Raman spectra and gene expression correspondences reveal global stoichiometry conservation architecture in cells

Ken-ichiro F. Kamei,[1*] Koseki J. Kobayashi-Kirschvink,[2,3] Takashi Nozoe,[1,4,5] Hidenori Nakaoka,[6] Miki Umetani,[1,4,5] Yuichi Wakamoto[1,4,5*]

[1]Department of Basic Science, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo 153-8902, Japan
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]Laser Biomedical Research Center, G. R. Harrison Spectroscopy Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Research Center for Complex System Biology, The University of Tokyo, Tokyo 153-8902, Japan
[5]Universal Biology Institute, The University of Tokyo, Tokyo 153-8902, Japan
[6]Department of Gene Mechanisms, Graduate School of Biostudies, Kyoto University, Kyoto 606-8501, Japan

*To whom correspondence should be addressed; E-mail: kenichiro_kamei@cell.c.u-tokyo.ac.jp (K.F.K.), cwaka@mail.ecc.u-tokyo.ac.jp (Y.W.).

Cellular Raman spectra are optical signatures reflecting comprehensive molecular composition of cells. It has been shown recently that condition-dependent cellular transcriptome profiles are inferable from the changes in cellular Raman spectra. However, it remains elusive whether the other layers of omics are also inferable from Raman spectra and what constraints in gene expression profiles enable such inference. Here we show that proteome profiles of

***Escherichia coli* across diverse conditions are predictable from cellular Raman spectra using simple ordinary linear regression. Employing methods in graph theory, we characterize core and peripheral stoichiometrically-conserved groups in the proteomes and reveal that major changes in cellular Raman spectra reflect low-dimensional proteome structure constructed from the global stoichiometry conservation relations. Mathematical analysis suggests a proportionality between stoichiometry conservation strength and condition-specificity of gene expression, which turns out universal across omics classes and organisms. Furthermore, stoichiometry conservation strength correlates with the essentiality and evolutionary conservation of each gene. These results provide an omics-level interpretation of cellular Raman spectra and uncover an architectural principle of global omics profiles.**

Biological cells change their gene expression and metabolic profiles globally depending on their biological contexts and external conditions. Although cells are composed of myriad biomolecules, the changes in their abundance are often correlated and could be restricted to low-dimensional manifolds (*1–6*). However, what kinds of biological principles and architectural constraints give rise to the low dimensionality of global molecular profiles of cells remains to be characterized.

Vibrational spectroscopy such as Raman spectroscopy might help us investigate such global constraints. Raman spectroscopy is a light scattering technique that measures energy shifts of light caused by interaction with sample molecules. Raman spectra are obtainable non-destructively even from biological samples such as individual cells. In principle, cellular Raman spectra are optical signatures conveying comprehensive molecular composition of targeted cells (*7–10*). Although molecular profiling of cells by direct spectral decomposition is usually intractable, we have previously shown that condition-dependent transcriptome profiles of cells

are inferable from cellular Raman spectra on the basis of their statistical correspondence (*11*). Notably, such Raman-spectroscopic transcriptome inference was based on dimension-reduced Raman spectra. Therefore, an inspection of the correspondence might hint at the principles and constraints leading to dimensional reduction of omics profiles.

In this report, we first reveal that, like transcriptomes, condition-dependent proteome profiles of *Escherichia coli* are also predictable from cellular Raman spectra. Next, we scrutinize the correspondence between Raman and proteome data, identifying several stoichiometrically-conserved groups (SCGs) whose expression tightly correlates with the major changes in cellular Raman spectra. We finally reveal a quantitative constraint between stoichiometry conservation strength and condition-specificity of gene expression levels, which turns out general across different omics layers and organisms.

RESULTS

Statistical correspondence between Raman spectra and proteomes To examine the correspondence between Raman spectra and proteomes in *E. coli*, we reproduced 15 environmental conditions for which absolute quantitative proteome data are already available (*12*) and measured Raman spectra of *E. coli* cells under those conditions (Fig. 1A and 1B). The culture conditions we adopted include (i) exponential growth phase in minimal media with various carbon sources, (ii) exponential growth phase in rich media, (iii) exponential growth phase with various stressors, and (iv) stationary phases (Table S1). We measured Raman spectra of single cells sampled from each condition and focused on the fingerprint region of biological samples, where the signals from various biomolecules concentrate (spectral range of 700–1800 cm^{-1}, Fig. 1B and Fig. S1). The Raman spectra were classified on the basis of the environmental conditions using a simple linear classifier, linear discriminant analysis (LDA) (*7, 8, 13*) (Fig. 1C–1E). This classifier calculates the most discriminatory axes by maximizing the ratio of between-condition

variance to within-condition variance and reduces the dimensions of Raman data to $m - 1$, where m is the number of conditions (see sections 1.1 and 2.1 in (*14*)).

The result shows that Raman spectral points from different environmental conditions are distinguishable in the LDA space (Fig. 1C–1E). For example, the first and second LDA axes clearly distinguish the conditions "LB" and "stationary3days" (Fig. 1C), and the third axis distinguishes "Glucose42C" and "GlycerolAA" (Fig. 1D). Notably, the first principal axis LDA1 correlated with growth rate significantly (Pearson correlation $r = 0.81 \pm 0.09$, Fig. S1). Visualizing the Raman LDA data by embedding them on a 2D plane using t-distributed Stochastic Neighbor Embedding (t-SNE) (*15*) confirms that the points for each condition form a distinctive cluster (Fig. 1F). These results imply that positions in the Raman LDA space reflect condition-dependent differences in cellular physiological states.

We next asked whether these Raman spectral differences in the LDA space could be linked to the different proteome profiles. To examine this, we hypothesized linear correspondence between the proteome vector $\hat{\mathbf{p}}_j$ and the low-dimensional Raman vector $\hat{\mathbf{r}}_j$ in condition j,

$$\hat{\mathbf{p}}_j = \mathrm{B} \cdot \begin{bmatrix} 1 \\ \hat{\mathbf{r}}_j \end{bmatrix}. \tag{1}$$

B is a matrix that connects $\hat{\mathbf{p}}_j$ and $\hat{\mathbf{r}}_j$. We calculated $\hat{\mathbf{r}}_j$ as the average of the low-dimensional LDA Raman data of single cells in condition j since the proteomes were measured for cell populations.

We conducted leave-one-out cross-validation (LOOCV) to verify this linear correspondence (Fig. 1G). We excluded one condition (here, j) as a test condition and estimated B as $\mathrm{B}^{\mathrm{est}}_{-j}$ by simple ordinary least squares (OLS) regression using the data of the rest of the conditions. We thereby estimated the proteome in condition j as $\hat{\mathbf{p}}^{\mathrm{est}}_j = \mathrm{B}^{\mathrm{est}}_{-j} \cdot \begin{bmatrix} 1 \\ \hat{\mathbf{r}}_j \end{bmatrix}$.

The proteome profile estimated using the first four major LDA axes (LDA1-LDA4) agreed well with the actual proteome data under most conditions (Fig. 1H and S2; see section 1.2

in (*14*) for the estimation with all the LDA axes). Changing the condition to exclude, we estimated the proteomes for all the 15 conditions and calculated the overall estimation error by the Euclidean distance $\sum_j \|\hat{\mathbf{p}}_j^{\mathrm{est}} - \hat{\mathbf{p}}_j\|^2$. The result shows that the overall estimation error is significantly small ($p = 0.00005$ by permutation test (*16–18*)). Adopting other distance measures does not change the conclusion (Table S2 and S3). These results therefore validate the assumption of linear correspondence between cellular Raman spectra and proteomes and confirm that condition-dependent changes in proteomes can be inferred from the corresponding low-dimensional Raman spectra.

Stoichiometry conservation of the proteins in the ISP COG class The regression matrix B considered above determines how the proteomes relate to the Raman LDA axes. Therefore, analyzing B provides some insights into constraints on condition-dependent changes in the proteomes.

The matrix B is represented as $\mathrm{B} = [\mathbf{b}_0\ \mathbf{b}_1\ \cdots\ \mathbf{b}_{m-1}]$, where $\mathbf{b}_k = (b_{1k}\ b_{2k}\ \cdots\ b_{nk})^\top$ is the $(k+1)$-th column of B ($0 \leq k \leq m-1$) and $n =$ 2,058 is the number of protein species in the proteome data. We first asked whether any shared features might exist in the coefficients of B depending on biological functions of corresponding proteins. We then classified the proteins according to functional annotations of Clusters of Orthologous Group (COG) classes (*19–21*) and found that, for the proteins belonging to the "information storage and processing" (ISP) COG class, their coefficients corresponding to different LDA axes are approximately proportional to those for the constant term; i.e., $b_{lk} \approx c_k b_{l0}$, where c_k is a constant common to the proteins in the ISP COG class (Fig. 2A). The ISP COG class contains various proteins involved in processing genetic information such as translation, transcription, DNA replication, and DNA repair (*12*). Simple calculations show that these proportionality relationships imply that the proteins in the ISP COG class conserves their mutual abundance ratios, i.e., stoichiometry, irrespective of en-

vironmental conditions (see section 1.3 in (*14*)).

Since this is an implication from the Raman-proteome correspondence, we next examined the stoichiometry conservation only with the proteome data, evaluating the expression levels with Pearson correlation coefficients for all the pairs of the conditions for each COG class (Fig. 2B and Fig. S3). For the proteins in the ISP COG class, the correlation coefficients were close to 1, whereas those in the other COG classes were significantly weaker depending on condition pairs. Therefore, stoichiometry conservation is stronger in the ISP COG class than in the other COG classes. Remarkably, neither shared transcription factors nor chromosome locations can account for the observed stoichiometry conservation of many protein pairs (Fig. 2C and 2D), implying multi-level regulation and coordination of their abundance.

We consulted other public quantitative proteome data of *Mycobacterium tuberculosis* (*22*), *Mycobacterium bovis* (*22*), and *Saccharomyces cerevisiae* (*23*) under environmental perturbations and consistently found strong stoichiometry conservation of the ISP COG class (Fig. S3). Furthermore, the same trend was observed for the genotype-dependent expression changes in *E. coli* proteomes (*12*) (Fig. S3).

Identifying stoichiometrically-conserved groups Inspired by the existence of a large class of proteins that conserves their stoichiometry, we considered a systematic way to extract stoichiometrically-conserved groups (SCGs) without relying on artificial functional classification of COG. Focusing only on the proteome data, we evaluated stoichiometry conservation by the cosine similarity of expression levels across conditions for all the pairs of proteins in the proteome (Fig. 3A) and extracted groups in each of which the component proteins exhibit coherent expression change patterns by setting a high threshold of cosine similarity (≥ 0.995, Fig. 3B).

The largest SCG (SCG 1) included many proteins in the ISP COG class (91 out of 191 SCG

1 members), such as ribosomal proteins and RNA polymerase, and also proteins in the other COG classes (Fig. 3B, Table S4). We call this largest SCG *homeostatic core*, as it constitutes the largest stoichiometry-conserving unit in cells. We found that the abundance of each protein in the homeostatic core (SCG 1) increased approximately linearly with the growth rate in each condition (Fig. 3C). This relationship is reminiscent of the "growth law": The total ribosomal contents for translation increase linearly with growth rate (*24–26*). The linear increase in the abundance of each protein in Fig. 3C indicates that the growth law is valid even at the single-gene level for a large class of ribosomal and non-ribosomal proteins in the homeostatic core (Fig. S4) (see section 3.1 in (*14*)).

Though not evenly distributed, the gene loci of the proteins in the homeostatic core are scattered throughout the chromosome (Fig. 3D). Therefore, localization of gene loci to a single or a small number of operons is not likely a cause of the observed stoichiometry conservation.

The proteins in the second largest SCG (SCG 2) are expressed at high levels in the fast growth conditions, especially in the "LB" condition (Fig. 3C). The SCG 2 includes many proteins in the metabolism COG class (21 out of 26 SCG 2 members) (Table S5), and their abundance increases approximately exponentially with growth rate (Fig. 3C). We also identified other condition-specific small SCGs, such as a group most expressed in the "GlycerolAA" condition (SCG 3) (Table S6), a group mainly expressed in the "Fructose" condition (SCG 4) (Table S7), and a group most expressed in the stationary phase conditions (SCG 5) (Table S8) (Fig. 3C).

Functional relevance of stoichiometry conservation To understand the overall strength of stoichiometry conservation of the proteins in the different SCGs, we calculated the sum of cosine similarity, $d_i = \sum_j \cos\theta_{\mathbf{p}_i\mathbf{p}_j}$, for each protein species i, where $\cos\theta_{\mathbf{p}_i\mathbf{p}_j}$ is cosine similarity between the expression level vectors of protein i and protein j, and the sum is taken over all

the protein species (see section 1.5 in (*14*)). We refer to d_i as "stoichiometry conservation centrality."

The proteins in the homeostatic core had high centrality scores (Fig. 4A). Therefore, these proteins tend to have more connections with other proteins in terms of stoichiometry conservation. On the other hand, the proteins in the condition-specific SCGs tend to have low centrality scores among all the proteins (Fig. 4A), which suggests that their stoichiometry conservation is localized within each SCG.

The stoichiometry conservation centrality is functionally relevant because it correlates with gene essentiality. Fractions of essential genes almost monotonically decrease with the ranks of centrality score (Fig. 4B and S5). We also noted that the stoichiometry conservation correlates with the evolutionary conservation of proteins (Fig. 4C and S5). Here, we used the number of orthologs in 150 biological species across the three domains determined by OrthoMCL-DB (*27*) as a proxy for the evolutionary conservation strength of each protein.

Comparable correlations of stoichiometry conservation centrality with gene essentiality and evolutionary conservation were also found for the *S. pombe* transcriptome data (Fig. S5). In addition, fractions of coding genes almost monotonically decreased with ranks of centrality score in the *S. pombe* data (Fig. S5). These results suggest that each protein's strength of stoichiometry conservation partly reflects its biological and functional importance in cells.

Revealing global stoichiometry conservation architecture of the proteomes with csLE To gain further insights into the linkage between stoichiometry conservation constraints and cellular Raman spectra, we next analyzed the proteomes using a method similar to Laplacian eigenmaps (LE) (*28*). We consider a symmetric matrix A with its (i, j) entry is $\cos\theta_{\mathbf{p}_i\mathbf{p}_j}$ (Fig. 4D). The entire proteome structure can be represented using the eigenvectors of normalized A. Major differences of this method from the ordinary LE are that we consider an edge for all node pairs

and that we adopt cosine similarity for weighting edges. This method places the proteins with higher cosine similarity closer in the resulting space (see sections 1.5 and 2.1 (*14*)); we call this method cosine similarity LE (csLE).

In this csLE space Ω_{LE}, the stoichiometry conservation centrality of the proteins decreased from center to periphery (Fig. 4E), which confirms that it indeed measures the extent to which each protein is close to the center in the entire stoichiometry conservation architecture. Furthermore, the proteins formed polyhedral distributions with the cluster of the proteins in the homeostatic core at the center and the clusters of the proteins in the other condition-specific SCGs at distinct vertices (Fig. 4F). This distribution is consistent with the fact that the condition-specific SCGs are the components whose expression patterns are distant from the homeostatic core and also between each other.

Representing the proteomes using the Raman LDA axes Given that the analysis of the Raman-proteome regression coefficients B initially led us to identify the SCGs and the stoichiometry conservation architecture in the proteome data, major changes in cellular Raman spectra characterized by LDA might reflect the expression changes of these coherently-expressed components.

The coefficients in the regression matrix B must satisfy $\frac{b_{ik}}{b_{i0}} = \frac{b_{jk}}{b_{j0}}$ for all k ($1 \leq k \leq m-1$) for the pair of protein i and protein j that conserve their stoichiometry (see sections 1.3 and 2.1 in (*14*)). Noting this property, we constructed another proteome space Ω_{B}, assigning each protein species i a coordinate $\left(\frac{b_{i1}}{b_{i0}}\ \frac{b_{i2}}{b_{i0}}\ \cdots\ \frac{b_{i(m-1)}}{b_{i0}}\right)$. As in Ω_{LE}, a pair of proteins with strong stoichiometry conservation are expected to position closely in this proteome space Ω_{B}.

We then found that the distribution of the proteins in Ω_{B} closely resembled the one in Ω_{LE} when visualized using the first few major axes (Fig. 4F and 5A). This similarity is nontrivial because Ω_{LE} is constructed only from the proteome data, whereas Ω_{B} depends on the Raman

LDA space.

We remark that each axis of Ω_{B} is directly linked to the corresponding Raman LDA axis. Consequently, the orthants in Ω_{B} where the condition-specific protein species reside agree with those in the Raman LDA space where the cellular Raman spectra under corresponding conditions reside (Fig. S9) (see sections 1.7 and 2.1 in (*14*)). Indeed, we find such orthant agreement between the proteins in the condition-specific SCGs (SCG 2–SCG 5) and those of the cellular Raman spectra in the LDA space (Fig. 5B and 5C). This straightforward correspondence between Ω_{B} and the Raman LDA space allows us to examine the relationship between changes in cellular Raman spectra and omics components' stoichiometry conservation architecture by comparing the two proteome structures in Ω_{B} and Ω_{LE}.

Omics-level interpretation of cellular Raman spectra and a quantitative constraint between expression generality and stoichiometry conservation centrality To understand rigorously what the similarity of the proteome structures in Ω_{B} and Ω_{LE} signifies (Fig. 5C and 5D), we clarified the mathematical relation between the coordinates of the proteins in these two spaces (Fig. 5E; see sections 2.1 and 2.2 in (*14*) for detail). We then characterized the two mathematical conditions that must be satisfied simultaneously (Fig. 5E).

The first condition is that major axes of the Raman LDA space and those of the proteome csLE space correspond (Fig. 5E). Consequently, cellular Raman spectra under a condition accompanying the expression of a condition-specific SCG must be significantly different from those under conditions with the expression of other condition-specific SCGs in a manner distinguishable by LDA. We verified this first condition with the data (Fig. S8).

The second condition is that the stoichiometry conservation centrality of each protein species d_i must be proportional to $g_i := \|\mathbf{p}_i\|_1/\|\mathbf{p}_i\|_2$, where $\|\mathbf{p}_i\|_1$ and $\|\mathbf{p}_i\|_2$ are the L1 and L2 norms of the expression level vector of protein i across conditions (Fig. 5E).

g_i can be interpreted as *the expression generality score*. When g_i is large, the protein i is expressed generally across conditions; when g_i is small, this is expressed only under specific conditions (Fig. S7) (see section 1.9 in (*14*)). Therefore, the proportionality between d_i and g_i indicates that the proteins with high stoichiometry conservation centrality must be expressed non-specifically to conditions. We also verified this condition with the data, confirming that it is indeed satisfied (Fig. 6A and S8).

The spread of the points from the proportionality diagonal line of the *E. coli* proteome data in Fig. 6A were found related to the growth rate under the condition where each protein is expressed the most (see section 2.2 in (*14*) for a detailed analysis on the origin of the deviation). Consequently, one can envisage a growth-rate-dependent expression pattern of each protein on the basis of its relative position in this g_i-d_i plot (Fig. 6B and 6C). For example, both BamB and YqjD are expressed non-specifically to the conditions with nearly identical expression generality scores. However, BamB is expressed at higher levels under fast growth conditions, whereas YqjP is expressed at higher levels under slow growth conditions due to their relative positions to the proportionality line. Similar growth rate dependence is observed for PaaE and DgoA, but with more prominent condition-specificity because these proteins are characterized by their low expression generality scores. These growth-rate-dependent deviation patterns might hint at a new growth law that governs *the total relative expression changes* of the proteome components (see section 2.2 in (*14*) for detailed discussion).

Generality We also examined the generality of the aforementioned two conditions using the Raman and proteome data of *E. coli* strains with different genotypes (BW25113, MG1655, and NCM3722) under two culture conditions (*12*) and the Raman and transcriptome data of *S. pombe* under 10 culture conditions (*11*). Applying csLE to the omics data, we again found similar omics structures between Ω_{LE} and Ω_{B} when visualized using the first few major axes,

with homeostatic cores at the centers and condition-specific SCGs at the vertices (Fig. S10 and S11).

Proportionality between stoichiometry conservation centrality and expression generality score was also confirmed in both additional datasets (Fig. S6). We further used publicly available quantitative proteome data of *M. tuberculosis*, *M. bovis*, and *S. cerevisiae* (*22, 23*) to examine this relation and confirmed that the proportinality is universally held (Fig. S6 and S12). Almost no deviation from the proportionality line existed in the *S. cerevisiae* proteome data measured for the cells in different media but cultured in chemostats with an identical dilution rate (thus, identical growth rate), which is consistent with the result of *E. coli* in which the deviations were related to the growth rate differences.

DISCUSSION

A Raman spectrum obtained from a single cell is a superposition of the spectra of all of its constituent biomolecules. Therefore, cellular Raman spectra potentially contain rich information on essential state differences in targeted cells. The fact that both transcriptomes and proteomes are inferable from cellular Raman spectra, as demonstrated in this and previous (*11*) studies, endorses this speculation. The detailed analyses of the relationship between Raman and omics data have identified functionally-relevant constraints on omics changes and provided an interpretation of cellular Raman spectra. Specifically, it has been revealed that major changes in cellular Raman spectra distinguishable by LDA reflect the changes in omics profiles under the constraints of stoichiometry conservation. This correspondence would help us interpret global changes in cellular Raman spectra by translating them into the differences in omics profiles.

We remark that linearity in our formulation enabled us to find the rigorous connection between the two omics spaces Ω_{B} and Ω_{LE}. Unlike the original LE, we adopted cosine similarity as weights of edges between all node pairs to measure expression stoichiometry conservation of

proteins. This modification was indispensable in terms of interpretation; relative proximity of positions in Ω_{LE} reflects the strength of stoichiometry conservation. We also remark that simple principal component analysis (PCA) applied to the normalized *E. coli* proteome data also finds a similar low-dimensional proteome structure (Fig. S5) (see section 1.10 in (*14*)). Therefore, besides interpretability, omics structures in Ω_{LE} might reflect dominant relationships among omics components commonly characterized by several methods of omics representation.

Stoichiometry conservation is plausibly crucial for cellular functions and physiology. For example, the enzymes involved in evolutionarily-conserved metabolic pathways conserve their stoichiometry across microorganism species despite their diverse transcriptional and translational rates (*29*). It is suggested that stoichiometry conservation is achieved by optimizing the metabolic flux for fast growth (*30*). Furthermore, a ribosome-targeting antibiotic causes an imbalance of ribosomal proteins and growth arrest in *E. coli*, but the balance is restored alongside growth recovery through physiological adaptation (*31*). These results suggest that breakage of stoichiometric balance among core components could impose significant fitness cost.

It is intriguing to ask how cells conserve stoichiometry among the components in each SCG. Especially, the homeostatic core (SCG 1) contains many components whose gene loci are scattered throughout the genome. It is known that both transcriptional and translational negative auto-regulation contributes to controlling the stoichiometry of many ribosomal proteins (*32–37*). The genes for the ribosomal proteins are scattered in multiple operons and co-regulated with many other non-ribosomal proteins, such as RNA polymerase subunits, translation initiation/elongation factors, and transmembrane transporters (*38*). Therefore, the stoichiometry-conserving mechanisms established for ribosomes might be partially exploited for the stoichiometry conservation within the homeostatic core.

The proportionality between stoichiometry conservation centrality and expression generality score suggests that proteins with high stoichiometry conservation centrality govern basal

cellular functions required under any conditions. In fact, both essential genes and evolutionary-conserved genes are enriched in the omics fractions with high centrality scores. On the contrary, proteins of low centrality scores might have been acquired in later stages of the evolution and exploited to survive or increase fitness under specific conditions. Such hierarchy in the stoichiometry conservation centrality among core and peripheral processes might promote the adaptability of cells since cells can respond to diverse environments without restructuring a large body of the functional homeostatic core. This architectural principle in omics might underlie the robusstness and adaptability of biological cells.

Fig. 1. Estimation of proteomes from Raman spectra. **(A)** The experimental design. We cultured *E. coli* cells under 15 different conditions and measured single cells' Raman spectra. We then examined the correspondence between the measured Raman spectra and the absolute quantitative proteome data reported by Schmidt *et al.* (*12*). **(B)** Representative Raman spectra from single cells, one from the "Glucose" condition, and the other from the "LB" condition. The fingerprint region and representative peaks are annotated. **(C to E)** Cellular Raman spectra in LDA space. The dimensionality of the spectra is reduced to 14 $(= 15 - 1)$. Each point represents a spectrum from a single cell, and each ellipse shows the 95% concentration ellipse for each condition. Their projection to the LDA1-LDA2 plane (C), the LDA1-LDA3 plane (D), and the LDA1-LDA4 plane (E) are shown. **(F)** Visualization of the 14 dimensional LDA space embedded in two dimensional space with t-SNE. **(G)** The scheme of leave-one-out cross validation. The Raman and proteome data of one condition (here j) are excluded, and the matrix B is estimated using the data of the rest of the conditions as $\mathrm{B}^{\mathrm{est}}_{-j}$. The proteome data under the condition j is estimated from the Raman data $\hat{\mathbf{r}}_j$ with $\mathrm{B}^{\mathrm{est}}_{-j}$ and compared with the actual data to calculate estimation errors. **(H)** Comparison of measured and estimated proteome data. The plot for the "Glucose" condition is shown as an example. Each dot corresponds to one protein species. The straight line indicates $x = y$. Proteins with negative estimated values are not shown.

Fig. 2. A stoichiometrically-conserved protein group identified by an analysis of the Raman-proteome coefficient matrix. **(A)** Scatterplots of Raman-proteome transformation coefficients. The horizontal axes are constant terms (b_0) in all the plots. The vertical axis is coefficients for LDA1 (b_1), LDA2 (b_2), LDA3 (b_3), or LDA4 (b_4) in each plot. ISP COG class proteins are indicated in yellow. Yellow solid straight lines are least squares regression lines passing through the origins for ISP proteins. Insets are enlarged views of area around the ori-

gins. In this figure, we used the average of $\mathrm{B}^{\mathrm{est}}_{-i}$ as an estimate of B. **(B)** Similarity of expression patterns between culture conditions for each COG class. We divided the proteome into COG classes (*20, 21*), and calculated Pearson's correlation coefficient of expression patterns for all the combinations of culture conditions. Since the data are from 15 conditions, there are 105 (= 15·14/2·1) points for each COG class in the graph. The box-and-whisker plots summarize the distributions of the points. The lines inside the boxes denote the medians, the top and bottom edges of the boxes do the 25th percentiles and 75th percentiles, respectively. The numbers of protein species are 376 for the Cellular Processes and Siginaling COG class, 354 for the ISP COG class, and 840 for the Metabolism COG class. See Fig. S3 for the evaluation with Pearson correlation coefficient of log abundances and with cosine similarity. Fig. S3 also contains figures directly showing expression-level changes of different protein species across conditions for each COG class. **(C)** Examples of stoichiometry-conserving proteins in the ISP COG class. The horizontal axis represents the abundance of RplF under 15 conditions and the vertical axis represents those of several ISP COG class proteins. These proteins are also contained in the *homeostatic core* defined later (see Fig. 3). The solid straight lines are linear regression lines with intercept of zero. **(D)** Examples of abundance ratios of non-ISP COG class proteins. The horizontal axis represents the abundance of RplF under 15 conditions and the vertical axis represents those of compared non-ISP COG class proteins. Crp belongs to the Cellular Processes and Signaling COG class; the other proteins belong to the Metabolism COG class. In both (C) and (D), we selected the proteins expressed from distant loci on the chromosome. All sigma factors participating in the regulation of the proteins examined in (C) and (D) are listed to the right of the gene name legends. All transcription factors known to regulate multiple genes listed here are shown in the right diagrams. Arrows show activation; bars indicate inhibition; and squares represent that a transcription factor activates or inhibits depending on other factors. The information on gene regulation and functions was obtained from EcoCyc (*38*) in August 2022. The

error bars are standard errors calculated by using the data of (*12*). The inset shows the positions of the genes on the *E. coli* chromosome determined based on ASM75055v1.46 (*39*). No genes are in the same operon.

Fig. 3. Extracting SCGs from proteome data. (A) Quantifying stoichiometry conservation by cosine similarity. We consider an expression vector for each protein species whose elements represent its abundance under different conditions. The cosine similarity between the expression vectors of two protein species becomes nearly 1 when they conserve mutual stoichiometry strongly across conditions, whereas lower than 1 when their expression patterns are incoherent. **(B)** Extracted SCGs. We extracted proteins with high cosine similarity relationships. Each node represents a protein species. An edge connecting two nodes represents that the expression patterns of the two connected protein species have high cosine similarity exceeding a threshold of 0.995. Proteins that have no edge with the other proteins are not shown. The largest and the second largest protein groups, which we refer to as SCG 1 and SCG 2, respectively, are indicated by shaded polygons. **(C)** Expression patterns of the extracted SCGs. The horizontal and vertical axes represent growth rate and protein abundance, respectively. Line-connected points represent expression-level changes of different protein species across conditions. The inset for SCG 2 shows the total abundances of SCG 2 proteins with a log-scaled vertical axis. Error bars are standard errors. **(D)** The gene loci of the homeostatic core (SCG 1) proteins on the chromosome. Magenta dots are nodes (genes), and gray lines are edges (high cosine similarity relationships). We determined the gene loci based on ASM75055v1.46 (*39*).

Fig. 4. A proteome structure characterized by global stoichiometry conservation relationships. (A) Distributions of stoichiometry conservation centrality values for all the proteins (gray), the homeostatic core (SCG 1) proteins (magenta), and the proteins belonging to the

other SCGs (cyan). **(B)** Correlation between stoichiometry conservation centrality and gene essentiality. The proportion of essential genes within each class of stoichiometry conservation ranking is shown. The list of essential genes was downloaded from EcoCyc (*38*). **(C)** Correlation between stoichiometry conservation centrality and evolutionary conservation. The strength of evolutionary conservation of each protein species was estimated by the number of orthologs found in the OrthoMCL core species (*27*). The distributions of the number of orthologs were compared between the high and low stoichiometry conservation centrality fractions (threshold, 1300). The high centrality fraction tends to have more orthologs (one-sided Brunner-Munzel test, $p = 1.78 \times 10^{-10}$). **(D)** Representation of the proteomes as a graph. A node corresponds to a protein species, and the weight of an edge is taken as the cosine similarity between the expression vectors of the two connected protein species. The matrix A can specify the whole graph. Note that the diagonal elements of A are ones, which were introduced just for simplicity. **(E)** Cosine similarity LE (csLE) structure in a three-dimensional space. Each dot represents a different protein species and is color-coded on the basis of its stoichiometry conservation centrality value. We selected the axes considering the structural similarity to the Raman-based proteome structure in Ω_{B} (See Fig. 5). **(F)** The csLE structure in a three-dimensional space. The views from two different angles are shown. Each gray dot represents a different protein species. The proteins belonging to each SCG are indicated with distinct markers.

Fig. 5. Raman-based proteome structure and its similarity to stoichiometry-based proteome structure. **(A)** Proteome structure determined by Raman-proteome coefficients visualized in a three-dimensional space. The views from two different angles are shown. Each gray dot represents a protein species. The proteins belonging to each SCG are indicated with distinct markers. We note that SCGs are defined without referring to Raman data (Fig. 3). **(B–D)** Similarity among the distribution of LDA Raman spectra (B), the proteome structure determined

by Raman-proteome coefficients (C), and the proteome structure determined by stoichiometry conservation (D). **(E)** Mathematical relation between the coordinates of the proteins in Ω_{B} (C) and Ω_{LE} (D). The two conditions, one between $\mathbf{b}_0$ and $\mathbf{b}_0^{\mathrm{est}}$ (cyan), and the other with Θ (magenta), must hold for the similarity between the two proteome structures (yellow), as described in the gray box. $\overset{\star}{\propto}$ denotes column-wise proportionality.

Fig. 6. Proportionality between stoichiometry conservation centrality and expression generality. **(A)** Relationships between stoichiometry conservation centrality (d_i) and expression generality (g_i). Each gray dot represent a protein species. The proteins belonging to each SCG are indicated with distinct markers. The dashed lines are $y = n, x = 1, \sqrt{m}$ $(n = 2058, m = 15)$. The solid lines represent $y = \left\{\left(\sum_{j=1}^{n} d_j\right)/m\right\}^{1/2} x$ (see section 2.2 in (*14*)). The deviation of a point from the solid line is related to the growth rate under the condition where each protein is expressed the most. **(B)** The same plot as (A) in black and white. Overlaid red circles indicate proteins featured in (C). **(C)** Expression patterns of the proteins indicated by red circles in (B) across conditions. The condition differences are shown by the growth rate differences on the horizontal axes. The arrangement of the plots for the proteins corresponds to their relative positions in (B).

Acknowledgments

We thank Matthias Heinemann and Silke Bonsing-Vedelaar for detailed information on the *E. coli* culture conditions; Doeke R. Hekstra, Tetsuya J. Kobayashi, Takafumi Miyamoto, John Russell, and Ian Hunt-Isaak for reading the manuscript and providing critical comments; Kunihiko Kaneko and the members of the Wakamoto Lab for discussion. This work was supported by JST CREST Grant Number JPMJCR1927 (Y.W.); JST ERATO Grant Number JPMJER1902 (Y.W.); JSPS KAKENHI Grant Numbers 19J22448 (K.F.K.) and 21K20672 (T.N.).

References

1. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proceedings of the National Academy of Sciences* **95**, 14863 (1998).

2. E. Segal, *et al.*, *Nature Genetics* **34**, 166 (2003).

3. S. Bergmann, J. Ihmels, N. Barkai, *Physical Review E* **67**, 031902 (2003).

4. G. Heimberg, R. Bhatnagar, H. El-Samad, M. Thomson, *Cell Systems* **2**, 239 (2016).

5. S. Biswas, *et al.*, *Nature Communications* **8** (2017).

6. L. Keren, *et al.*, *Molecular Systems Biology* **9**, 701 (2013).

7. R. Goodacre, *et al.*, *Microbiology* **144**, 1157 (1998).

8. Y.-S. Huang, T. Karashima, M. Yamamoto, T. Ogura, H.-o. Hamaguchi, *Journal of Raman Spectroscopy* **35**, 525 (2004).

9. T. Ichimura, *et al.*, *PLoS ONE* **9** (2014).

10. A. Germond, *et al.*, *Communications Biology* **1**, 85 (2018).

11. K. J. Kobayashi-Kirschvink, *et al.*, *Cell Systems* **7**, 104 (2018).

12. A. Schmidt, *et al.*, *Nature Biotechnology* **34**, 104 (2016).

13. T. De Bie, N. Cristianini, R. Rosipal, *Handbook of geometric computing* (Springer, 2005), pp. 129–167.

14. Supplementary materials.

15. L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008).

16. R. A. Fisher, *The design of experiments* (Oliver And Boyd; Edinburgh; London, 1937).

17. E. J. Pitman, *Supplement to the Journal of the Royal Statistical Society* **4**, 119 (1937).

18. B. Phipson, G. K. Smyth, *Statistical Applications in Genetics and Molecular Biology* **9** (2010).

19. R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).

20. R. L. Tatusov, *et al.*, *BMC Bioinformatics* **4**, 1 (2003).

21. M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, *Nucleic Acids Research* **43**, D261 (2015).

22. O. T. Schubert, *et al.*, *Cell Host & Microbe* **18**, 96 (2015).

23. P.-J. Lahtvee, *et al.*, *Cell Systems* **4**, 495 (2017).

24. F. C. Neidhardt, B. Magasanik, *Biochimica et Biophysica Acta* **42**, 99 (1960).

25. M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, T. Hwa, *Science* **330**, 1099 (2010).

26. H. Bremer, P. P. Dennis, *EcoSal Plus* **3** (2008).

27. F. Chen, A. J. Mackey, C. J. Stoeckert Jr, D. S. Roos, *Nucleic Acids Research* **34**, D363 (2006).

28. M. Belkin, P. Niyogi, *Neural Computation* **15**, 1373 (2003).

29. J.-B. Lalanne, *et al.*, *Cell* **173**, 749 (2018).

30. J.-B. Lalanne, G.-W. Li, *Elife* **10**, e69222 (2021).

31. Y. Koganezawa, M. Umetani, M. Sato, Y. Wakamoto, *Elife* **11**, e74486 (2022).

32. M. Nomura, J. L. Yates, D. Dean, L. E. Post, *Proceedings of the National Academy of Sciences* **77**, 7084 (1980).

33. D. Dean, J. L. Yates, M. Nomura, *Nature* **289**, 89 (1981).

34. M. Kaczanowska, M. Rydén-Aulin, *Microbiology and Molecular Biology Reviews* **71**, 477 (2007).

35. C. Portier, M. Grunberg-Manago, *Translational Regulation of Gene Expression 2* (Springer, 1993), pp. 23–47.

36. L. V. Aseev, A. A. Levandovskaya, L. S. Tchufistova, N. V. Scaptsova, I. V. Boni, *RNA* **14**, 1882 (2008).

37. B. Roy, *et al.*, *Communications Biology* **3**, 1 (2020).

38. I. M. Keseler, *et al.*, *Nucleic Acids Research* **45**, D543 (2017).

39. K. L. Howe, *et al.*, *Nucleic Acids Research* **48**, D689 (2020).

Fig. 1

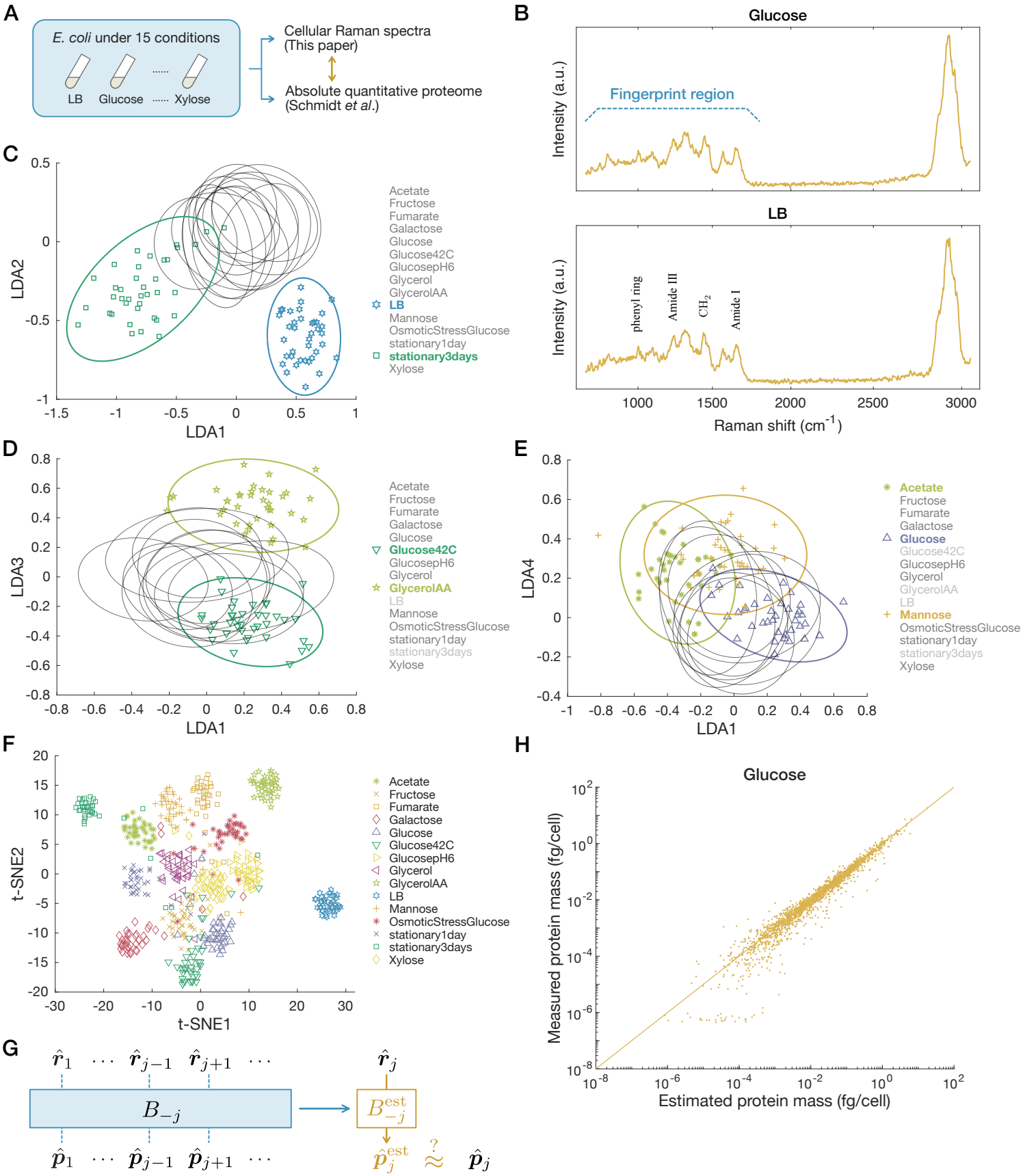

A
E. coli under 15 conditions
LB
Glucose
Xylose
Cellular Raman spectra
(This paper)
Absolute quantitative proteome
(Schmidt et al.)
B
Glucose
Fingerprint region
Intensity (a.u.)
LB
phenyl ring
Amide III
CH2
Amide I
Raman shift (cm-1)
C
LDA1
LDA2
Acetate
Fructose
Fumarate
Galactose
Glucose
Glucose42C
GlucosepH6
Glycerol
GlycerolAA
LB
Mannose
OsmoticStressGlucose
stationary1day
stationary3days
Xylose
D
LDA1
LDA3
E
LDA1
LDA4
F
t-SNE1
t-SNE2
G
H
Glucose
Measured protein mass (fg/cell)
Estimated protein mass (fg/cell)

Fig. 2

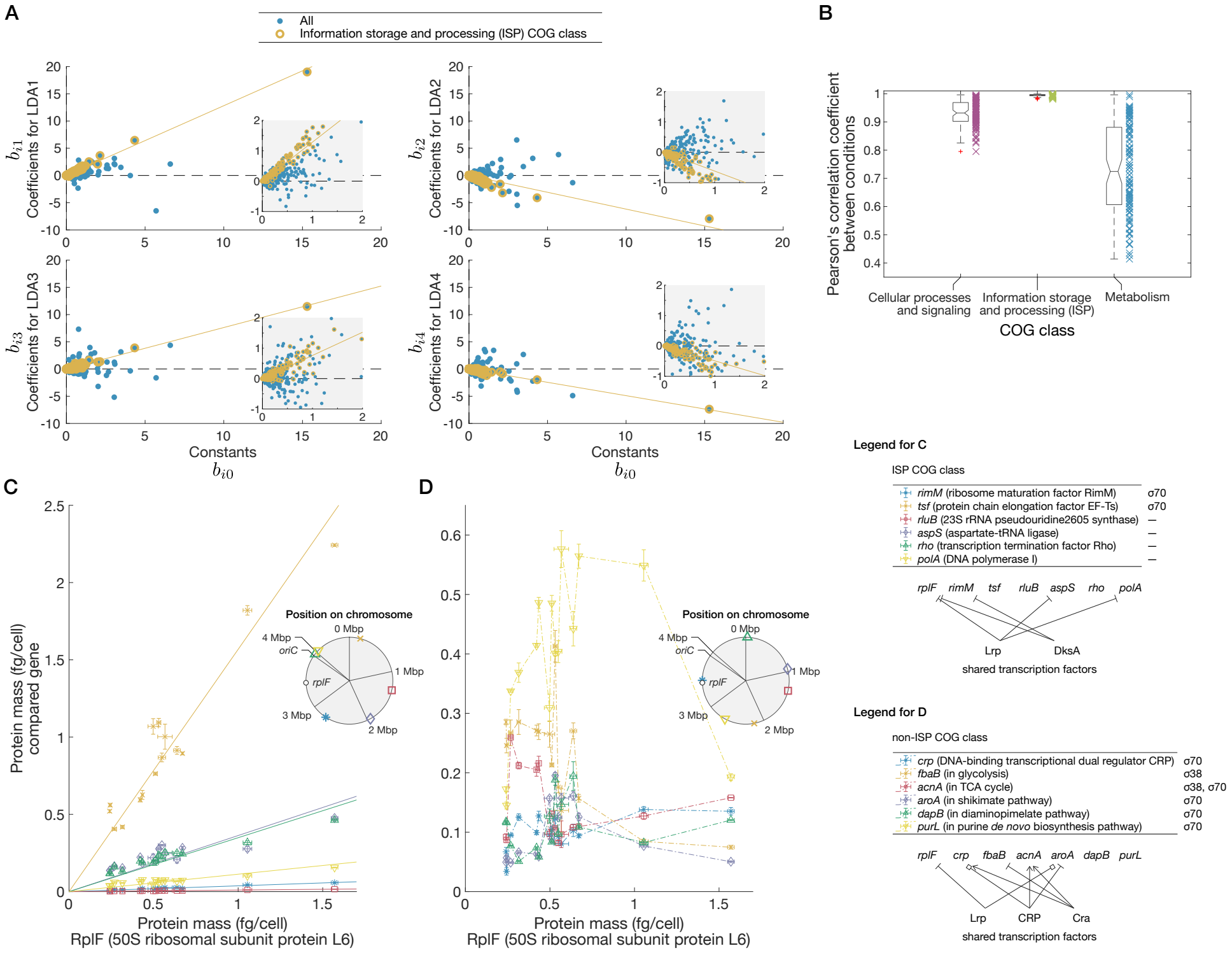

A
All
Information storage and processing (ISP) COG class
Coefficients for LDA1
Coefficients for LDA2
Coefficients for LDA3
Coefficients for LDA4
Constants
B
Pearson's correlation coefficient between conditions
Cellular processes and signaling
Information storage and processing (ISP)
Metabolism
COG class
C
Protein mass (fg/cell) compared gene
Protein mass (fg/cell)
RplF (50S ribosomal subunit protein L6)
Position on chromosome
0 Mbp
1 Mbp
2 Mbp
3 Mbp
4 Mbp
oriC
rplF
D
Legend for C
ISP COG class
rimM (ribosome maturation factor RimM) σ70
tsf (protein chain elongation factor EF-Ts) σ70
rluB (23S rRNA pseudouridine2605 synthase) —
aspS (aspartate-tRNA ligase) —
rho (transcription termination factor Rho) —
polA (DNA polymerase I) —
rplF rimM tsf rluB aspS rho polA
Lrp DksA
shared transcription factors
Legend for D
non-ISP COG class
crp (DNA-binding transcriptional dual regulator CRP) σ70
fbaB (in glycolysis) σ38
acnA (in TCA cycle) σ38, σ70
aroA (in shikimate pathway) σ70
dapB (in diaminopimelate pathway) σ70
purL (in purine de novo biosynthesis pathway) σ70
rplF crp fbaB acnA aroA dapB purL
Lrp CRP Cra
shared transcription factors

Fig. 3

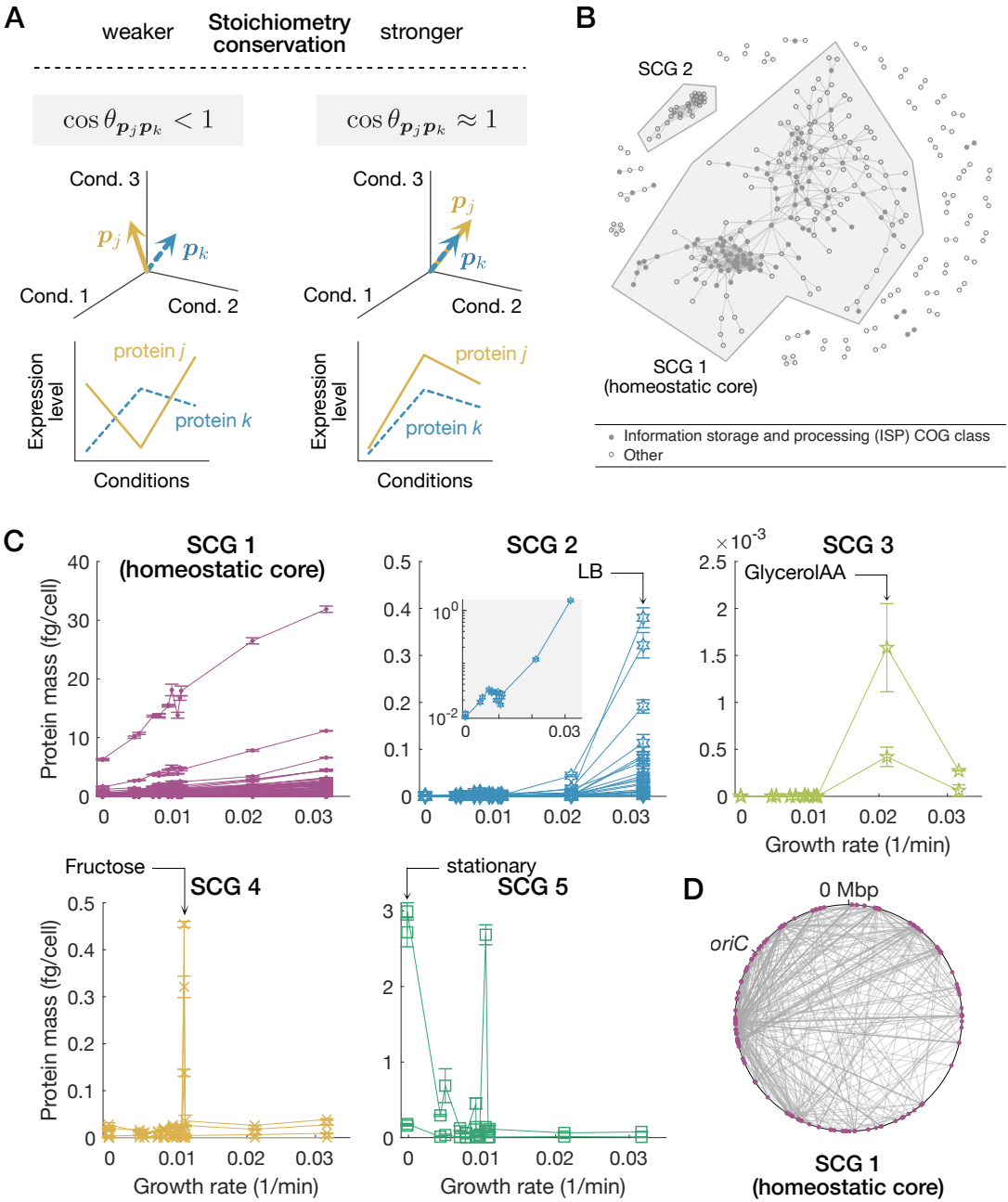

A
weaker
Stoichiometry conservation
stronger
$\cos\theta_{\boldsymbol{p}_j\boldsymbol{p}_k} < 1$
$\cos\theta_{\boldsymbol{p}_j\boldsymbol{p}_k} \approx 1$
Cond. 3
Cond. 1
Cond. 2
$\boldsymbol{p}_j$
$\boldsymbol{p}_k$
Expression level
protein j
protein k
Conditions
B
SCG 2
SCG 1
(homeostatic core)
Information storage and processing (ISP) COG class
Other
C
SCG 1
(homeostatic core)
SCG 2
LB
SCG 3
GlycerolAA
Protein mass (fg/cell)
Growth rate (1/min)
Fructose
SCG 4
stationary
SCG 5
D
0 Mbp
oriC
SCG 1
(homeostatic core)

Fig. 4

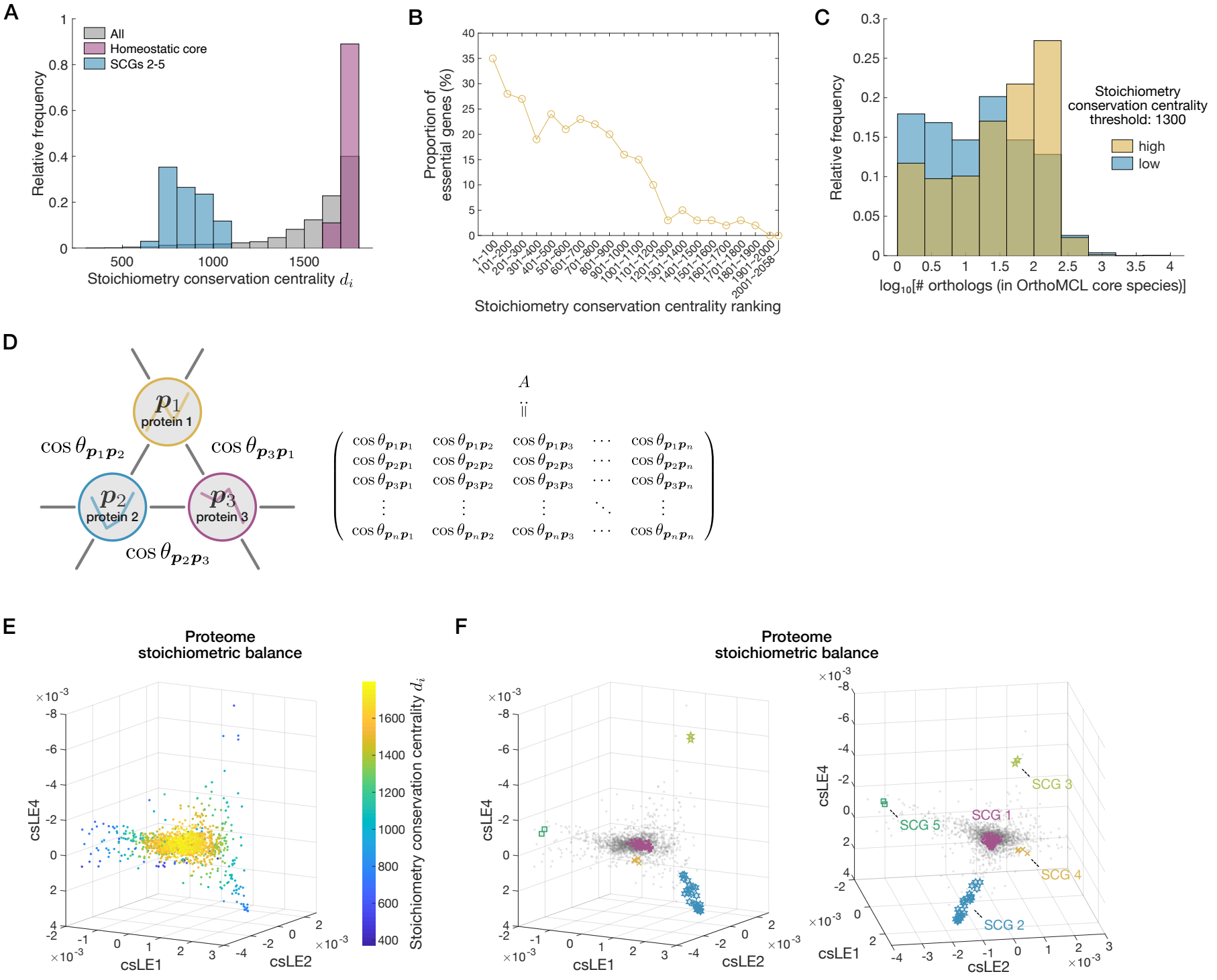
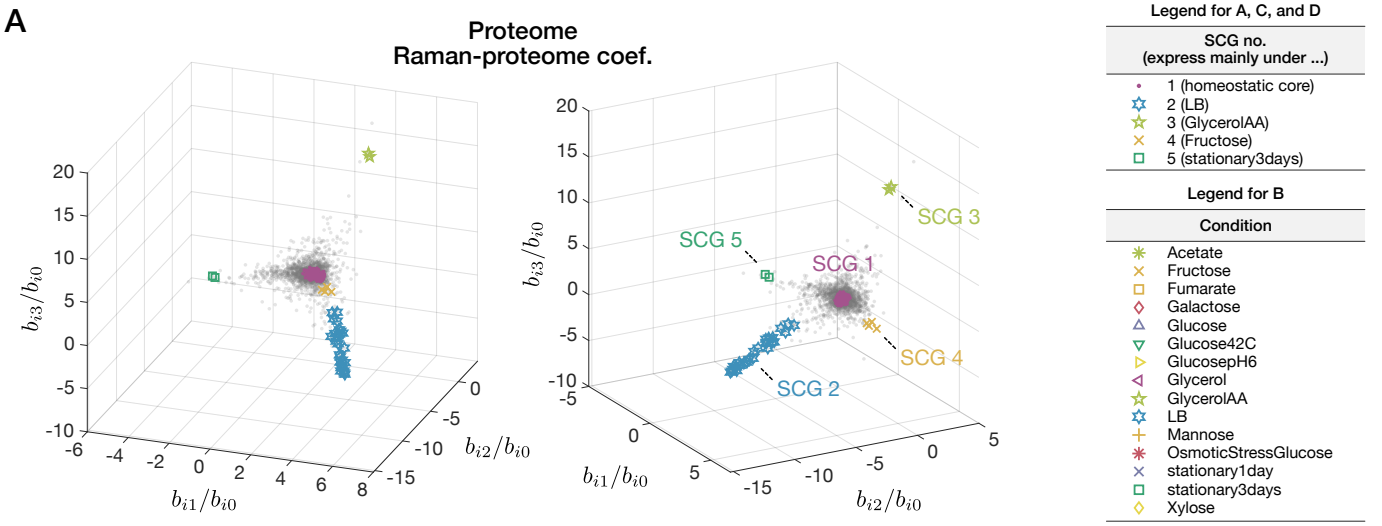
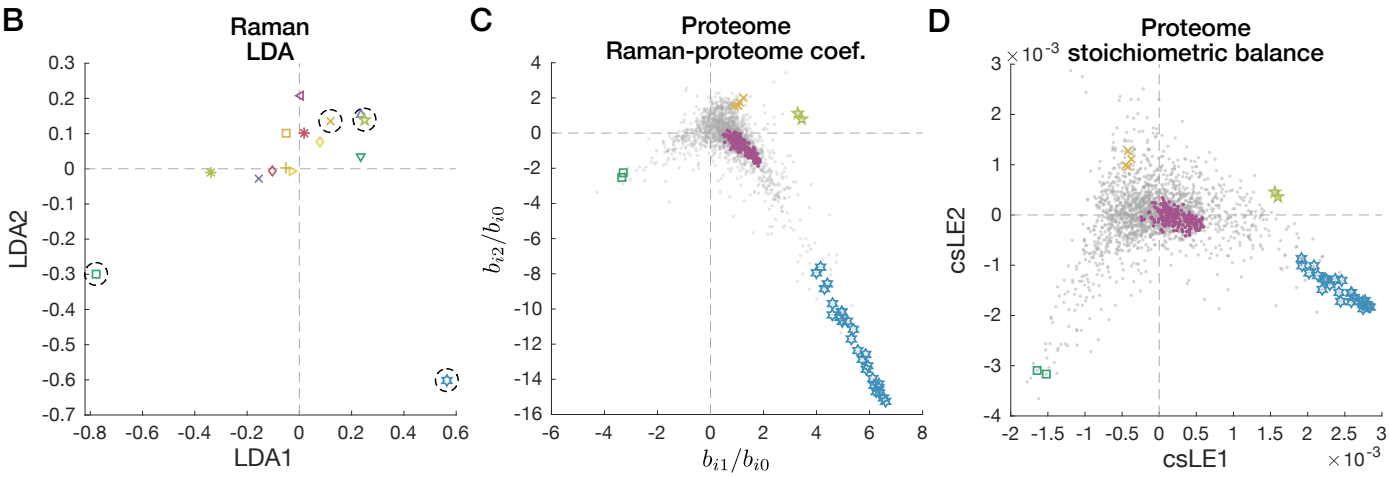

A
All
Homeostatic core
SCGs 2-5
Relative frequency
Stoichiometry conservation centrality d_i
B
Proportion of essential genes (%)
Stoichiometry conservation centrality ranking
C
Relative frequency
Stoichiometry conservation centrality threshold: 1300
high
low
log10[# orthologs (in OrthoMCL core species)]
D
protein 1
protein 2
protein 3
$\cos\theta_{p_1p_2}$
$\cos\theta_{p_3p_1}$
$\cos\theta_{p_2p_3}$
A
E
Proteome stoichiometric balance
csLE1
csLE2
csLE4
Stoichiometry conservation centrality d_i
F
Proteome stoichiometric balance
SCG 1
SCG 2
SCG 3
SCG 4
SCG 5

Fig. 5





E

LDA Raman-based matrix which contains protein coordinates in Ω_{B}

Stoichiometry-based matrix which contains protein coordinates in Ω_{LE}

$$\sqrt{m}\,\mathrm{diag}\left(\boldsymbol{b}_0\right) B_E^{\mathrm{norm}} \Sigma_{R_E}^{\mathrm{norm}} = \mathrm{diag}\left(\boldsymbol{b}_0^{\mathrm{est}}\right) B_E^{\mathrm{est,norm}} \Sigma_{\mathrm{LE}} \Theta$$

... Supplementary materials equation (2.138)

Vector which contains constant terms for Raman-proteome transformation

Vector which contains estimated constant terms for Raman-proteome transformation

Orthogonal transformation (e.g. rotation)

$B_E^{\mathrm{norm}} \overset{\star}{\propto} B_E^{\mathrm{est,norm}} \quad \Leftrightarrow$

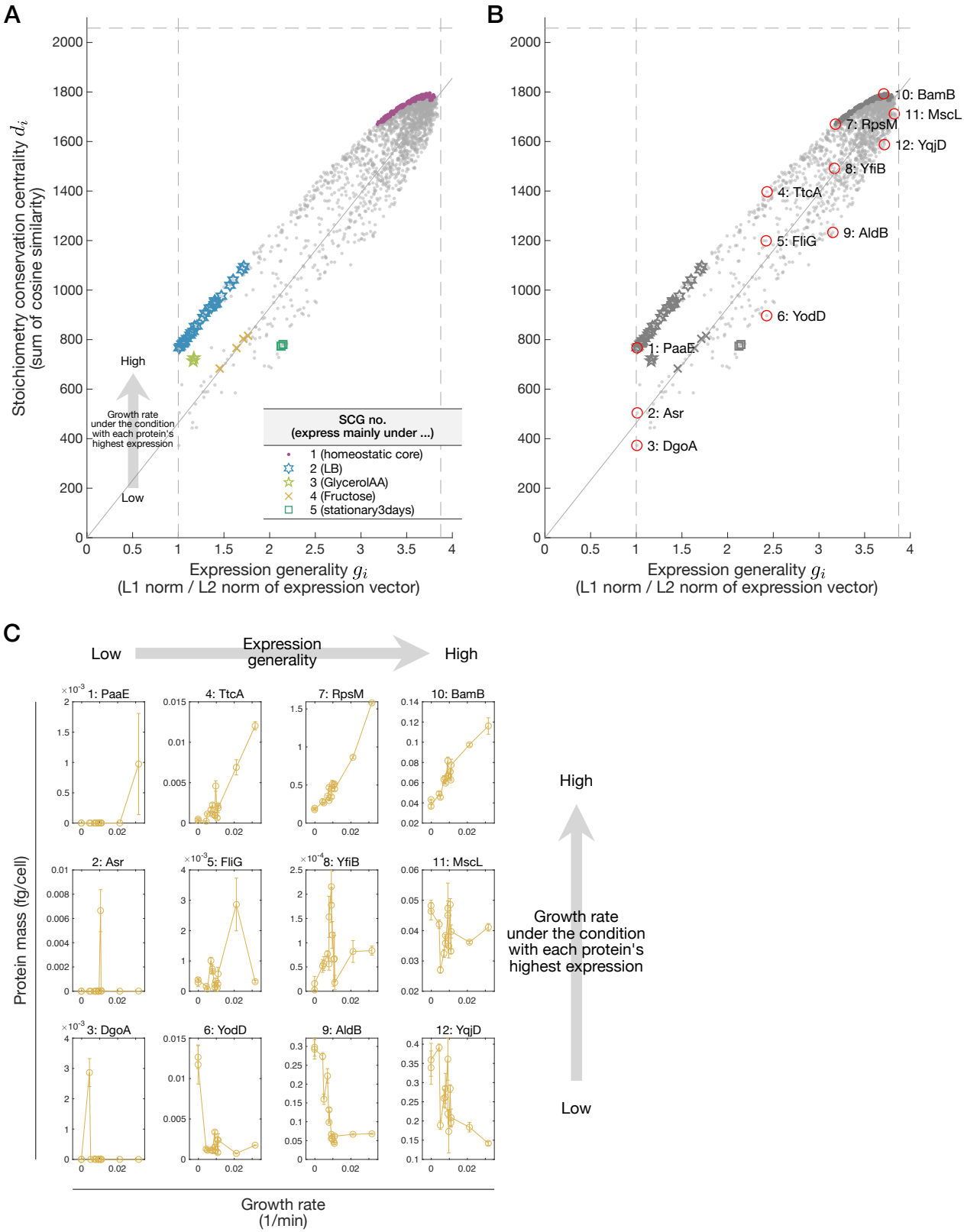
$\boldsymbol{b}_0 \propto \boldsymbol{b}_0^{\mathrm{est}}$

→ Proportionality between
d_i (stoichiometry conservation centrality) and
g_i (expression generality)
(Fig. 6 and supplementary figures)

$\Theta = I$ (identity transformation)

→ Commonality between
LDA axes (distinction of cellular Raman spectra) and
csLE axes (stoichiometric characterization of gene expression patterns)
(Supplementary figures)

Fig. 6


A
B
C
Stoichiometry conservation centrality d_i
(sum of cosine similarity)
Expression generality g_i
(L1 norm / L2 norm of expression vector)
High
Low
Growth rate
under the condition
with each protein's
highest expression
SCG no.
(express mainly under ...)
1 (homeostatic core)
2 (LB)
3 (GlycerolAA)
4 (Fructose)
5 (stationary3days)
1: PaaE
2: Asr
3: DgoA
4: TtcA
5: FliG
6: YodD
7: RpsM
8: YfiB
9: AldB
10: BamB
11: MscL
12: YqjD
Low
Expression
generality
High
Protein mass (fg/cell)
Growth rate
(1/min)