates the score for each submodel (where score = inlier num/total samples * Q^2) and retains the submodels with scores above the 'min. score' parameter. The 'min. score' parameter is lowered every 10000 iterations until models satisfying the criteria start to be found. After 1000 models have been retained, the procedure terminates. HCA is then performed on the matrix of model coefficients to

cluster models into groups with similar coefficient profiles. All parameters and settings used in this study are provided in Table S3.

Fig. 3. Hierarchical Cluster Analysis (HCA) of model coefficients.

The HCA results are visualized in the heatmap shown here. Each row corresponds to metabolite coefficients of a model retained by RANSAC-PLS. Coefficients are color coded from black (negative) to white (positive), with gray indicating coefficients close to zero. The dendrogram cut-off distance for delineating clusters was determined by visual inspection of the heatmap.

Fig. 4. Average PLS coefficient profiles for each HCA cluster.

The values of PLS coefficients for models in each HCA cluster were averaged and displayed as shown here. (A) Cluster 1. (B) Cluster 2. (C) Cluster 3. Error bars represent standard deviation across the cluster. A large positive or negative value, together with a relatively small error bar indicates that a metabolite's coefficient is consistently large across models in that cluster. Coefficients with large values and small error bars are marked by black (negative) and white (positive) respectively. Dark gray indicates metabolites that have been previously investigated in (3). The matrix of PLS coefficients with models sorted by HCA cluster are provided in Table S4.

Fig. 5. Important metabolites obtained from RANSAC-PLS, relevant metabolic pathways and corresponding genes.

Metabolites with positive coefficients in the HCA clusters (Fig. 4) are shown in **bold**, and were targeted for increased accumulation by the deletion of reactions related to their degradation/utilization. Conversely, metabolites with negative coefficients (underlined) were targeted for decrease by deletion of reactions related to their production.

Fig. 6. Tolerance and metabolite levels for the new strains.

(A) Growth rates under 1.5% (v/v) 1-butanol stress condition for the new strains, taken as the measure of tolerance. (B) Pyroglutamic acid levels of *oxp1*Δ compared to control strain. (C) Trehalose levels of *ath1*Δ and *nth1*Δ compared to control strain. (D) Valine levels of *bat2*Δ compared to control strain. Columns represent average values and error bars represent standard deviations from n=4 biological replicates. All measurement values are shown as relative values to the control strain *his3*Δ (dark gray). The tolerance measurement and metabolite quantification values are provided in Table S5 and Table S6 respectively.

Fig. S1. Effect of removing ill-fitting strain from OPLS model.

(A) Observed vs Predicted plots of original and trimmed OPLS model. *stp2*Δ samples to be excluded are marked on the plot to the left. Two PLS model diagnostics are compared: RMSEE indicates model error (lower means more accurate) while Q² indicates robustness of model predictions under cross-validation (higher means more robust). Lower RMSEE and higher Q² indicated that the trimmed dataset produced a better OPLS model. (B) Comparison of coefficient plots for original model and trimmed models. Error bars coefficient plots represent standard errors estimated from cross validation (CV-error). The negative coefficient of Cadaverine became much larger after removal of *stp2*Δ, indicating that it was negatively correlated with tolerance in the remaining 18 strains.

Table 1. Important metabolites indicated by RANSAC-PLS and corresponding gene deletion targets.

Metabolite	Correlation with tolerance	Deletion target
Trehalose	Positive	<i>ath1Δ</i> or <i>nth1Δ</i> to increase trehalose
Valine	Negative	<i><u>bat1Δ</u></i> or <i>bat2Δ</i> to decrease valine
Cadaverine	Negative	<i><u>spe1Δ</u></i> to decrease cadaverine
Lysine	Positive	<i><u>spe1Δ</u></i> to increase lysine
Pyroglutamic acid	Positive	<i>oxp1Δ</i> to increase pyroglutamic acid

Strains in bold were obtained and subjected to tolerance measurement and metabolome analysis. Underlined strains could not be studied due to unavailability or incompatibility with our cultivation methods (see text).

Fig. 1

Fig. 2

Fig. 3

Fig. 4

Fig. 5

Fig. 6