On the impact of biomass composition in constraint-based flux analysis

Meiyappan Lakshmanan[1], Sichang Long[1], Kok Siong Ang[1], Nathan Lewis[2,3], Dong-Yup Lee[1,4,*]

[1]Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), 20 Biopolis Way, #06-01, Centros 138668, Singapore

[2]Departments of Pediatrics and Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA

[3]Novo Nordisk Foundation Center for Biosustainability at the University of California San Diego, La Jolla, CA 92093, USA

[4]School of Chemical Engineering, Sungkyunkwan University, 2066, Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do 16419, Republic of Korea

Correspondence to: D.-Y. Lee (e-mail: dongyuplee@skku.edu)

ABSTRACT

The biomass equation is a critical component in genome-scale metabolic models (GEMs): It is one of most widely used objective functions within constraint-based flux analysis formulation, describing cellular driving force under the growth condition. The equation accounts for the quantities of all known biomass precursors that are required for cell growth. Most often than not, published GEMs have adopted relevant information from other species to derive the biomass equation when any of the macromolecular composition is unavailable. Thus, its validity is still questionable. Here, we investigated the qualitative and quantitative aspects of biomass equations from GEMs of eight different yeast species. Expectedly, most yeast GEMs borrowed macromolecular compositions from the model yeast, *Saccharomyces cerevisiae*. We further confirmed that the biomass compositions could be markedly different even between phylogenetically closer species and none of the high throughput omics data such as genome, transcriptome and proteome provided a good estimate of relative amino acid abundances. Upon varying the stoichiometric coefficients of biomass components, subsequent flux simulations demonstrated how predicted *in silico* growth rates change with the carbon substrates used. Furthermore, the internal fluxes through individual reactions are highly sensitive to all components in the biomass equation. Overall, the current analysis clearly highlight that biomass equation need to be carefully drafted from relevant experiments, and the *in silico* simulation results should be appropriately interpreted to avoid any inaccuracies.

1 Introduction

Constraint-based modelling methods, such as flux balance analysis (FBA), are popular approaches for analyzing cellular metabolic behaviors *in silico* (Bordbar et al., 2014). Unlike the kinetic modelling, it does not involve complex kinetic parameters and just requires information on metabolic reaction stoichiometry and mass balances around the metabolites under pseudo-steady state assumption (Lewis et al., 2012). Such simplicity of FBA enabled the development and use of hundreds of genome-scale *in silico* models for several species across all three domains of life for the study of microbial evolution, metabolic engineering, drug discovery, context-specific analysis of high throughput omics data and for the investigation of cell-cell interactions (Bordbar et al., 2014).

FBA is an optimization-based approach where a particular objective function is maximized or minimized given a series of constraints such as mass balance, thermodynamic, and enzyme capacity for determining the underlying steady-state fluxes. Among objective functions that have been used to interrogate the metabolic states and cellular behaviors, the maximization of biomass production has been most commonly used in FBA simulations with a principal hypothesis that microbial cells typically strive to grow as fast as possible under exponential growth conditions (Feist and Palsson, 2011; Schuetz et al., 2007). Therefore, almost all reconstructed GEMs include an artificial reaction that accounts for the stoichiometric proportions of various metabolites that make up the macromolecular biomass components, e.g. protein, DNA, RNA and lipid, of the cell. Generally, this biomass equation is derived based on their compositional measurements for the target organism whose GEM being reconstructed. However, since it is not easy to obtain such measurements, researchers often borrowed relevant data from the closest organism or made appropriate estimations from high-throughput data while drafting the biomass equation. However, the formulation of a biomass reaction needs utmost care, since ignoring/missing some of the required metabolites and/or using a wrong

stoichiometric coefficient for certain metabolites may have significant impacts on the model predictions.

Indeed, the accuracy of the biomass equation substantially influences most FBA results as reported in the previous studies; they have examined the global applicability of the biomass equation at various growth rates and substrates (Pramanik and Keasling, 1998), sensitivity of various components in growth predictions (Dikicioglu et al., 2015; Feist et al., 2007; Koduru et al., 2017), variations in gene essentiality predictions (Xavier et al., 2017) and the accuracy of biomass compositions in terms of elemental balances (Chan et al., 2017). Nevertheless, several questions surrounding the biomass equation still remain unclear. For example, do phylogenetically closer organisms have similar macromolecular compositions? How reliable is the estimation of biomass composition from omics datasets? Which particular components influence FBA results the most? Do the effects of biomass coefficients vary across environmental conditions? To answer these questions, here, we first examined the source of biomass compositions used in genome-scale models of eight different yeast species and the variations among them. Then, we analysed the quantitative variations between the macromolecular compositions obtained from omics datasets and the experimentally measured ones. We also investigated the impact of variations in the stoichiometric coefficients of the biomass equation on various flux analysis results such as predicted *in silico* growth rates, essential reactions and flux spans of all the reactions in the GEM across various environmental conditions, thereby identifying key biomass components that have the greatest impact on the FBA results.

2 Results

2.1 Biomass composition can neither be approximated using macromolecular composition from other species nor from the "-omics" data of same species

We first compared the biomass equations from genome-scale models of eight phylogenetically different yeast species: *Candida glabrata* (Xu et al., 2013), *Candida tropicalis* (Mishra et al., 2016), *Kluveromyces lactis* (Dias et al., 2014), *Pichia pastoris* (Chung et al., 2010), *Saccharomyces cerevisiae* (Mo et al., 2009), *Schizosaccharomyces pombe* (Sohn et al., 2012), *Scheffersomyces stipitis* (Balagurunathan et al., 2012) and *Yarrowia lipolytica* (Kavšcek et al., 2015) (see **Supplementary Methods**; **Figure 1A**). Each biomass equation contains metabolites that can be classified into any of the six major categories: carbohydrate, protein, RNA, DNA, lipid and cofactors. A total of 68 metabolites excluding cofactors are involved, and 42% of them are present in all biomass equations (**Supplementary Table 1**). These compounds include the 20 amino acids and 8 nucleotide compounds, representing the protein and nucleotides, respectively. On the other hand, phospholipid and carbohydrate compounds varied the most across GEMs. We also examined the source of biomass composition used in each GEM to assess their quality. Several components were borrowed from one model to another with *S. cerevisiae* being the common ancestor (**Figure 1B**). Inevitably, such heavy cross-referencing of macromolecular components led to overall biomass composition with a very similar level of resolution (**Figure 1C**). Noticeably, except the *S. stipitis*, all other GEMs derive at least a part of their composition from *S. cerevisiae* due to data scarcity. Note that it is a common practice to derive biomass compositions from phylogenetically closer organisms when relevant data is not available. Thus, we compared the variations among the various biomass components, especially amino acids and fatty acids, across yeast species and checked whether this assumption is valid. Surprisingly, we found significant variation in the concentrations of both amino acids (% w/w) and fatty acids (% mol/mol) even between the

phylogenetically closer species (**Figure 1D and 1E**). For example, concentrations of lysine, methionine, glycine and arginine (as measured by % w/w) were markedly different between *C. tropicalis* and *S. stipitis*, despite clustering together in the phylogenetic tree. Such observations clearly raise the question over the adaptation of amino acid concentrations from one species to another as it is done in *Y. lipolytica*, where the values were taken from *S. cerevisiae*.

Apart from experimentally determining the individual monomer composition in proteins, it was also suggested that this information could also be estimated from genome sequence based on the codon usage as it is calculated in the *K. lactis* GEM. We also checked whether this assumption is acceptable by comparing the experimentally measured amino acid compositions of *S. cerevisiae* (Bruinenberg et al., 1983) and *P. pastoris* (Carnicer et al., 2009) with that of the compositions estimated from their genome (Cherry et al., 2014; De Schutter et al., 2009), transcriptome (Adhikari and Cullen, 2014; Wang et al., 2013) and proteome (Ghaemmaghami et al., 2003; Renuse et al., 2014). Not expectedly, none of the high-throughput data provided a good estimate. Instead, comparable values can be obtained when the highly expressed transcripts and proteins are used as input to calculate the amino acid relative abundances (**Figure 2A and 2B**).

2.2 Internal fluxes are highly sensitive to biomass composition compared to growth rate and reaction essentiality predictions

Next, we examined the sensitivity of predicted *in silico* growth rates upon variations in the stoichiometric coefficients of the biomass equation across various environments, thereby identifying key biomass components with the greatest influence on the FBA results (see **Supplementary Methods**). While the growth rate predictions are most sensitive to the carbohydrate and amino acid compositions due to their large stoichiometric coefficients in biomass equation, typically, the deviation is slightly magnified in reduced substrates such as acetate and glycerol than that of glucose, and also higher under aerobic condition than

anaerobic (**Figure 3**). Collectively, these results indicate that the effect of any error in biomass equation could be more significant across different environmental conditions. We also tested the importance of individual component in biomass equation by analyzing the change in essential reactions across diverse environments. These analyses results showed that while there is a notable change in number of essential reactions when a biomass component is omitted as reported earlier (Xavier et al., 2017), the number remained almost same across all environments, thus emphasizing that more or less same number of essential reactions are required to synthesize a particular component (**Supplementary Figure S1**).

We then explored the effect of macromolecular compositions on internal flux distribution by examining the flux spans of all the reactions upon variations in the stoichiometric coefficients of biomass equation. Unlike the variations in growth rates which was mostly influenced by protein and carbohydrate, the flux spans were significantly changed by all the components of biomass equation in *S. cerevisiae* (**Figure 4**). Similar observations were made in all other organisms investigated as well (**Supplementary Figures S2-S7**). Such variations in the allowable fluxes across all reactions clearly indicate that the biomass composition is most critical for predicting accurate internal flux distributions when we apply parsimonious FBA (Lewis et al., 2010), flux variability analysis (Mahadevan and Schilling, 2003) and Monte Carlo flux sampling (Schellenberger and Palsson, 2009).

Finally, to examine whether the changes in biomass components are dominant over the variation in metabolic network content in FBA, we performed all the above mentioned analyses in two different versions of *S. cerevisiae* GEMs, *i*MM904 (Mo et al., 2009) and *i*ND705 (Duarte et al., 2004), with same biomass composition. The resultant variations in the metabolic network content have much more effect than the changes in biomass compositions (see **Supplementary Figure S8**). Most importantly, differences in network contents also affected the identification of essential reactions under various environments. Such observations are

possibly due to the incomplete nature of previous versions of the models, which could be further refined and updated in subsequent ones. Hence, accurate biomass composition can serve as a good template for better model-refinement during the reconstruction process.

3 Conclusions

Overall, this study highlighted that changes in biomass composition have a greater influence on internal flux distributions than on *in silico* growth rate and reaction essentiality predictions. Although the predicted growth rates are most sensitive to amino acid and carbohydrate components owing to their large stoichiometric coefficients, the internal flux distributions are significantly influenced by all the components including small molecule cofactors. Moreover, the biomass equation is arguably the best cellular objective describing the phenomenon of natural selection. For example, microbes may gain or lose metabolic reactions across diverse environments to improve their fitness (Nam et al., 2011). Therefore, it is highly recommended for new and existing GEMs to carefully draft and revise the biomass equation, respectively, through relevant experiments and standardized computational platforms (Lachance et al., 2018) in an unbiased manner as the macromolecular compositions could be markedly different even between phylogenetically closer organisms and the relative abundances obtained from high-throughput omics data are generally poor.

Acknowledgements

This work was supported by the Biomedical Research Council of A*STAR (Agency for Science, Technology and Research), Singapore, the Next-Generation BioGreen 21 Program (SSAC, No. PJ01334605), Rural Development Administra-tion, Republic of Korea, and NIGMS (R35 GM119850).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Adhikari H, Cullen PJ. 2014. Metabolic Respiration Induces AMPK- and Ire1p-Dependent Activation of the p38-Type HOG MAPK Pathway. *PLoS Genet.* **10**:e1004734.

Balagurunathan B, Jonnalagadda S, Tan L, Srinivasan R. 2012. Reconstruction and analysis of a genome-scale metabolic model for *Scheffersomyces stipitis*. *Microb. Cell Fact.* **11**:27.

Bordbar A, Monk JM, King ZA, Palsson BO. 2014. Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* **15**:107–20.

Bruinenberg PM, Van Dijken JP, Scheffers W a. 1983. A theoretical analysis of NADPH production and consumption in yeasts. *J. Gen. Microbiol.* **129**:953–964.

Carnicer M, Baumann K, Töplitz I, Sánchez-Ferrando F, Mattanovich D, Ferrer P, Albiol J. 2009. Macromolecular and elemental composition analysis and extracellular metabolite balances of *Pichia pastoris* growing at different oxygen levels. *Microb. Cell Fact.* **8**:65.

Chan SHJ, Cai J, Wang L, Simons-Senftle MN, Maranas CD. 2017. Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. *Bioinformatics* **33**:3603–3609.

Cherry JM, Karra K, Costanzo MC, Wong ED, Binkley G, Miyasato SR, Fisk DG, Engel SR, Nash RS, Balakrishnan R, Hitz BC, Simison M, Dietrich FS, Dwight SS, Skrzypek MS, Lloyd P, Weng S. 2014. The reference genome sequence of *Saccharomyces cerevisiae* : Then and now. *Genes|Genomes|Genetics* **4**:389–398.

Chung BKS, Selvarasu S, Andrea C, Ryu J, Lee H, Ahn J, Lee H, Lee D. 2010. Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. *Microb. Cell Fact.* **9**:50.

Dias O, Pereira R, Gombert AK, Ferreira EC, Rocha I. 2014. iOD907, the first genome-scale metabolic model for the milk yeast *Kluyveromyces lactis*. *Biotechnol. J.* **9**:776–790.

Dikicioglu D, Kırdar B, Oliver SG. 2015. Biomass composition: the "elephant in the room" of metabolic modelling. *Metabolomics* **11**:1690–1701.

Duarte NC, Herrgård MJ, Palsson BØ. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**:1298–309.

Feist AM, Palsson BØ. 2011. The Biomass Objective Function. *Curr. Opin. Microbiol.* **13**:344–349.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**:121.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* **425**:737–741.

Kavšcek M, Bhutada G, Madl T, Natter K. 2015. Optimization of lipid production with a genome-scale model of *Yarrowia lipolytica*. *BMC Syst. Biol.* **9**.

Koduru L, Kim Y, Bang J, Lakshmanan M, Han NS, Lee D-Y. 2017. Genome-scale modeling and transcriptome analysis of *Leuconostoc mesenteroides* unravel the redox

governed metabolic states in obligate heterofermentative lactic acid bacteria. *Sci. Rep.* **7**:15721.

Lachance J-C, Monk JM, Lloyd CJ, Seif Y, Palsson BO, Rodrigue S, Feist AM, King ZA, Jacques P-E. 2018. BOFdat: generating biomass objective function stoichiometric coefficients from experimental data. *PLoS Comput Biol*. **15**(4): e1006971.

Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, Weitz KK, Eils R, König R, Smith RD, Palsson BØ. 2010. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**:390.

Lewis NE, Nagarajan H, Palsson BO. 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nat. Rev. Microbiol.* **10**:291–305.

Mahadevan R, Schilling CH. 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**:264–276.

Mishra P, Park G-Y, Lakshmanan M, Lee H-S, Lee H, Chang MW, Ching CB, Ahn J, Lee D-Y. 2016. Genome-scale metabolic modeling and in silico analysis of lipid accumulating yeast *Candida tropicalis* for dicarboxylic acid production. *Biotechnol. Bioeng.* **113**:1993–2004.

Mo ML, Palsson BO, Herrgård MJ. 2009. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**:37.

Nam H, Conrad TM, Lewis NE. 2011. The role of cellular objectives and selective pressures in metabolic pathway evolution. *Curr. Opin. Biotechnol.* **22**:595–600.

Pramanik J, Keasling JD. 1998. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol. Bioeng.* **60**:230–238.

Renuse S, Madugundu AK, Kumar P, Nair BG, Gowda H, Prasad TSK, Pandey A. 2014. Proteomic analysis and genome annotation of *Pichia pastoris*, a recombinant protein expression host. *Proteomics* **14**:2769–2779.

Schellenberger J, Palsson BØ. 2009. Use of randomized sampling for analysis of metabolic networks. *J. Biol. Chem.* **284**:5457–61.

Schuetz R, Kuepfer L, Sauer U. 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**:119.

De Schutter K, Lin YC, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouzé P, Van De Peer Y, Callewaert N. 2009. Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nat. Biotechnol.* **27**:561–566.

Sohn S, Kim T, Lee JH, Lee S. 2012. Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of in silico/in vivo mutant growth. *BMC Syst. Biol.* **6**:49.

Wang X, Lin Y, Wang B, Pan L, Ye Y, Han S, Liang S, He M, Zheng S. 2013. Comprehensive structural annotation of *Pichia pastoris* transcriptome and the response to various carbon sources using deep paired-end RNA sequencing. *BMC Genomics* **13**:738.

Xavier JC, Patil KR, Rocha I. 2017. Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. *Metab. Eng.* **39**:200–208.

Xu N, Liu L, Zou W, Liu J, Hua Q, Chen J. 2013. Reconstruction and analysis of the genome-scale metabolic network of *Candida glabrata*. *Mol. Biosyst.* **9**:205–16.

FIGURES

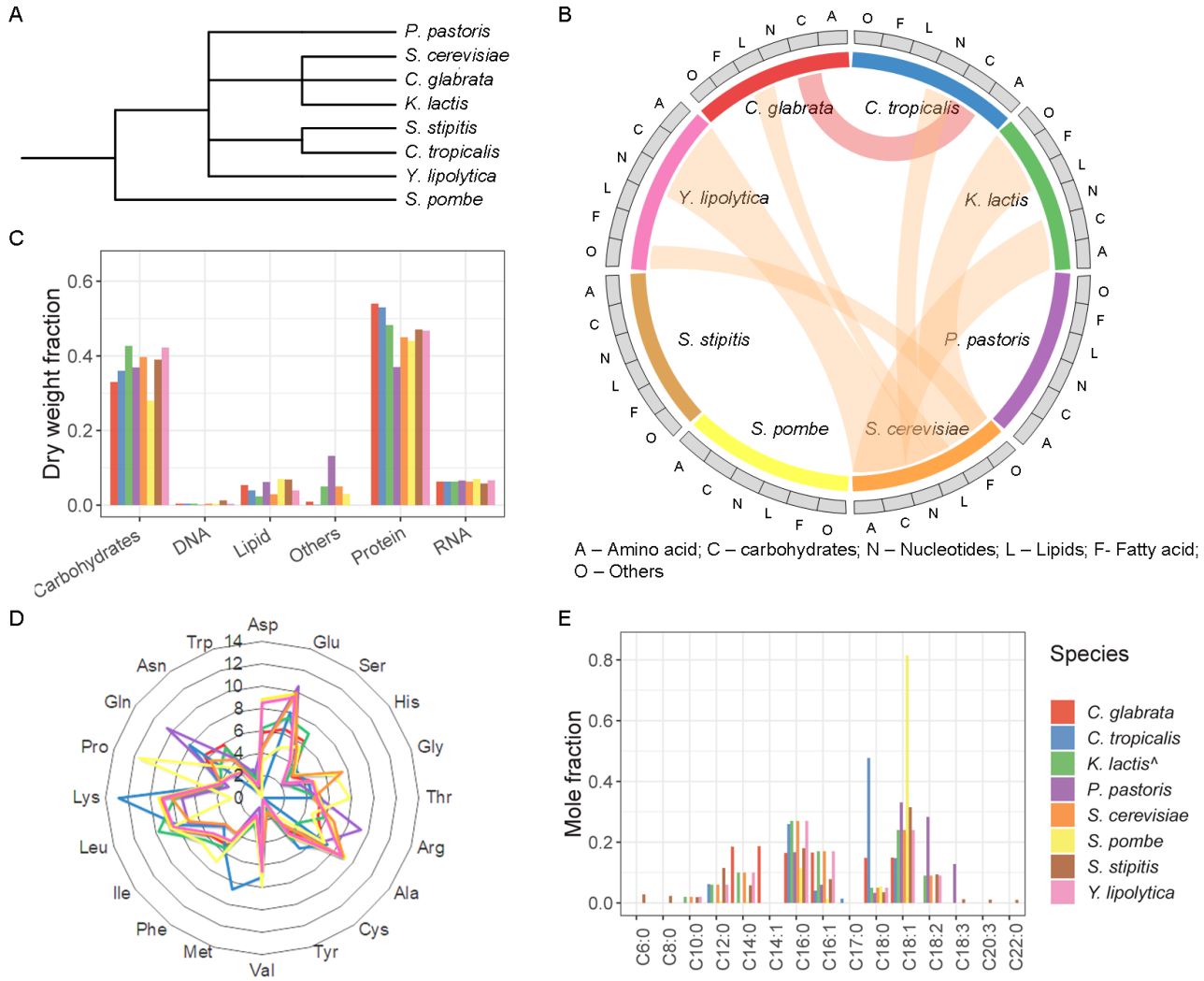


Figure 1. Global assessment of biomass equations from various yeast GEMs. (A) Phylogenetic tree of the yeasts whose GEMs are examined in this study, (B) Borrowing of biomass composition between yeast GEMs, (C) overall biomass composition accounted in GEMs, (D) and (E) relative amino acid and fatty acid compositions of yeast GEMs, respectively.

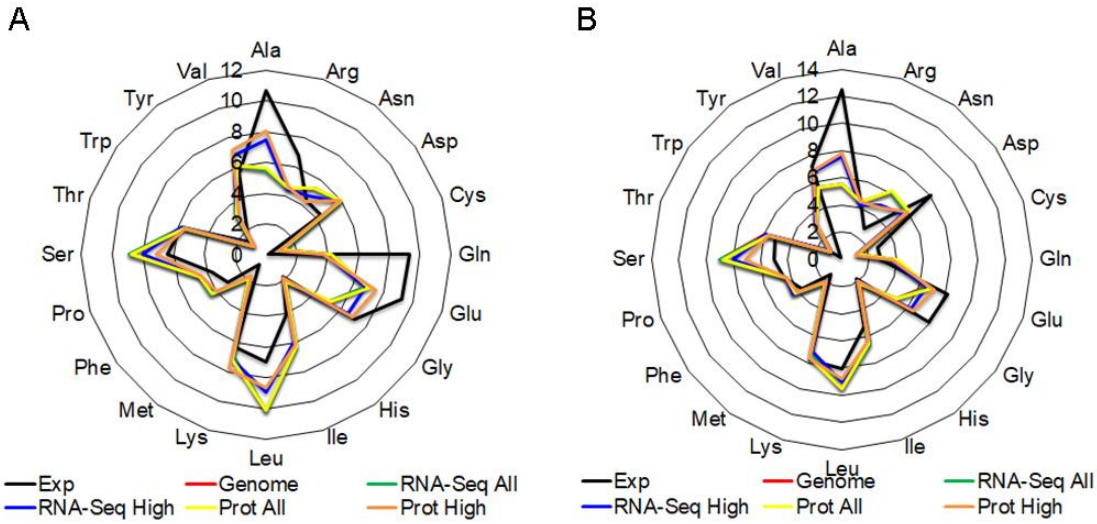


Figure 2. Comparison of amino acid composition from experiments and omics data. Relative differences in amino acid compositions obtained from experiments and estimated from various high-throughput data in (A) *S. cerevisiae* and (B) *P. pastoris*, respectively.

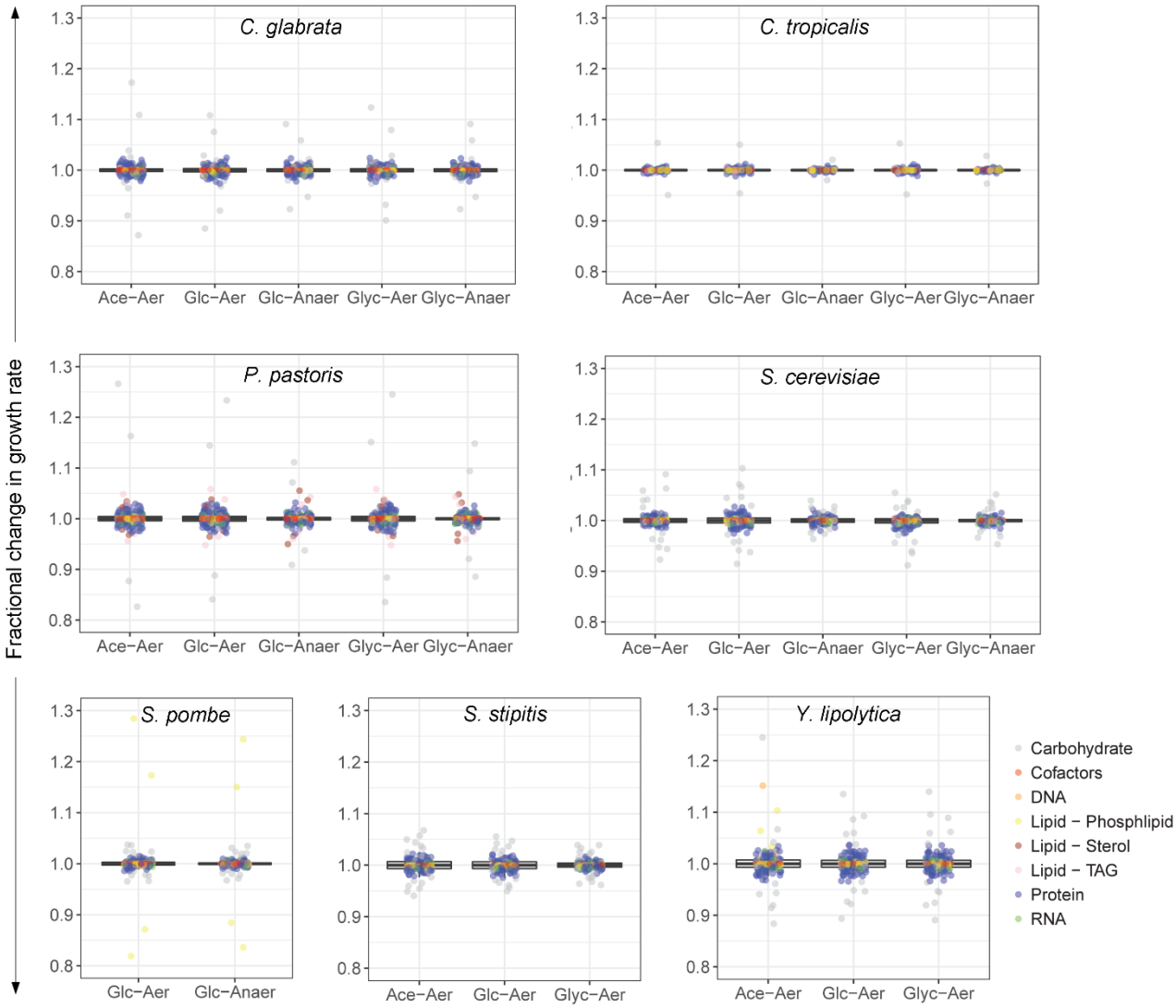


Figure 3. Impact of varying individual biomass components stoichiometric coefficients in predicted *in silico* growth rates in various yeast GEMs. The fractional change in predicted *in silico* growth rates is shown upon varying the stoichiometric coefficient in the range of 25% to 175% in steps of 25%. For visualization purposes, each biomass component is grouped into any of the following six categories: carbohydrate, protein, RNA, DNA, lipid and cofactors. Note that none of the yeast species grow under anaerobic conditions with acetate as carbon source. Similarly, *S. stipitis* and *Y. lipolytica* does not grow in anaerobic conditions. *S. pombe* cannot grow in glycerol and acetate.

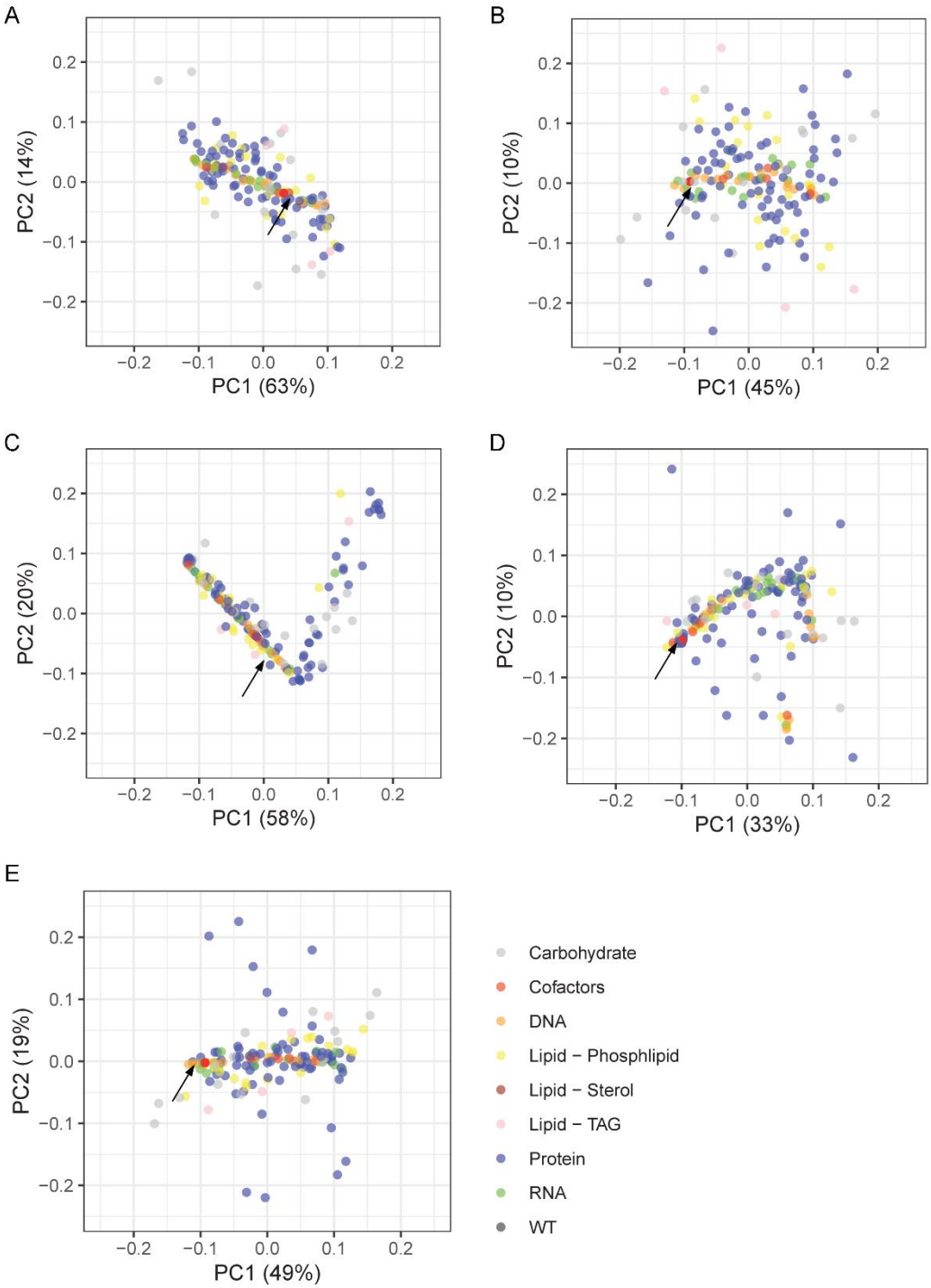
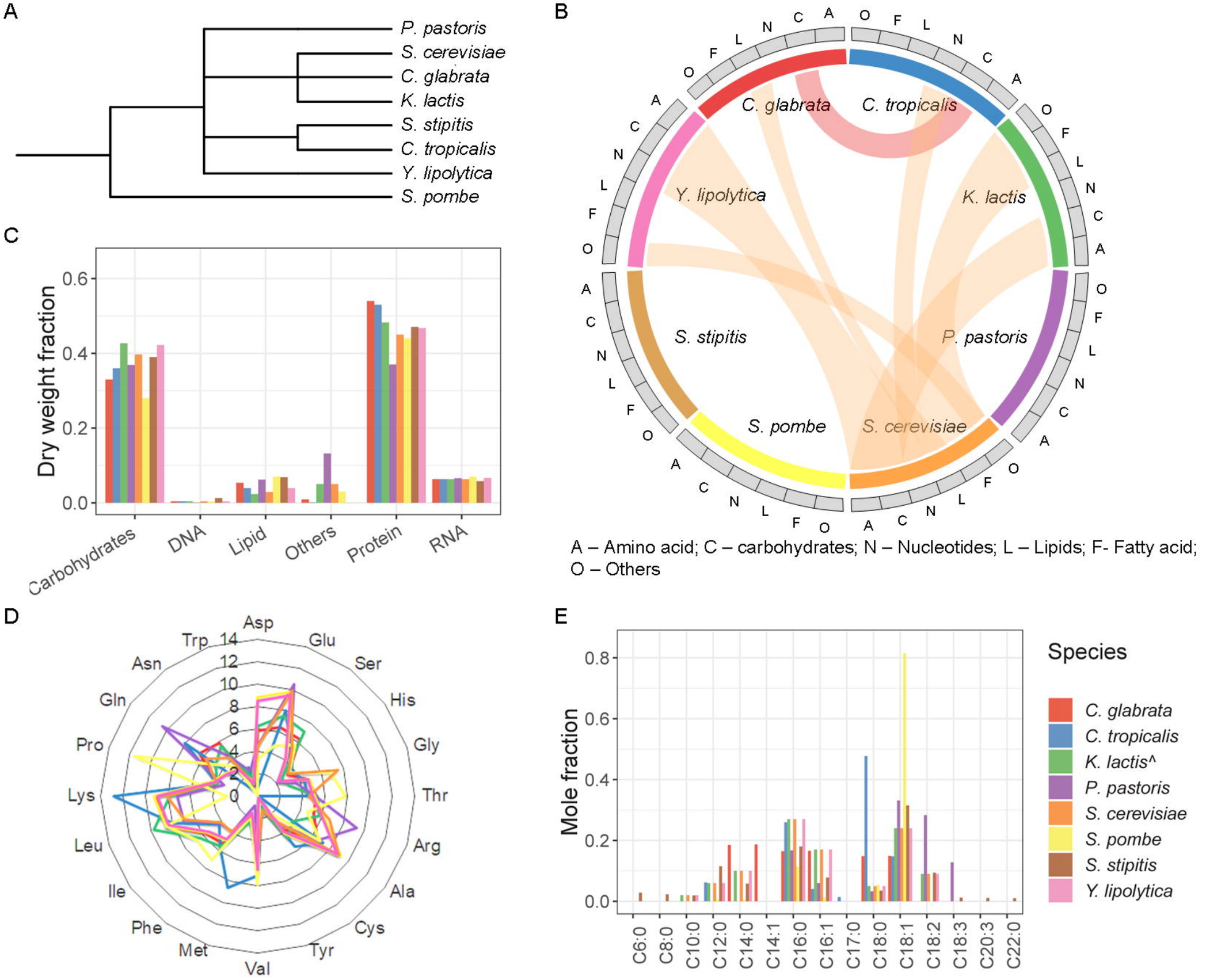
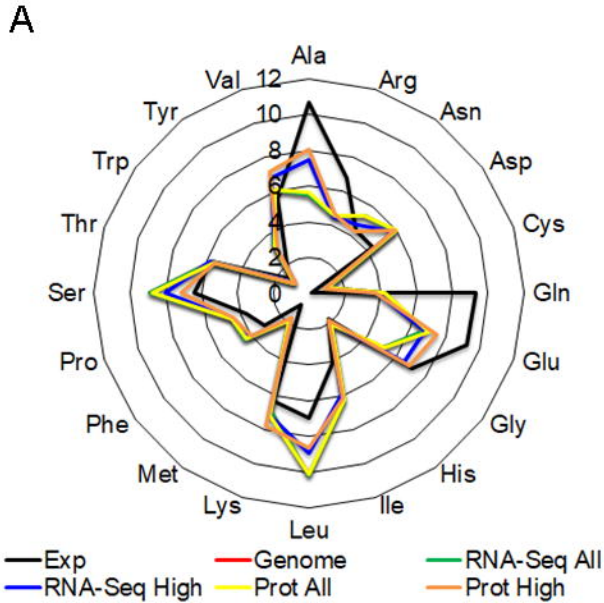
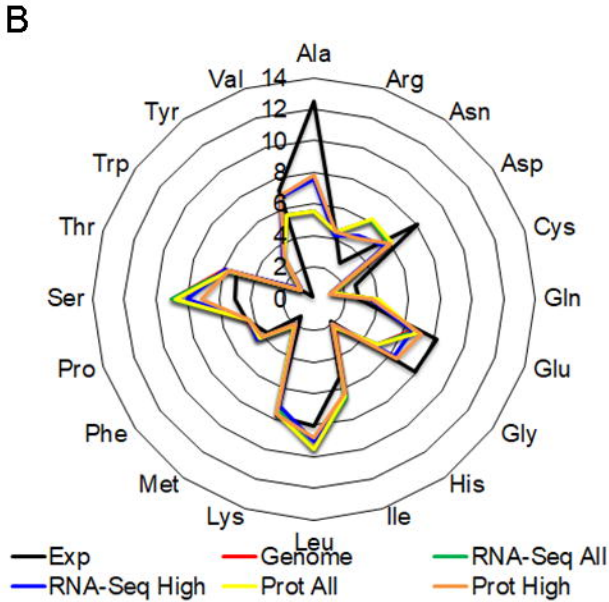
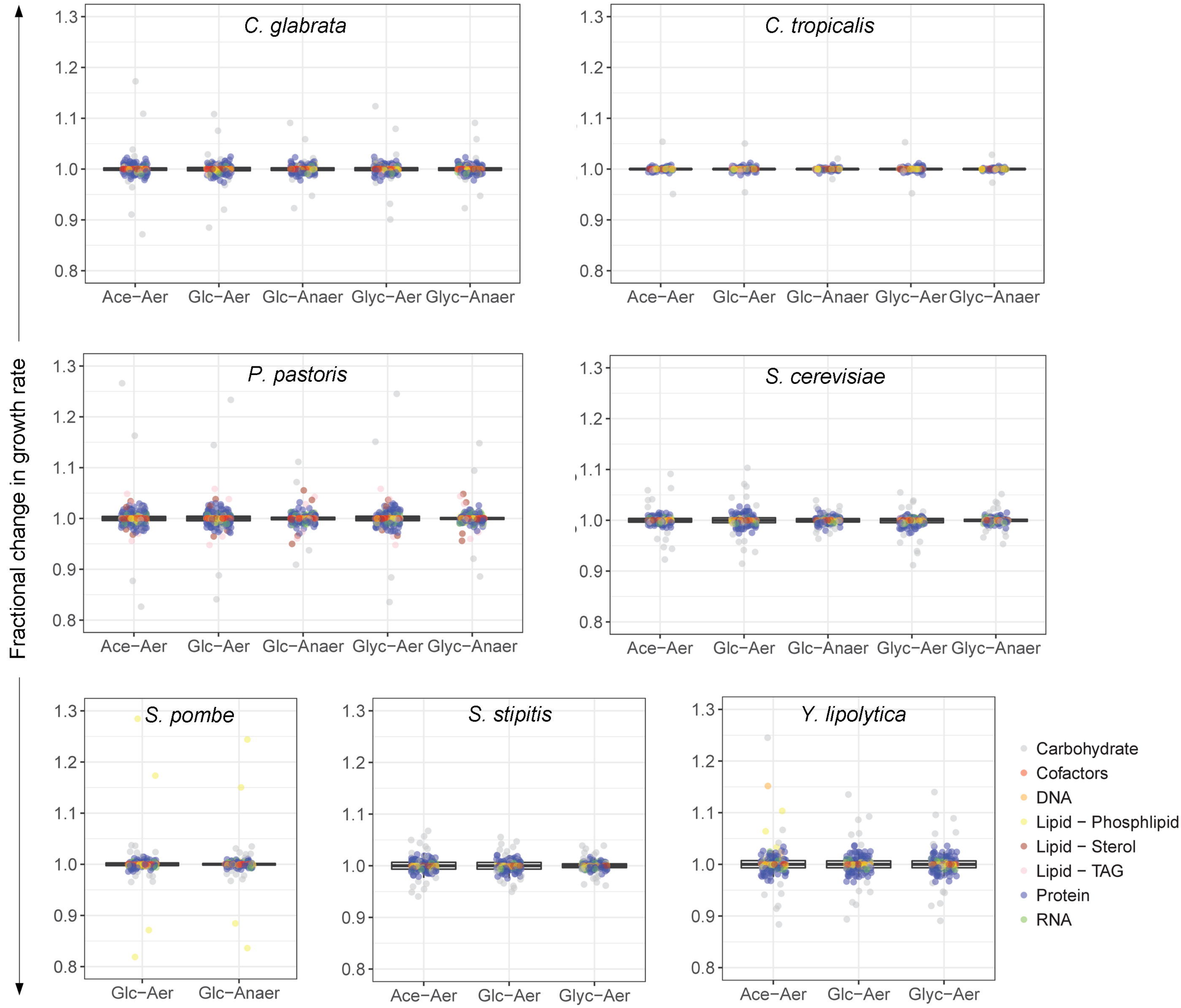


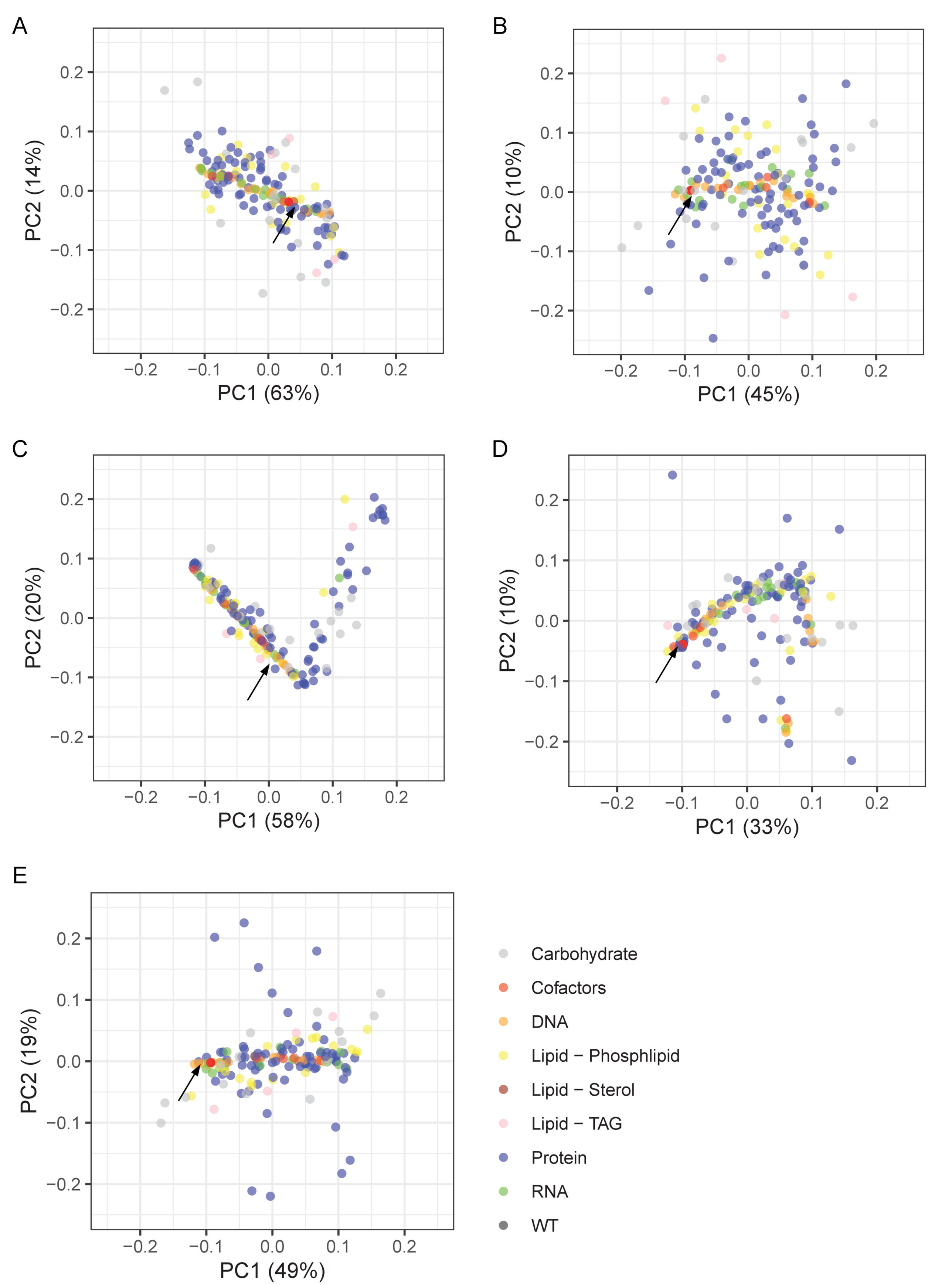
Figure 4. Effect of varying individual biomass components stoichiometric coefficients in internal flux distributions. Flux spans of all reactions in the GEM were first computed by FVA using *S. cerevisiae i*MM904 GEM in 5 environmental conditions: (A) glucose – aerobic, (B) glucose – anaerobic, (C) glycerol – aerobic, (D) glycerol – anaerobic and (E) acetate - aerobic. For visualization purposes, principal component analysis was performed with resulting flux spans. Each biomass component is grouped into any of the following six categories: carbohydrate, protein, RNA, DNA, lipid and cofactors. The flux span of wild type (WT) without any change in biomass composition in the plot is shown with an arrow. Note that *S. cerevisiae* cannot grow under anaerobic conditions with acetate as carbon source.


A
P. pastoris
S. cerevisiae
C. glabrata
K. lactis
S. stipitis
C. tropicalis
Y. lipolytica
S. pombe
B
C. glabrata
C. tropicalis
Y. lipolytica
K. lactis
S. stipitis
P. pastoris
S. pombe
S. cerevisiae
A – Amino acid; C – carbohydrates; N – Nucleotides; L – Lipids; F- Fatty acid; O – Others
C
Dry weight fraction
0.0
0.2
0.4
0.6
Carbohydrates
DNA
Lipid
Others
Protein
RNA
D
Asp
Glu
Ser
His
Gly
Thr
Arg
Ala
Cys
Tyr
Val
Met
Phe
Ile
Leu
Lys
Pro
Gln
Asn
Trp
0
2
4
6
8
10
12
14
E
Mole fraction
0.0
0.2
0.4
0.6
0.8
C6:0
C8:0
C10:0
C12:0
C14:0
C14:1
C16:0
C16:1
C17:0
C18:0
C18:1
C18:2
C18:3
C20:3
C22:0
Species
C. glabrata
C. tropicalis
K. lactis^
P. pastoris
S. cerevisiae
S. pombe
S. stipitis
Y. lipolytica


A
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
0
2
4
6
8
10
12
Exp
Genome
RNA-Seq All
RNA-Seq High
Prot All
Prot High


B
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
0
2
4
6
8
10
12
14
Exp
Genome
RNA-Seq All
RNA-Seq High
Prot All
Prot High


Fractional change in growth rate
C. glabrata
C. tropicalis
P. pastoris
S. cerevisiae
S. pombe
S. stipitis
Y. lipolytica
1.3
1.2
1.1
1.0
0.9
0.8
Ace−Aer
Glc−Aer
Glc−Anaer
Glyc−Aer
Glyc−Anaer
Carbohydrate
Cofactors
DNA
Lipid − Phosphlipid
Lipid − Sterol
Lipid − TAG
Protein
RNA


A
PC2 (14%)
PC1 (63%)
B
PC2 (10%)
PC1 (45%)
C
PC2 (20%)
PC1 (58%)
D
PC2 (10%)
PC1 (33%)
E
PC2 (19%)
PC1 (49%)
Carbohydrate
Cofactors
DNA
Lipid − Phosphlipid
Lipid − Sterol
Lipid − TAG
Protein
RNA
WT