Computational and Structural Biotechnology Journal 21 (2023) 3736–3745

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Mitigating biomass composition uncertainties in flux balance analysis using ensemble representations

Yoon-Mi Choi [a,b], Dong-Hyuk Choi [a], Yi Qing Lee [a], Lokanand Koduru [c], Nathan E. Lewis [d], Meiyappan Lakshmanan [b,e,*], Dong-Yup Lee [a,f,**]

[a] School of Chemical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, Republic of Korea
[b] Bioprocessing Technology Institute (BTI), Agency for Science, Technology and Research (A*STAR), Singapore
[c] Institute of Molecular and Cell Biology (IMCB), Agency for Science, Technology and Research (A*STAR), Singapore
[d] Departments of Pediatrics and Bioengineering, University of California, La Jolla, San Diego, USA
[e] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, and Centre for Integrative Biology and Systems medicinE (IBSE), Indian Institute of Technology Madras, Chennai, Tamil Nadu, India
[f] Bitwinners Pte. Ltd., Singapore

ARTICLE INFO

Keywords:
Flux balance analysis (FBA)
Parsimonious flux balance analysis (pFBA)
Biomass equation
Ensemble modeling
Macromolecular composition

ABSTRACT

The biomass equation is a critical component in genome-scale metabolic models (GEMs): it is used as the *de facto* objective function in flux balance analysis (FBA). This equation accounts for the quantities of all known biomass precursors that are required for cell growth based on the macromolecular and monomer compositions measured at certain conditions. However, it is often reported that the macromolecular composition of cells could change across different environmental conditions and thus the use of the same single biomass equation in FBA, under multiple conditions, is questionable. Herein, we first investigated the qualitative and quantitative variations of macromolecular compositions of three representative host organisms, *Escherichia coli*, *Saccharomyces cerevisiae* and *Cricetulus griseus*, across different environmental/genetic variations. While macromolecular building blocks such as RNA, protein, and lipid composition vary notably, changes in fundamental biomass monomer units such as nucleotides and amino acids are not appreciable. We also observed that flux predictions through FBA is quite sensitive to macromolecular compositions but not the monomer compositions. Based on these observations, we propose ensemble representations of biomass equation in FBA to account for the natural variation of cellular constituents. Such ensemble representations of biomass better predicted the flux through anabolic reactions as it allows for the flexibility in the biosynthetic demands of the cells. The current study clearly highlights that certain component of the biomass equation indeed vary across different conditions, and the ensemble representation of biomass equation in FBA by accounting for such natural variations could avoid inaccuracies that may arise from *in silico* simulations.

1. Introduction

Flux balance analysis (FBA) is a popular approach for analyzing cellular metabolic behaviors *in silico* [1]. Unlike dynamic modelling, which requires detailed kinetic parameters, FBA simply uses the information on metabolic reaction stoichiometry and mass balances around the metabolites, under pseudo-steady state assumption [2,3]. Such simplicity of FBA and the availability of massive amounts of genome sequences from public databases have enabled the development of thousands of genome-scale metabolic models (GEMs) for a multitude of species across all three domains of life [4]. These GEMs have been successfully applied in various studies including microbial evolution, metabolic engineering, drug targeting, context-specific analysis of high throughput omics data and the investigation of metabolic interactions among cells and/or organisms [4,5].

FBA is an optimization-based approach where a particular cellular objective is maximized or minimized while simultaneously constraining the mass balance, thermodynamic and enzyme capacity of a metabolic

* Correspondence to: Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, and Centre for Integrative Biology and Systems medicinE (IBSE), Indian Institute of Technology Madras, Chennai, 600 036, Tamil Nadu, India.
** Corresponding author at: School of Chemical Engineering, Sungkyunkwan University, Suwon-si, Gyeonggi-do, Republic of Korea.
E-mail addresses: meiyappan@iitm.ac.in (M. Lakshmanan), dongyuplee@skku.edu (D.-Y. Lee).

https://doi.org/10.1016/j.csbj.2023.07.025
Received 17 March 2023; Received in revised form 4 July 2023; Accepted 19 July 2023
Available online 23 July 2023

network to determine the plausible steady-state fluxes [6]. Maximization of biomass production has been the most commonly used objective function in FBA with a principal hypothesis that living cells typically strive to grow as fast as possible, particularly under the exponential growth phase [7,8]. Almost all reconstructed GEMs include an artificial reaction, referred to as the 'biomass equation', that accounts for the stoichiometric proportions of various compounds that make up macromolecules of the cellular biomass, e.g., protein, DNA, RNA, carbohydrate, and lipid, for the use as objective function during FBA. In most cases, the macromolecular and monomer compositions used to derive this biomass equation are empirically determined under a certain experimental condition and assumed to remain similar across a wide range of growth environments. However, it is well-documented that cellular volume and the compositions of macromolecular components may vary depending on growth conditions and/or genetic makeup of cells [9–11]. For example, the RNA/protein ratio of *Escherichia coli* exhibits robust correlation with their growth phase and culture conditions such as nutrient utilization/depletion and waste-product accumulation [12]. Recent studies also demonstrated variation in protein and lipid composition in immortalized mammalian cell lines derived from the same original tissue [11], and dynamic nature of macromolecular composition in photoautotroph, *Chlorella vulgaris*, in phototropic and autotrophic conditions [13]. Therefore, the perseverative use of a biomass equation formulated from a single compositional dataset is questionable as the FBA phenotype predictions are sensitive to any variations in individual biomass components [14–20].

Efforts were recently undertaken to address the uncertainty associated with the biomass equation formulation in FBA. Lachance et al. [21] proposed "BOFdat", in which genetic algorithm was used to obtain a particular biomass equation from a series of biomass equations which were stochastically generated by including/removing certain metabolites to best-fit the known gene essentiality data [21]. While this approach utilizes data from multiple environmental conditions to draft the biomass equation with relevant metabolites that make up the cell, it still cannot address the quantitative variations within each macromolecular/monomer component(s). In order to address the quantitative variations of biomass compositions, another study proposed two approaches [22]: 1) an optimal set of trade-off weights were assigned to multiple biomass equations so that it can fit the maximal growth rate; and 2) the coefficients of metabolites in the biomass equation are estimated by interpolation of known sets of macromolecular/monomer compositional data measured under different environmental conditions, assuming linearity between compositions and the environmental changes. The key limitation of this approach is the theoretical treatment of biomass variations, i.e., linear variation across environments, as macromolecules vary only within a certain range and not necessarily exhibit a linear relationship. The development of genome-scale metabolism and expression models (ME-models) systematically eliminated the use of a biomass equation by expanding the scope through the inclusion of pathways that constitute the cellular transcription and translation, and thus able to predict compositions of proteome and transcriptome with the constant structural composition [23–25]. However, it is often difficult to collect all the required kinetic parameters such as transcription/translation rates and catalytic turnover constants specific to each mRNA or enzyme to establish a good quality ME-model.

To address the uncertainties in the biomass equation and make it applicable across multiple environmental conditions, we first need to find answers to the following open questions: Do all macromolecular and monomer compositions in biomass vary across environmental conditions? If so, how significant are those variations? Do phylogenetically close organisms have similar monomer compositions? How reliable is the estimation of biomass composition from omics datasets? How much does such a natural variation of biomass composition impact model predictions? In this study, we first examined the variations in biomass compositions in three representative host organisms, namely *E. coli*, *Saccharomyces cerevisiae* and Chinese hamster (*Cricetulus griseus*) ovary (CHO) cells to answer all the above-mentioned questions. We also investigated the quantitative variation between monomer compositions obtained from omics datasets and the experimentally measured ones. Based on the analysis results, we newly propose the use of an ensemble of biomass equations within the FBA framework, to better capture the natural variation in biomass compositions.

2. Methods

2.1. Compilation of macromolecule and monomer compositions and the estimation of their natural variations

Macromolecular and monomer composition data for three representative species, *E. coli*, *S. cerevisiae*, and CHO cells, was collected through a targeted literature search. Data was collected from various studies which reported macromolecular and monomer composition across various conditions including changes in environmental conditions including temperature, dilution rates, oxygen concentrations, media compositions, and growth phases. Similar data was also collected from multiple mutants, strains, and cell lines of the same species. In addition, we also obtained the monomer composition for the species that are phylogenetically close to either of *E. coli*, *S. cerevisiae*, or CHO cells. The full list of biomass composition data and its source are available in Supplementary File S1.

Natural variability of all macromolecules and monomers were calculated using the coefficient of variation (CV). It was calculated by dividing the standard deviation of collected and/or processed composition data (mass % or mole %) by the mean value of the corresponding biomass component. The CVs of monomers necessitated an additional step, where the average of multiple monomer CVs was taken to represent a final CV. The variability of monomers in DNA was simply determined based on multiple guanine-cytosine (GC)-content data crawled with a query of each species name of the three organisms from NCBI genome database. Note that the estimation of ribonucleotides composition from omics data was not taken into consideration when we evaluated the CVs.

2.2. Estimation of monomer composition from omics-data

We collected multi-omics data of *E. coli*, *S. cerevisiae*, and CHO cells to estimate ribonucleotide and amino acid composition. Genome data were obtained from NCBI RefSeq database [26], transcriptomic datasets were collected from the Sequence Read Archive [27] and NCBI Gene Expression Omnibus (GEO) [28], and proteomic data was downloaded from PaxDB [29]. Note that amino acid composition is not estimated from proteome for CHO cells due to lack of whole proteomic data for CHO cells in PaxDB. The source information of each omics data used in this study is listed in Supplementary File S2.

For the transcriptome and proteome data, list of genes that were expressed were used to extract corresponding coding sequences to estimate the ribonucleotide and amino acid compositions. These gene lists were then classified into two categories: all expressed and top 10% highly expressed genes. In case of genome data, all known genes encoded were considered. The gene sets from each category were first labelled according to their respective source: Genome, Transcriptome-all, Transcriptome-high, Proteome-all, and Proteome-high for estimating amino acid composition, and Genome, Transcriptome-all and Transcriptome-high for ribonucleotide composition. While the coding sequences of genes were considered as it is for genome and transcriptome data, a corresponding amino acid sequence was first obtained using the ExPasy translate tool. Then the frequency of monomer i (p_M^i) is obtained from the coding sequences as follows:

$$p_M^i = \begin{cases} \frac{\theta^i}{\theta_{All}} \forall i \in G, & if\ Genome, Transcriptome-all, Proteome-all \\ \frac{\theta^i}{\theta_{Top10\%}} \forall i \in G, & if\ Transcriptome-high,\ Proteome-high \end{cases}$$

$$G = \begin{cases} \{1,2,3,4\} & if\ M = Ribonucleotide \\ \{1,2,\ldots 20\}, & if\ M = Amino\ acid \end{cases}$$

$$G = \begin{cases} \{1,2,3,4\}, \\ \{1,2,\ldots 20\}, \end{cases} if \quad M = Ribonucleotide$$

$$if \quad M = Amino \quad acid$$

where θ^i stands for the occurrence of monomer i, and θ_{All} is the total number of monomers. Total number of monomers in genome was based on coding sequences of all genes while in transcriptome and proteome data it was based on expressed genes only. $\theta_{Top10\%}$ denotes the total number of monomer units encoded by top 10% expressed genes in transcriptome or proteome data while θ_{All} considers all genes expressed (count >10).

2.3. Sensitivity analysis of growth rates and intracellular metabolic fluxes upon biomass composition variations

We used parsimonious flux balance analysis (pFBA) [30] for analyzing the sensitivity of growth rate and intracellular fluxes upon varying biomass compositions. In this method, the GEM is first converted into irreversible one by converting all reversible reactions into two reactions in which one represent forward and the reverse direction, respectively. Then, the maximum possible growth rate is simulated by solving the following linear programming problem:

$$\max \quad Z = v_{biomass} \tag{P1}$$

$$s.t \quad S_{irrev} \bullet v_{irrev} = 0$$

$$0 \le v_{irrev,j} \le v_{\max,j}$$

where S_{irrev} is a stoichiometric matrix where all reactions are represented as irreversible, $v_{irrev,j}$ is the irreversible flux through reaction j, $v_{\max,j}$ is the maximum allowable flux through reaction j, and $v_{biomass}$ denotes the flux through biomass equation. Subsequently, the following linear problem is solved where the sum of all irreversible fluxes is minimized while simultaneously constraining the previously obtained flux through biomass Eq. (P1) as the lower bound

$$\min \sum_j v_{irrev,j} \tag{P2}$$

$$s.t. \quad v_{biomass} \ge Z$$

$$S_{irrev} \bullet v_{irrev} = 0$$

$$0 \le v_{irrev,j} \le v_{\max}$$

To investigate how flux predictions are affected by variations in biomass composition, we modified the coefficient of a specific component in a reference biomass equation. The reference equation was established based on the average composition of biomass. We altered the coefficient of the target component by 25% of the average mass fraction (g/g dry cell weight, DCW) and normalized the equation so that the total sum of all components equaled 1gDCW. We maintained the original relative amounts of the other macromolecules in the equation. Furthermore, we maintained the original monomer compositions when varying the mass fractions of the macromolecules, and vice versa. The biomass equation (B) can be represented as follows:

$$\boldsymbol{B}(p,d,r,c,l,o) = \boldsymbol{P}(p, q_{amino \quad acid}) + \boldsymbol{D}(d, q_{DNA}) + \boldsymbol{R}(r, q_{RNA}) + \boldsymbol{C}(c, q_{carb}) + \boldsymbol{L}(l, q_{fa}) + \boldsymbol{O}(o, q_{others})$$

where *p, d, r, c, l, o* denotes the macromolecular weight fraction of protein, DNA, RNA, carbohydrate, lipid, and others such as ions/co-factors, respectively. The macromolecular synthesis equations, symbolized by ***P, D, R, C, L, O***, are functions of q vectors representing the composition of monomers or ions.

By varying the biomass components iteratively, we first evaluated the change in *in silico* growth rate predictions using problem (**P1**). Next, we sought to estimate the sensitivity of metabolic flux distributions of all reactions to change in each biomass component. Flux vectors were obtained under two conditions: the reference condition, which was based on the average biomass composition, and an altered biomass composition condition where we modified the biomass equations. To obtain the sensitivity (s), we first obtained flux vectors (v) by solving problem (**P2**). We then calculated sensitivity (s) using equation as follows:

$$v_{ref} = v, when \quad \boldsymbol{B}(p_{mean}, d_{mean}, r_{mean}, c_{mean}, l_{mean})$$

$$s = \sum_i^N \frac{1}{2} \left(\frac{|v_{i,ref} - v_{i,\min}|}{v_{i,ref}} + \frac{|v_{i,ref} - v_{i,\max}|}{v_{i,ref}} \right)$$

where $p_{mean}, d_{mean}, r_{mean}, c_{mean}, l_{mean}$ are the average fractions of protein, DNA, RNA, carbohydrate, and lipid, respectively. N is the number of reactions in a GEM. To calculate the sensitivity of protein, for instance, we increased or decreased its mass fraction by 25% of the average fraction value and normalized the fraction of the other components (*d', r', c', l'*), as shown below:

$$v_{\min} = v, when \quad \boldsymbol{B}\left((1-0.25) \times p_{mean}, d', r', c', l'\right)$$

$$v_{\max} = v, when \quad \boldsymbol{B}((1+0.25) \times p_{mean}, d', r', c', l')$$

Finally, we divided the absolute sensitivity by the sum of absolute sensitivities for all macromolecules and monomers to obtain the relative sensitivity.

All simulations were performed using the COBRA Toolbox v2.0 [31] implemented within MATLAB with Gurobi 7 as the optimization solver. The following GEMs were used for each organism: *E. coli* - *i*ML1515 [32], *S. cerevisiae* - Yeast8.0.0 [33] and CHO cells - *i*CHO2291 [34].

2.4. Ensemble representations of biomass

In order to represent biomass equation as ensembles and implement pFBA using each of the biomass equation, the first step is to establish a single set of 'reference' macromolecular/monomer compositions, which can be determined by calculating the mean of multiple measurements. The subsequent step involves randomly generating a set of 'n' number, e.g., 5000, of biomass equations by altering the composition of the individual biomass components which either vary across environmental conditions or highly sensitive to flux predictions. Our analysis revealed that either all macromolecules vary significantly or are sensitive to flux predictions and thus need to be varied. On the other hand, we noted that compositions of all monomers, except fatty acids, neither vary across conditions nor found to be sensitive to flux predictions and need not be varied in biomass equations. The variation in each biomass composition should be carried out within a specified range indicated by the coefficient of variation (CV). Once the 'n' number of biomass equations is generated, pFBA is implemented 'n' times, with each iteration using a different biomass equation, and the resulting flux distribution is analyzed as range of fluxes for each reaction.

3. Results

3.1. Assessing the validity of common assumptions made while drafting the biomass equation

To understand the natural variations within biomass components, we first compiled relevant macromolecular and monomer composition data for three highly divergent and representative organisms, *E. coli* (a prokaryote), *S. cerevisiae* (a unicellular eukaryote), and CHO cells (immortalized cells derived from a multicellular eukaryote) from literature. In total, we obtained 352 (105, 127 and 120 for *E. coli*, *S. cerevisiae* and CHO cells, respectively) different cellular composition data from various studies which were measured across various environmental and/or genetic conditions: conditions which reported the cells grown in different substrates, different levels of oxygenation and data from mutants (see Methods, Supplementary File S1). We then calculated the coefficient of variation (CV) of each macromolecule within unit biomass, i.e., 1 g of DCW, and each monomer within a particular unit macromolecule for each species. This analysis revealed that macromolecules show larger variation than monomers across the three species (CV = 6–87% compared to that of 2–50%). Within macromolecules, lipids, which are biomolecular compounds involved in long term energy

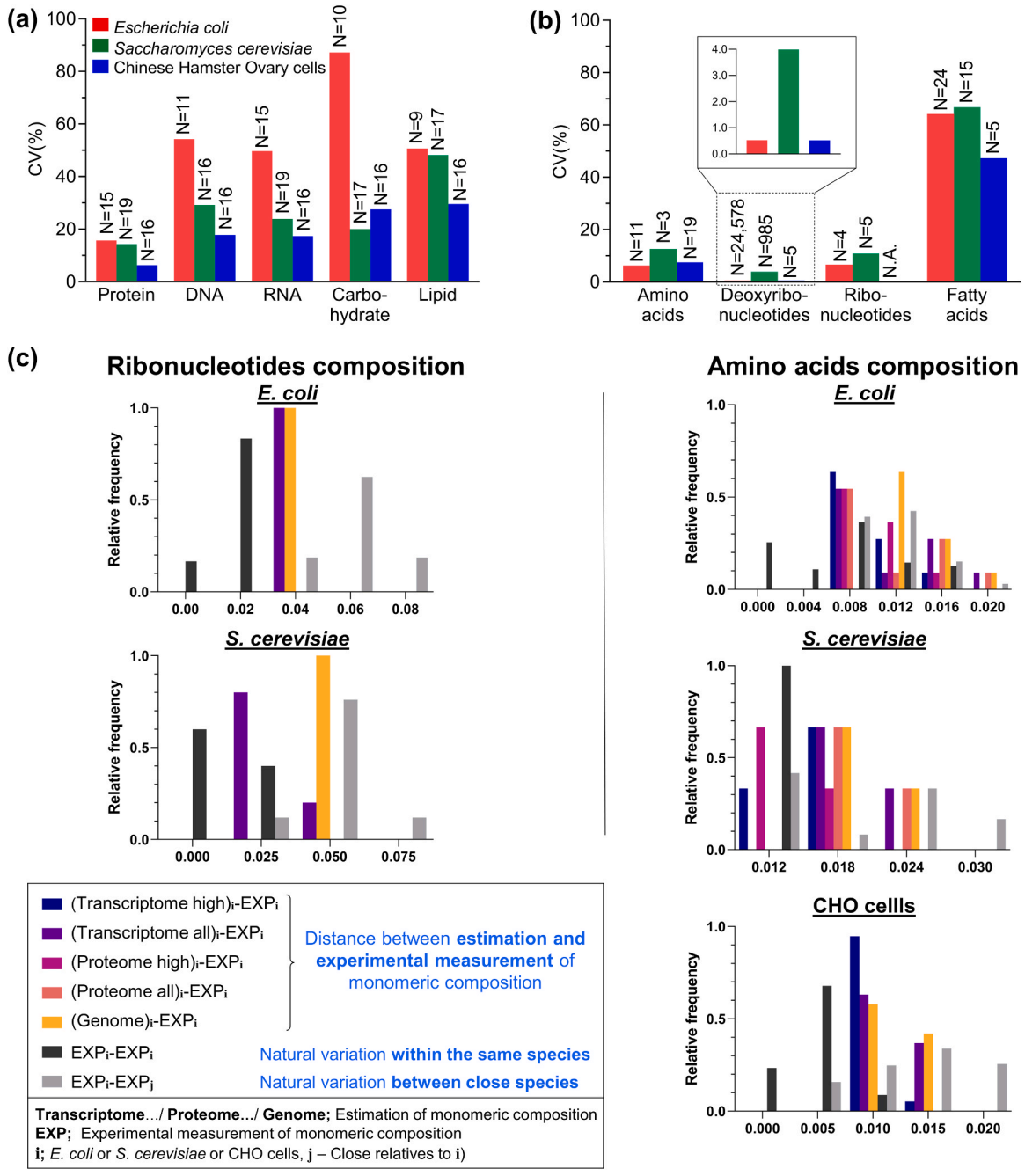

Fig. 1. Inherent variations in macromolecules and monomers compositions across various species and comparison of monomer composition from experiments and omics data. The coefficient of variations (CVs) of five macromolecules classes (a) and 4 monomer groups (b) is provided for three species: *E. coli*, *S. cerevisiae*, and CHO cells. N denotes the number of each biomass composition data. The bar graphs in (c) represent relative frequency distributions (y-axis) of the Euclidean expression distance (x-axis, root mean squared deviation, rmsd) between monomer (amino acid and ribonucleotide) compositions obtained from experiments and estimated values from various combinations of omics data.

storage and cellular membrane reconstitution, have substantially higher variability than other macromolecules in all three species. We also observed that all macromolecules showed increasing CVs in sequence of CHO cells, *S. cerevisiae*, and *E. coli* (Fig. 1**a**). Since prokaryotic cells are more adept in adapting to various niches through their faster exchange of various molecules with adjacent environments, their metabolism as well as cell growth can be stimulated much faster than eukaryotes, and thus could result in a faster turnover of cellular components, e.g., macromolecular composition [35]. Among monomers, while amino acids and nucleotides showed relatively less variability, fatty acids showed very large variations across different growth conditions (Fig. 1**b**). The high variability of fatty acids distributions could be attributed to the dynamic requirements of cellular membranes to adapt against perturbed environments, e.g., temperature, pH, salt, or dietary conditions [36–38]. In summary, the comparisons of biomass compositions from multiple conditions clearly showed that while monomers except fatty acids are relatively stable, distribution of macromolecules within unit biomass vary considerably.

It has been previously suggested that amino acid and nucleotide composition can be estimated from genome or transcriptome datasets [39,40]. Alternatively, previous studies have borrowed such data from phylogenetically close species: multiple yeast GEMs have used the nucleotide and/or monomer composition from *S. cerevisiae* [41–44]. We therefore evaluated which of these methods provide a better approximation for monomer composition. To do so, we first collected relevant multi-omics datasets: genome, transcriptome, and proteome for all three species (Supplementary File S2). We then estimated the ribonucleotide composition from whole transcriptome datasets and evaluated its "closeness" with the experimentally measured monomer composition across various conditions using the Euclidean distance metric – a measures of divergence between two datasets (see Methods). We also compared the experimentally measured and "-omics" estimated values with the data obtained from phylogenetically close species. The distance between experimentally measured ribonucleotide distributions (including the natural variation) and transcriptome data estimated ones was much lesser (indicated by a high relative frequency at low distances) compared to the distance calculated from phylogenetically close species (Fig. 1**c**). Next, we calculated the amino acid composition using genome, transcriptome, and proteome datasets and compared it with experimentally measured values. It is important to note that we utilized two combinations of transcriptome and proteome datasets: "all" and "highly expressed", respectively. Based on previous suggestions, we separately considered the highly expressed datasets (top 10%) of mRNA transcripts and proteins. This approach was adopted because highly expressed mRNA transcripts and proteins are known to contribute significantly to the overall proteome outputs. While the amino acid composition estimated from whole "-omics" datasets were distant from the experimentally measured ones (including natural variation), the distances from highly expressed transcripts/proteins were significantly less than that of experimental measurements from close organisms, and thus highlighting the estimation of amino acid composition from highly expressed transcripts/proteins is a good choice (Fig. 1**c**). Overall, our analysis indicates that "-omics" data, i.e., genome or transcriptome for ribonucleotide and highly expressed proteins or transcripts for amino acids, can be reliably used for deriving monomer composition within each macromolecule rather than borrowing it from close organisms.

3.2. Sensitivity of macromolecular/monomer composition in phenotype predictions

It has been earlier shown that FBA phenotype predictions are sensitive to any variations in individual biomass components. For the first time, Pramanik and Keasling showed that both growth rates and intracellular fluxes predicted from the metabolic model are sensitive to monomer compositions in *E. coli* [19]. Feist et al. later examined how growth rate predictions change in *E. coli* when the macromolecular composition is varied and observed no major influence [14]. Another study which performed similar sensitivity analysis in *S. cerevisiae* and reported biomass composition is indeed sensitive to flux predictions [15]. Such mixed observations in literature could be mainly due to lack of standardization in the sensitivity analyses performed; some studies focused only on the variations in intracellular fluxes while others focused on growth rate predictions, some studies analyzed the effect of changes in monomers while some other studies focused on macromolecular variations, etc. Therefore, we next examined the sensitivity of predicted *in silico* growth rates and intracellular metabolic fluxes upon varying both macromolecular and monomer compositions in all three species across diverse environments. Particularly, the sensitivity analysis was carried out under aerobic and anaerobic conditions in *E. coli*, in three different carbon sources (glucose, xylose and ethanol) in *S. cerevisiae* and three different cell lines in CHO cells by using the relevant GEMs (see Methods).

Initially, we examined the sensitivity of macromolecular and monomer composition on growth rate predictions. Our analysis indicated that protein compositions were the most sensitive among different macromolecules, while DNA was found to be the least sensitive, except *E. coli* (Supplementary Fig. S2). Such a trend is expected since proteins and lipids have the largest fraction (by mass) in dry cell weight while the composition of DNA is almost negligible. Interestingly, although protein fraction of *E. coli* was three times greater than RNA, the growth rate prediction in *E. coli* was more sensitive to RNA than protein. Our analysis also unraveled condition specific sensitivities of macromolecular composition. The sensitivity of protein predicted under aerobic condition was higher than that of anaerobic conditions in *E. coli*. Similarly, the sensitivities of protein and carbohydrate were lower in reduced substrate ethanol than that of glucose and xylose. Among monomers, only amino acid composition was observed to be relatively sensitive in growth rate predictions.

Subsequently, we explored the effect of macromolecular and monomer compositions on flux distribution of all the intracellular reactions (Supplementary Fig. S3). The flux sensitivity was quantified by comparing the predicted fluxes in the reference state, i.e., original biomass equation, and those in a perturbed state, i.e., modified biomass equation (see Methods). Similar to the variations in growth rates, macromolecular compositions were sensitive to reaction flux predictions than the monomers (Fig. 2). Among macromolecules, intracellular fluxes were highly sensitive to protein composition in all three species as more than 40% of overall sensitivity is accounted by them. This is expected since protein is the most abundant component in unit biomass, and the number of reactions required to synthesize the 20 proteinogenic amino acids constitutes a large metabolic network. Similarly, we noted lipids to be second most sensitive component, as they are usually second most abundant constituent in cells next to proteins in all three species despite their low abundance in unit biomass. Within monomers, amino acids had the largest sensitivity coefficients (but still negligible compared to macromolecules) while fatty acids had almost dispensable sensitivities. This is mainly because amino acids biosynthesis pathways are made up of diverse reactions across the metabolic network while all fatty acids of different chain lengths are synthesized using the same set of reactions within the same pathway. Finally, we noted CHO cells to be relatively more sensitive to all monomers than *E. coli* or *S. cerevisiae*.

3.3. Ensemble representation of biomass equation to account for natural variation in macromolecular and monomer components

By analyzing the natural variations and the sensitivity to flux predictions, we note that both these parameters could contribute to any potential errors in FBA. For example, a biomass component with high natural variability but low sensitivity, e.g., lipids, can result in more significant changes in flux prediction. On the other hand, components with small CV and high sensitivity, e.g., protein, should also be considered as critical since even a small measurement error may lead to

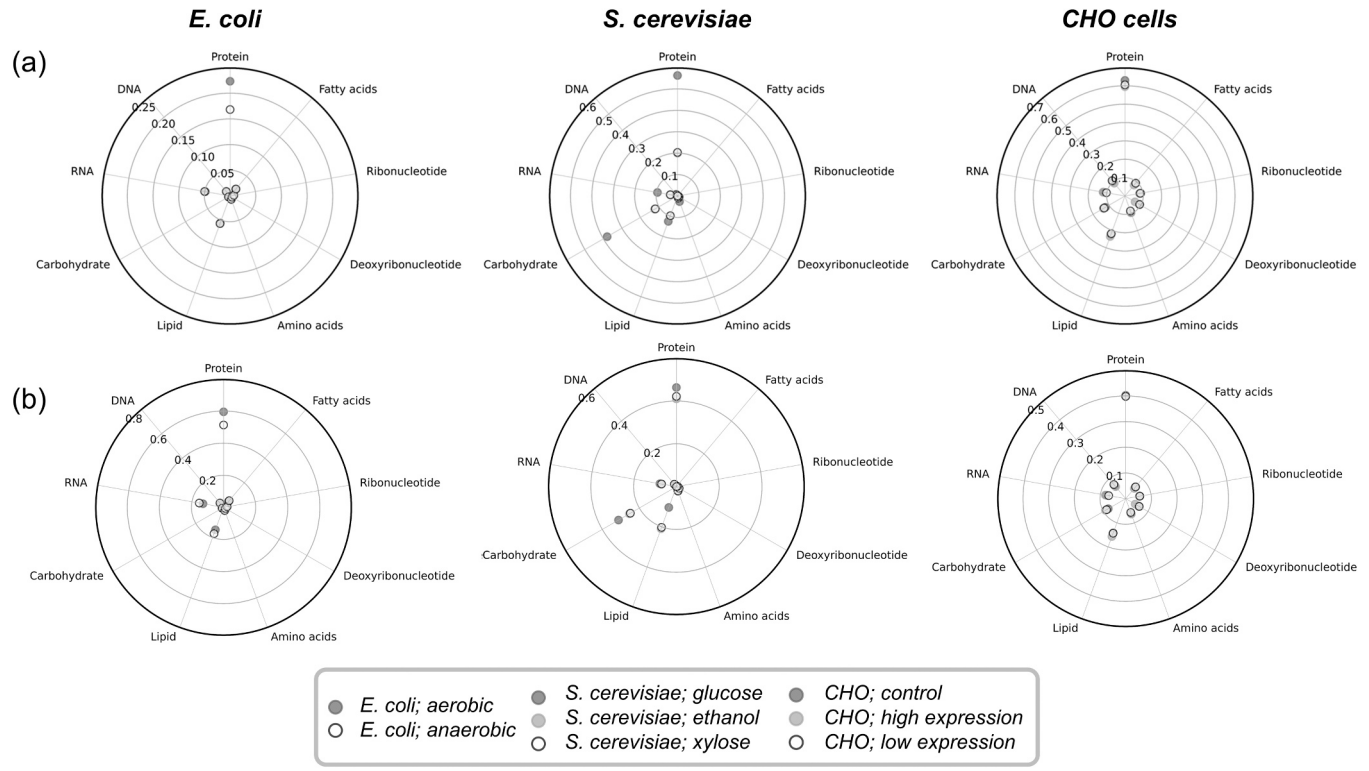

Fig. 2. Sensitivity of metabolic fluxes predicted by GEMs upon varying individual biomass components. The flux distributions of reference condition, i.e., default biomass equation, and the perturbed condition, i.e., biomass equation obtained by varying each biomass component by 25%, were compared to obtain absolute (a) and relative (b) sensitivities of biomass components under various conditions. A relative sensitivity was calculated by normalizing an absolute sensitivity by the sum of the absolute sensitivities of all macromolecules and monomers for every model condition.

large differences in flux predictions (Fig. 3). We thus propose to label certain biomass components as "critical", if it falls in any of the following three categories: 1) high CV & high sensitivity, 2) high CV & low sensitivity, and 3) low CV & high sensitivity. Based on this definition, we observed all the macromolecules were critical *in E. coli*. In *S. cerevisiae* and CHO cells, we observed all macromolecules except DNA and RNA to be critical. On the other hand, fatty acid was the only critical component observed in the monomer category.

To address both the natural variation in macromolecular components and potential experimental errors in the measurement of sensitive

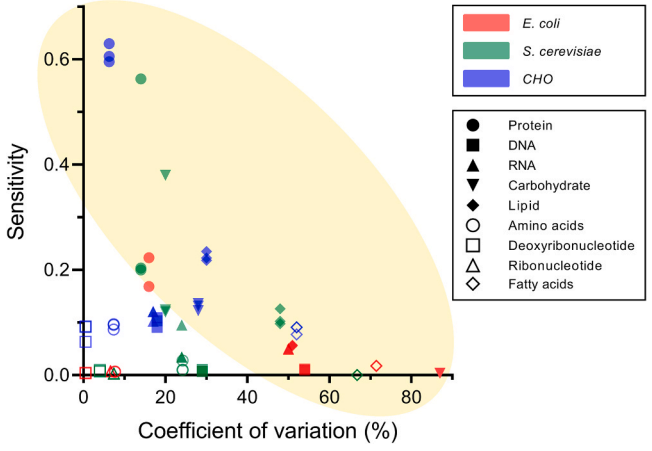

Fig. 3. Natural variation and the flux sensitivity of various biomass component. The distribution of biomass components based on their coefficient of variation (CV) on the x-axis and sensitivity on the y-axis. Biomass components exhibiting high CV or high sensitivity are identified as critical components within our method.

biomass constituents, we propose ensemble representation of biomass within the FBA framework (Supplementary Fig. S4). In this approach, the critical biomass components that either significantly vary or are highly sensitive to FBA results are sampled several times within a range, i.e., range of natural variation, and "n" number of biomass reactions with various combinations were generated and normalized to make up 1 g of cell in total (see Methods). Subsequently, FBA is implemented "n" times with each of the newly generated "n" biomass equation as the objective function. The distribution of fluxes obtained from "n" number of simulations are then analyzed to extract the plausible flux ranges for each reaction, similar to that in flux sampling approaches. Note that FBA with ensemble biomass is not a method with new types of constraints as in parsimonious FBA (pFBA) [30] or enzyme capacity constrained FBA (ecFBA) [34,45], rather it is an extension of FBA and can be incorporated into any constraint-based flux analysis method such as pFBA or ecFBA.

To demonstrate the utility of ensemble biomass representations in FBA, we implemented pFBA with ensemble biomass (pFBAwEB) and compared its performance to pFBA with a singular biomass equation which was derived from the average values of all macromolecular and monomer compositions collected. We used relevant experimentally measured growth phenotypic data of *E. coli* under various environmental conditions and genetic manipulations [46–49], and employed 5000 different biomass permutations to obtain 5000 flux solutions for each pFBAwEB simulation (see Supplementary Note and Supplementary Fig. S1 for details). In general, when evaluating the accuracy of growth rate predictions, pFBAwEB had smaller error values compared to pFBA. This improvement is attributed to the fact that pFBAwEB predicts a range of growth rates rather than a single value (two-sample Kolmogorov-Smirnov statistic 0.7 compared to that of 1, Fig. 4**a**).

To further evaluate the performance of pFBAwEB in estimating internal metabolic flux distributions, we compared it with other methods that address biomass composition uncertainty (BOFdat [21], BTW and HIP [22]) using the *E. coli* *i*ML1515 model [32], under aerobic and

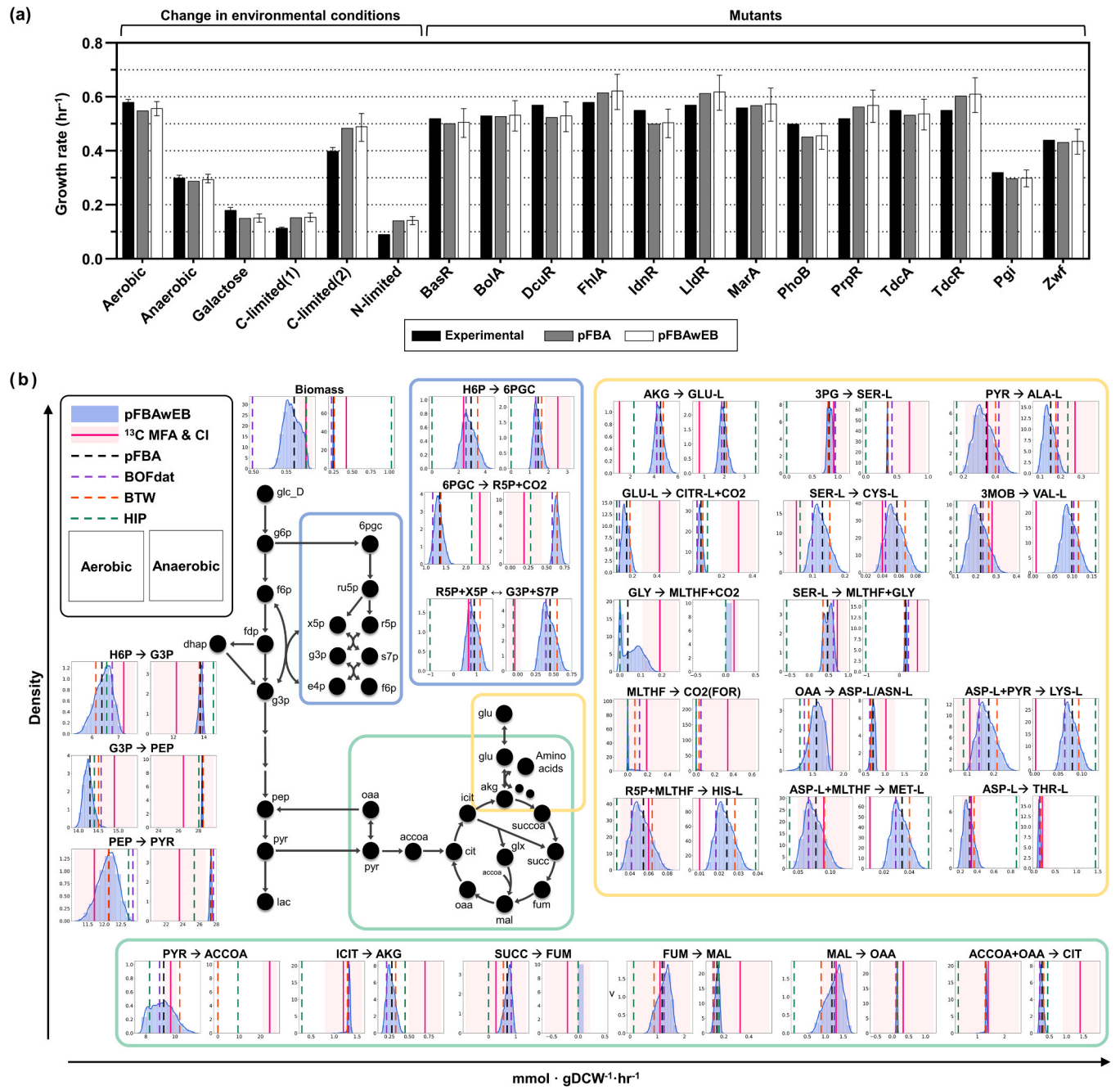

Fig. 4. Comparison of flux distributions with and without ensemble biomass. **(a)** Growth rate predictions by pFBA (grey bar) and pFBAwEB (white bar), compared to experimental measurements (black bar), across different environmental conditions and mutants. All predictions used the same reference biomass composition. **(b)** Detailed comparison of experimentally estimated fluxes (^{13}C MFA, solid magenta line with confidence interval range) with predicted fluxes by different methods: pFBAwEB (blue density plot), pFBA with the reference biomass composition (dashed black line), pFBA with BOFdat (dashed purple line) by Lachance et al. (2019), and pFBA with Biomass Trade-off Weighting (BTW) (dashed orange line) and Higher-dimensional-plane InterPolation (HIP) (dashed green line) by Schulz et al. [22]. The figure illustrates the central carbon metabolism, including the Pentose Phosphate Pathway (PPP, blue box), the TCA cycle (green box), and various amino acid metabolism pathways (yellow box), in *E. coli* under aerobic (left graph) and anaerobic (right graph) conditions.

anaerobic conditions. We used the same macromolecular composition as in pFBAwEB to derive a biomass equation for BOFdat. In the case of BTW, we generated different biomass equations for varying glucose and oxygen levels. For HIP, we obtained biomass equations for both aerobic and anaerobic conditions using four datasets of macromolecular composition, along with the corresponding glucose and oxygen uptake rates. The complete list of retrieved biomass equations for each method can be found in Supplementary File S3. We compared the resulting flux predictions from each method with experimentally measured flux data obtained through ^{13}C isotope labeling [50]. The flux predictions across glycolysis and fermentative pathways were more or less similar in all methods, while pFBAwEB, BOFdat and BTW performed better in simulating fluxes through anabolic pathways such as the Krebs cycle and amino acid biosynthetic reactions (Fig. 4**b**). The comparison of ^{13}C-flux analysis with pFBAwEB and other methods highlights two main strengths of biomass ensemble representations. First, pFBAwEB presents biologically relevant flux ranges by accounting for the uncertainty in the anabolic demands while other methods provide just a single solution which may have variable accuracy across the entire metabolic network. Second, despite using an ensemble biomass representation, the

predicted flux spans presented distinguishable patterns when comparing two different conditions, e.g., *E. coli* growing in aerobic vs. anaerobic conditions.

4. Discussion

Understanding the degree of natural variation in macromolecules and monomers is fundamental for mitigating the uncertainties in biomass equations and their influence in intracellular flux predictions using FBA. The current study investigated the natural variations in biomass of three representative organisms and gives a comprehensive overview on how cellular composition varies across organisms under different conditions. While most macromolecules varied significantly, the composition of monomers, except fatty acids, remained relatively stable across conditions. CV values of carbohydrate, DNA, and RNA were much greater than that of their corresponding monomers while protein and lipid showed less variability in most cases. Considering that the mean value and the CV are in an inverse relation, relatively small CV values of protein compared to other macromolecules are reasonable because protein accounts for the largest amount of cell weight (approx. 50% mass fraction). Unlike macromolecules, composition of monomers did not vary appreciably across different conditions, except fatty acids. Although perturbations in fatty acid composition brought almost negligible differences in growth rate predictions, its influence on internal fluxes was evident in CHO cells. These observations together clearly indicate the necessity of accurate measurements of various biomass components, particularly the macromolecular composition. To this end, Beck et al. (2018) reviewed multiple analytical techniques and suggested guidelines for accurate quantitative measurement of five major macromolecules [9].

In this study, we also examined how reliable it is to estimate biomass components from omics datasets and noted that it is better to use such estimates than borrowing this data from phylogenetically close organisms. Interestingly, even though omics data sets provide good estimate for monomer compositions, still, there exists difference between experimental and estimated values (Fig. 1**c**), particularly genome data estimates had poor concordance. This may stem from two main reasons. Firstly, not all the encoded genes get transcribed; only around 5% of the genome is transcribed into RNA at any given time [51]. Second, nucleotide, ribonucleotide and amino acid compositions are known to vary locally in individual members of a population. Nucleotide composition varies locally in different areas of a given genome [52], and consequently could result in mRNA transcripts with varying ribonucleotide composition. Similarly, amino acid composition differs across protein functional categories, probably related to consideration of translation rate [53]. Fortunately, even though monomer composition estimated from omics-data varies marginally from experimental values, it does not impact intracellular flux distribution significantly as the sensitivity of compositions of ribonucleotide and amino acids is negligible.

Not surprisingly, all biomass uncertainty methods rely on variation in macromolecular and monomer composition as fundamental input data. However, the key distinction among all the methods lies in their data requirements and prediction capabilities. BOFdat requires a single macromolecular composition as input and optionally incorporates gene essentiality data specific to the desired environmental condition in the final step to determine the suitable biomass equation. BTW, on the other hand, necessitates multiple biomass compositions obtained from measurements conducted under diverse environmental conditions. When we applied the BTW method using four distinct biomass equations derived from different macromolecular compositions with specific glucose and oxygen uptake rates, we observed that flux only flowed through one of the equations, deviating from the intended distribution across multiple biomass equations. HIP method relies on biomass composition data along with corresponding uptake rates of different substrates. For example, to simulate *E. coli* growth under aerobic and anaerobic conditions, multiple biomass composition datasets with corresponding glucose and oxygen uptake rates are required. Uniquely, pFBAwEB does not require any additional phenotypic data to simulate across the environmental conditions. Moreover, pFBAwEB provides flux spans representing the range of possible flux values for each reaction rather than single value, which allows us to better handle the inherent variability and uncertainty in cellular systems.

5. Conclusion

Overall, we conclude that among various components of biomass, all macromolecules except DNA, and fatty acids among monomers, vary considerably under different environments. Thus, these need to be accounted for its natural variation carefully while drafting the biomass equation. We also found that the omics data-estimated composition of monomers, is within the observed natural variation and could be reliably used. Based on such observations, we propose to use ensemble representations of biomass instead of a solitary equation in FBA, to account for both the natural variations and the intracellular flux sensitivities to cellular composition. While such ensemble biomass representations can be easily obtained for an organism with multiple cellular compositional measurements available, we suggest using the range of CVs reported in this study for each biomass component to derive an ensemble biomass equation when appropriate data is not available.

CRediT authorship contribution statement

Y.-M. Choi: Formal analysis, Implementation, Software, Visualization, Writing – original draft. **D.-H. Choi:** Software, Writing – original draft. **Y. Q. Lee:** Data curation. **L. Koduru:** Methodology, Writing – review & editing. **N. E. Lewis:** Conceptualization, Writing – review & editing. **M. Lakshmanan:** Conceptualization, Writing – review & editing, Supervision. **D.-Y. Lee:** Conceptualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Su Kyung Kim, Minouk Lee, Dongseok Kim for assistance in collecting relevant omics data. N.E.L. acknowledges funding support from National Institute of General Medical Sciences, United States [grant number R35 GM119850], National Institutes of Health, United States [grant number R35GM119850] and the Novo Nordisk Foundation [grant number NNF20SA0066621]. M. L. acknowledges funding support from A*STAR, Singapore through IAF-PP (HBMS) programme [grant number: H20H8a0003]. D.Y.L. acknowledges funding support from the Ministry of Health & Welfare, Republic of Korea [grant number HI19C1348], the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (iPET) through MAFRA programme [grant number 32136–05-1-HD050].

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.07.025.

References

[1] O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. Cell 2015;161:971–87. https://doi.org/10.1016/j.cell.2015.05.019.

[2] Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 2012;10:291–305. https://doi.org/10.1038/nrmicro2737.
[3] Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. Brief Bioinform 2009;10:435–49. https://doi.org/10.1093/bib/bbp011.
[4] Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol 2019;20:1–18. https://doi.org/10.1186/s13059-019-1730-3.
[5] Fang X, Lloyd CJ, Palsson BO. Reconstructing organisms in silico: genome-scale models and their emerging applications. Nat Rev Microbiol 2020;18:731–43. https://doi.org/10.1038/s41579-020-00440-4.
[6] Orth JD, Thiele I, Palsson BØ. What is flux balance analysis. Nat Biotechnol 2010;28:245–8. https://doi.org/10.1038/nbt.1614.
[7] Feist AM, Palsson BØ. The biomass objective function. Curr Opin Microbiol 2011;13:344–9. https://doi.org/10.1016/j.mib.2010.03.003.The.
[8] Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol 2007;3:119. https://doi.org/10.1038/msb4100162.
[9] Beck AE, Hunt KA, Carlson RP. Measuring cellular biomass composition for computational biology applications. Processes 2018;6:1–27. https://doi.org/10.3390/pr6050038.
[10] Volkmer B, Heinemann M. Condition-dependent cell volume and concentration of escherichia coli to facilitate data conversion for systems biology modeling. PLoS One 2011;6:e23126. https://doi.org/10.1371/JOURNAL.PONE.0023126.
[11] Széliová D, Ruckerbauer DE, Galleguillos SN, Petersen LB, Natter K, Hanscho M, et al. What CHO is made of: variations in the biomass composition of Chinese hamster ovary cell lines. Metab Eng 2020;61:288–300. https://doi.org/10.1016/j.ymben.2020.06.002.
[12] Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: Origins and consequences. Science (80-) 2010;330:1099–102. https://doi.org/10.1126/SCIENCE.1192588/SUPPL_FILE/SCOTT.SOM.PDF.
[13] Zuñiga C, Levering J, Antoniewicz MR, Guarnieri MT, Betenbaugh MJ, Zengler K. Predicting dynamic metabolic demands in the photosynthetic eukaryote chlorella vulgaris. Plant Physiol 2018;176:450–62. https://doi.org/10.1104/PP.17.00605.
[14] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 2007;3:121. https://doi.org/10.1038/msb4100155.
[15] Dikicioglu D, Kırdar B, Oliver SG. Biomass composition: the “elephant in the room” of metabolic modelling. Metabolomics 2015;11:1690–701. https://doi.org/10.1007/s11306-015-0819-2.
[16] Koduru L, Kim Y, Bang J, Lakshmanan M, Han NS, Lee D-Y. Genome-scale modeling and transcriptome analysis of Leuconostoc mesenteroides unravel the redox governed metabolic states in obligate heterofermentative lactic acid bacteria. Sci Rep 2017;7:15721. https://doi.org/10.1038/s41598-017-16026-9.
[17] Schinn SM, Morrison C, Wei W, Zhang L, Lewis NE. Systematic evaluation of parameters for genome-scale metabolic models of cultured mammalian cells. Metab Eng 2021;66:21–30. https://doi.org/10.1016/J.YMBEN.2021.03.013.
[18] Dinh HV, Sarkar D, Maranas CD. Quantifying the propagation of parametric uncertainty on flux balance analysis. Metab Eng 2022;69:26–39. https://doi.org/10.1016/j.ymben.2021.10.012.
[19] Pramanik J, Keasling JD. Effect of Escherichia coli biomass composition on central metabolic fluxes predicted by a stoichiometric model. Biotechnol Bioeng 1998;60:230–8.
[20] Xavier JC, Patil KR, Rocha I. Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. Metab Eng 2017;39:200–8. https://doi.org/10.1016/j.ymben.2016.12.002.
[21] Lachance J-C, Monk JM, Lloyd CJ, Seif Y, Palsson BO, Rodrigue S, et al. BOFdat: generating biomass objective function stoichiometric coefficients from experimental data. PLoS Comput Biol 2019. https://doi.org/10.1371/journal.pcbi.1006971.
[22] Schulz C, Kumelj T, Karlsen E, Almaas E. Genome-scale metabolic modelling when changes in environmental conditions affect biomass composition. PLOS Comput Biol 2021;17:e1008528. https://doi.org/10.1371/JOURNAL.PCBI.1008528.
[23] Lerman J a, Hyduke DR, Latif H, Portnoy V a, Lewis NE, Orth JD, et al. In silico method for modelling metabolism and gene product expression at genome scale. Nat Commun 2012;3:929. https://doi.org/10.1038/ncomms1928.
[24] Thiele I, Fleming RMT, Que R, Bordbar A, Diep D, Palsson BO. Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage. PLoS One 2012;7:e45635. https://doi.org/10.1371/journal.pone.0045635.
[25] O’Brien EJ, Lerman J a, Chang RL, Hyduke DR, Palsson BØ. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol 2013;9:693. https://doi.org/10.1038/msb.2013.52.
[26] O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189.
[27] Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res 2021:49. https://doi.org/10.1093/nar/gkaa967.
[28] Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002:30. https://doi.org/10.1093/nar/30.1.207.
[29] Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 2015:15. https://doi.org/10.1002/pmic.201400441.
[30] Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol 2010;6:390. https://doi.org/10.1038/msb.2010.47.
[31] Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 2011;6:1290–307. https://doi.org/10.1038/nprot.2011.308.
[32] Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes Escherichia coli traits. Nat Biotechnol 2017:35. https://doi.org/10.1038/nbt.3956.
[33] Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun 2019:10. https://doi.org/10.1038/s41467-019-11581-3.
[34] Yeo HC, Hong JK, Lakshmanan M, Lee D-Y. Enzyme capacity-based genome scale modelling of CHO cells. Metab Eng 2020;60:138–47. https://doi.org/10.1101/2019.12.19.883652.
[35] Sonea S, Mathieu LG. Major characteristics of the prokaryotic world. Prokaryotol Montr: Presses De l’Université De Montr 2000:29–72.
[36] Guerzoni ME, Lanciotti R, Cocconcelli PS. Alteration in cellular fatty acid composition as a response to salt, acid, oxidative and thermal stresses in Lactobacillus helveticus. Microbiology 2001;147:2255–64. https://doi.org/10.1099/00221287-147-8-2255/CITE/REFWORKS.
[37] Prakash O, Nimonkar Y, Shaligram S, Joseph N, Shouche YS. Response of cellular fatty acids to environmental stresses in endophytic Micrococcus spp. Ann Microbiol 2015;65:2209–18. https://doi.org/10.1007/S13213-015-1061-X/TABLES/4.
[38] Levental KR, Malmberg E, Symons JL, Fan YY, Chapkin RS, Ernst R, et al. Lipidomic and biophysical homeostasis of mammalian membranes counteracts dietary lipid perturbations to maintain cellular fitness. Nat Commun 2020;11:1–13. https://doi.org/10.1038/s41467-020-15203-1.
[39] Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 2010;5:93–121. https://doi.org/10.1038/nprot.2009.203.
[40] Santos S, Rocha I. A computation tool for the estimation of biomass composition from genomic and transcriptomic information. Adv Intell Syst Comput 2016;477:161–9. https://doi.org/10.1007/978-3-319-40126-3_17/COVER.
[41] Dias O, Pereira R, Gombert AK, Ferreira EC, Rocha I. iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. Biotechnol J 2014;9:776–90. https://doi.org/10.1002/biot.201300242.
[42] Mishra P, Park G-Y, Lakshmanan M, Lee H-S, Lee H, Chang MW, et al. Genome-scale metabolic modeling and in silico analysis of lipid accumulating yeast Candida tropicalis for dicarboxylic acid production. Biotechnol Bioeng 2016;113:1993–2004. https://doi.org/10.1002/bit.25955.
[43] Xu N, Liu J, Ai L, Liu L. Reconstruction and analysis of the genome-scale metabolic model of Lactobacillus casei LC2W. Gene 2015;554:140–7.
[44] Sohn SB, Graf AB, Kim TY, Gasser B, Maurer M, Ferrer P, et al. Genome-scale metabolic model of methylotrophic yeast Pichia pastoris and its use for in silico analysis of heterologous protein production. Biotechnol J 2010;5:705–15. https://doi.org/10.1002/biot.201000078.
[45] Sánchez BJ, Zhang C, Nilsson A, Lahtvee P, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol 2017;13:935. https://doi.org/10.15252/msb.20167411.
[46] Perrenoud A, Sauer U. Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in Escherichia coli. J Bacteriol 2005;187:3171–9. https://doi.org/10.1128/JB.187.9.3171-3179.2005/SUPPL_FILE/SUPPLEMENTAL_MATERIAL.XLS.
[47] Nanchen A, Schicker A, Sauer U. Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli. Appl Environ Microbiol 2006;72:1164–72. https://doi.org/10.1128/AEM.72.2.1164-1172.2006/SUPPL_FILE/SUPPLEMENTAL_TABLE_2.ZIP.
[48] Emmerling M, Dauner M, Ponti A, Fiaux J, Hochuli M, Szyperski T, et al. Metabolic flux responses to pyruvate kinase knockout in Escherichia coli. J Bacteriol 2002;184:152–64. https://doi.org/10.1128/JB.184.1.152-164.2002/ASSET/A58930CC-2B42-4EFD-83A9-EF6D2C00F488/ASSETS/GRAPHIC/JB0120763008.JPEG.
[49] Haverkorn Van Rijsewijk BRB, Nanchen A, Nallet S, Kleijn RJ, Sauer U. Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli. Mol Syst Biol 2011;7:477. https://doi.org/10.1038/MSB.2011.9.
[50] Chen X, Alonso AP, Allen DK, Reed JL, Shachar-Hill Y. Synergy between (13)C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in E. coli. Metab Eng 2011;13:38–48. https://doi.org/10.1016/J.YMBEN.2010.11.004.

[51] Frith MC, Pheasant M, Mattick JS. The amazing complexity of the human transcriptome. Eur J Hum Genet 2005:13. https://doi.org/10.1038/sj.ejhg.5201459.
[52] Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Dønsvik T, et al. Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genom 2010: 11. https://doi.org/10.1186/1471-2164-11-464.
[53] Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci USA 2002: 99. https://doi.org/10.1073/pnas.062526999.