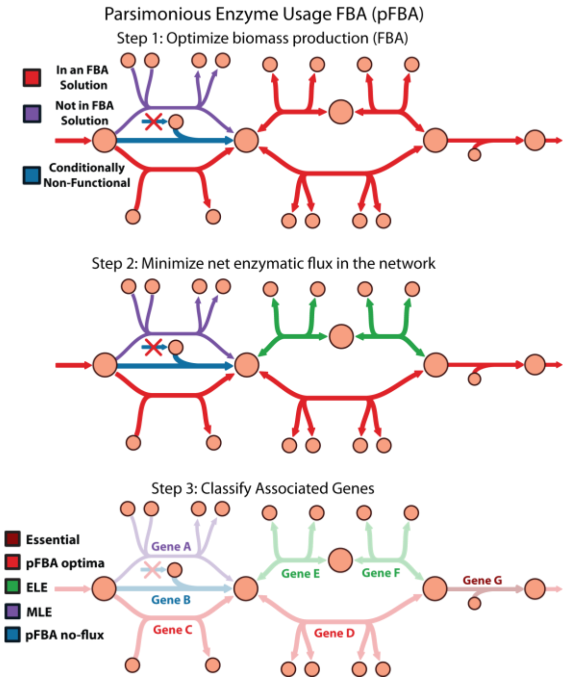
Supplement to: Omic Data from Evolved Strains are Consistent with Computed Optimal Growth States

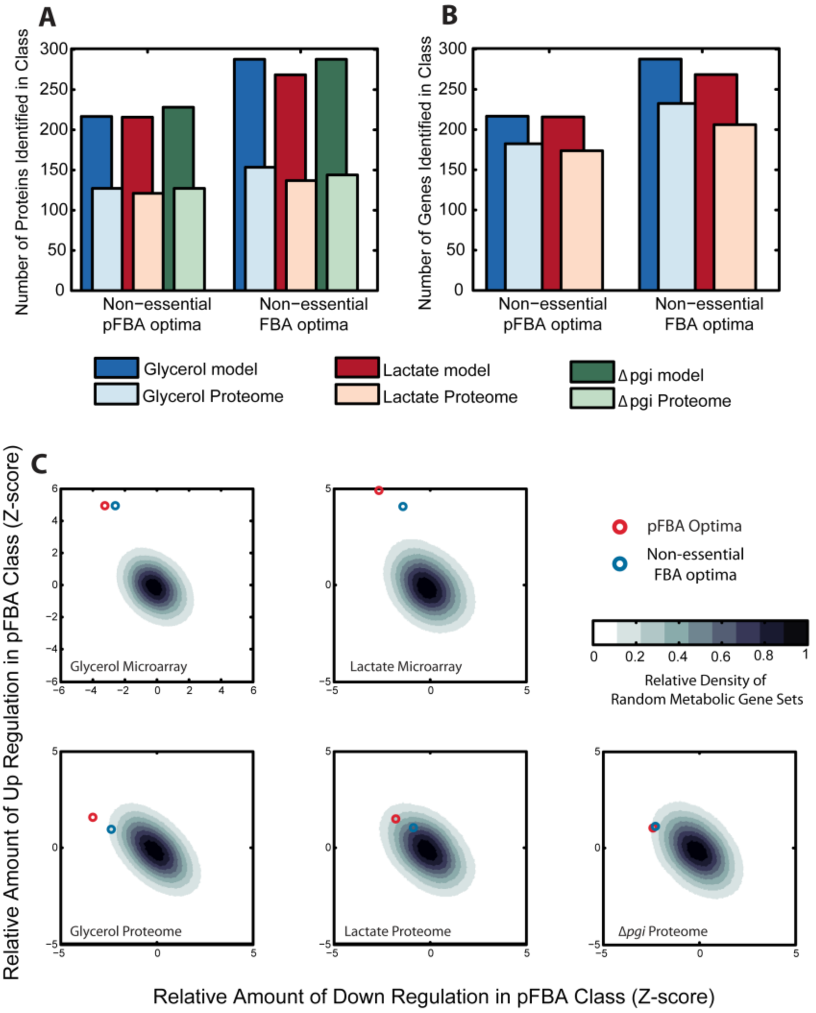
Nathan E. Lewis, Kim K. Hixson, Tom M. Conrad, Joshua A. Lerman, Pep Charusanti, Ashoka D. Polpitiya, Joshua N. Adkins, Gunnar Schramm , Samuel O. Purvine, Daniel Lopez-Ferrer, Karl K. Weitz, Roland Eils, Rainer König, Richard D. Smith, Bernhard Ø. Palsson

Contents

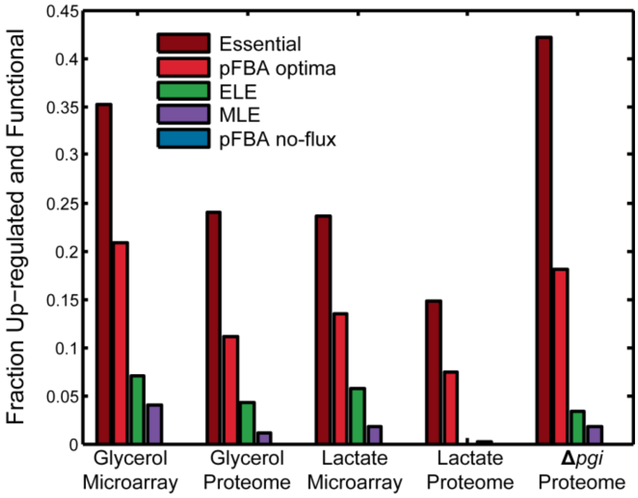
A. Supplementary Figures Referenced in the Main Text


Parsimonious Enzyme Usage FBA (pFBA)
Step 1: Optimize biomass production (FBA)
In an FBA Solution
Not in FBA Solution
Conditionally Non-Functional
Step 2: Minimize net enzymatic flux in the network
Step 3: Classify Associated Genes
Essential
pFBA optima
ELE
MLE
pFBA no-flux
Gene A
Gene B
Gene C
Gene D
Gene E
Gene F
Gene G

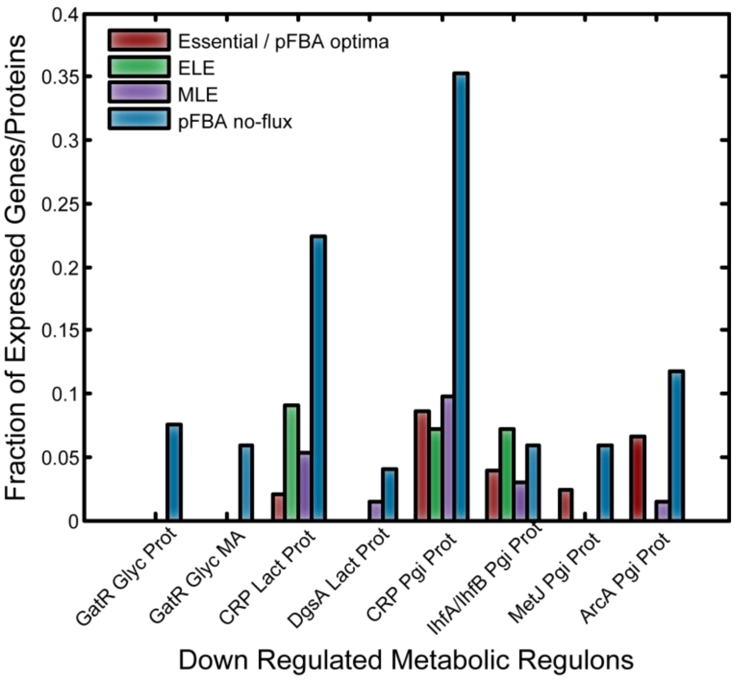
Supplementary Figure 1. Parsimonious Enzyme Usage FBA. In pFBA, the underlying assumption is that, under growth pressure, there is a selection for strains that can process the growth substrate the most rapidly and efficiently while using the minimum amount of enzyme. These assumptions are approximated by employing FBA to optimize the growth rate, followed by minimizing the net metabolic flux through all gene-associated reactions in the network. Therefore, pFBA finds the subset of genes and proteins that may contribute to the most efficient metabolic network topology under the given growth conditions. The genes and proteins are classified as 1) essential, 2) pFBA optima, 3) enzymatically less efficient, requiring more enzymatic steps than alternative pathways that meet the same cellular need (ELE), 4) metabolically less efficient, requiring a reduction in growth rate if used (MLE), or 5) unable to carry flux in the experimental conditions (pFBA no-flux). Here, Gene A, classified as MLE, represents an enzyme that uses a suboptimal co-factor to catalyze a reaction, thereby reducing the growth rate if used. Gene B, classified as pFBA no-flux, cannot carry a flux in this example since it is unable to take up or produce a necessary precursor metabolite. Genes E and F in this example require two different enzymes to catalyze the same transformation which Gene D can do alone; therefore they are classified as ELE. Gene G is essential, since its removal will stop the flux through all pathways. Genes C and D represent the most efficient (topologically and metabolically) pathway and therefore are part of the pFBA optima.



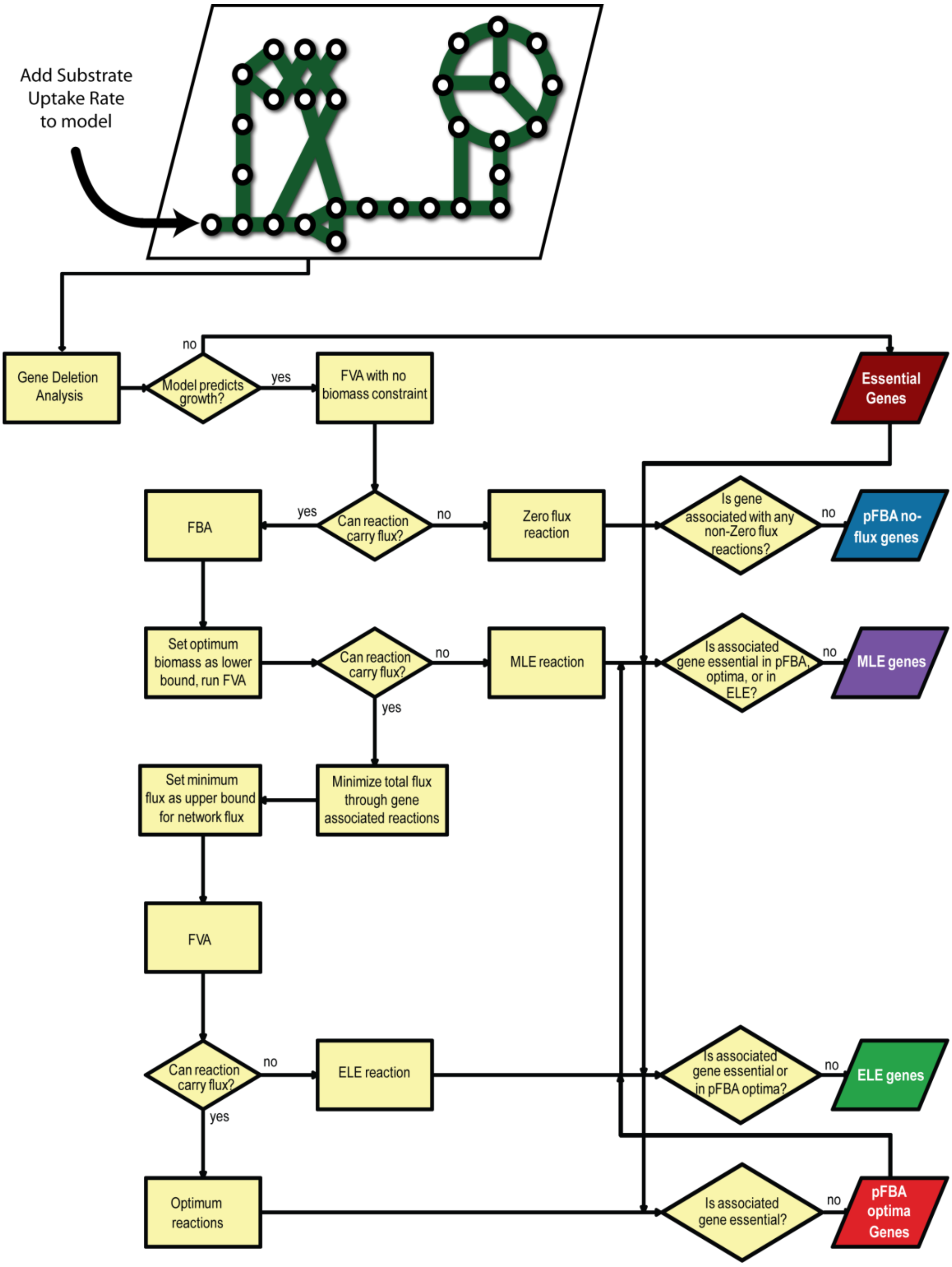
Supplementary Figure 2. pFBA shows a small improvement of normal FBA. pFBA is a variant of Flux Balance Analysis , in which an additional constraint is added that minimizes the total flux through all gene-associated reactions in the metabolic network. The addition of this constraint in pFBA reduces the number of non-essential metabolic genes that are predicted to be used in the optimal solutions. However, the pFBA optima show a slightly higher percent coverage by the (A) proteomic and (B) transcriptomic data than the non-essential FBA optima. Moreover, during the process of adaptive evolution (C), there is a slight increase in up-regulation of the pFBA optima, when compared with up-regulation within the non-essential FBA optima. The cloud represents the normalized distribution of the summed up and down regulated genes or proteins of randomly chosen differentially expressed genes, with x and y values representing the Z-score for the sum of down- and up-regulated gene fold-change, respectively. See Supplementary Analysis (page 20) for details.



Supplementary Figure 3.Few less-efficient genes are up-regulated and functional in up-regulation-optimized models. For each data set, models were generated that minimize the inclusion of non- and weakly up-regulated genes, while maintaining the ability to grow at 90% of the optimal WT growth rate. The optimization method added few functional MLE and ELE genes, suggesting that few up-regulated MLE and ELE genes are able to contribute to growth. See Supplementary Analysis (page 25) for details.



Supplementary Figure 4. Coverage of down regulated genes and proteins in metabolic regulons. To assess the effect of down regulation of a selection of down-regulated regulons, we evaluated how many genes or proteins from each regulon were down regulated in each pFBA class. From this, it is clear that metabolic regulons have a much greater effect on the suppression of genes and proteins in the pFBA no-flux class. See Supplementary Analysis (Page 31) for details.


Add Substrate Uptake Rate to model
Gene Deletion Analysis
Model predicts growth?
no
yes
FVA with no biomass constraint
Essential Genes
FBA
yes
Can reaction carry flux?
no
Zero flux reaction
Is gene associated with any non-Zero flux reactions?
no
pFBA no-flux genes
Set optimum biomass as lower bound, run FVA
Can reaction carry flux?
no
yes
MLE reaction
Is associated gene essential in pFBA, optima, or in ELE?
no
MLE genes
Set minimum flux as upper bound for network flux
Minimize total flux through gene associated reactions
FVA
Can reaction carry flux?
no
yes
ELE reaction
Is associated gene essential or in pFBA optima?
no
ELE genes
Optimum reactions
Is associated gene essential?
no
pFBA optima Genes

Supplementary Figure 5. A flowchart for the simulation and classification of genes in pFBA. Following the addition of experimentally measured substrate and oxygen uptake rates, pFBA was employed to predict pathway usage for the given conditions and to classify the genes, following the workflow demonstrated here.

B. Supplementary Methods

AMT tag method

The theoretical mass and the observed normalized elution time (NET) of each peptide identified by LC-MS/MS is used to construct a reference database of AMT tags, which serve as two-dimensional markers for identifying peptides in subsequent high resolution and high mass accuracy LC-MS analyses. A reference database of AMT tags for *Escherichia coli* had been generated through the exhaustive SCX fractionation and LC-MS/MS analysis described previously (Adkins et al, 2006). This approach to proteomics research is enabled by a number of published and unpublished in-group developed tools, which are available for download at http://omics.pnl.gov (Jaitly et al, 2006;Kiebel et al, 2006;Monroe et al, 2007;Monroe et al, 2008;Petritis et al, 2006). Prior to analysis the samples were subjected to a blocking and randomization treatment to minimize the effects of systematic biases and ensure the even distribution of known and unknown confounding factors across the entire experimental dataset. Peptides from each of the protein preparations were separated by an automated in-house designed reverse-phase capillary HPLC system as described elsewhere (Livesay et al, 2008). Eluate from the HPLC was directly electrosprayed into a 11.4 T FTICR mass spectrometer (LTQ-Orbitrap, Thermo Fisher Scientific, San Jose, CA) using electrospray ionization (ESI) with emitters described previously (Kelly et al, 2007) and the ESI interface modified with an electrodynamic ion funnel (Page et al, 2006). Three biological replicates for each sample were analyzed and relevant information such as the elution time from the capillary LC column, the abundance of the signal (peak height from each peptide elution profile of the most abundant charge state), and the monoisotopic mass (determined from charge state and the high accuracy *m/z* measurement) of each feature observed in the 11.4 T FTICR-MS was used to match the peptide identifications contained within the AMT tag database.

Each biological replicate was analyzed in triplicate and the order was randomized as in Latin Squares design to minimize bias. Isotopic clusters in the spectra were identified using the software tool Decon2LS (Jaitly et al, 2009). The monoisotopic masses of these isotopic clusters were then grouped into LC-MS features (i.e. potential peptides) and aligned against an arbitrarily chosen baseline dataset, using the LCMSWARP algorithm (Zimmer et al, 2006) in order to correct for chromatographic variations. Then the LC-MS features that are commonly observed at least in three datasets were clustered based on a mass tolerance of

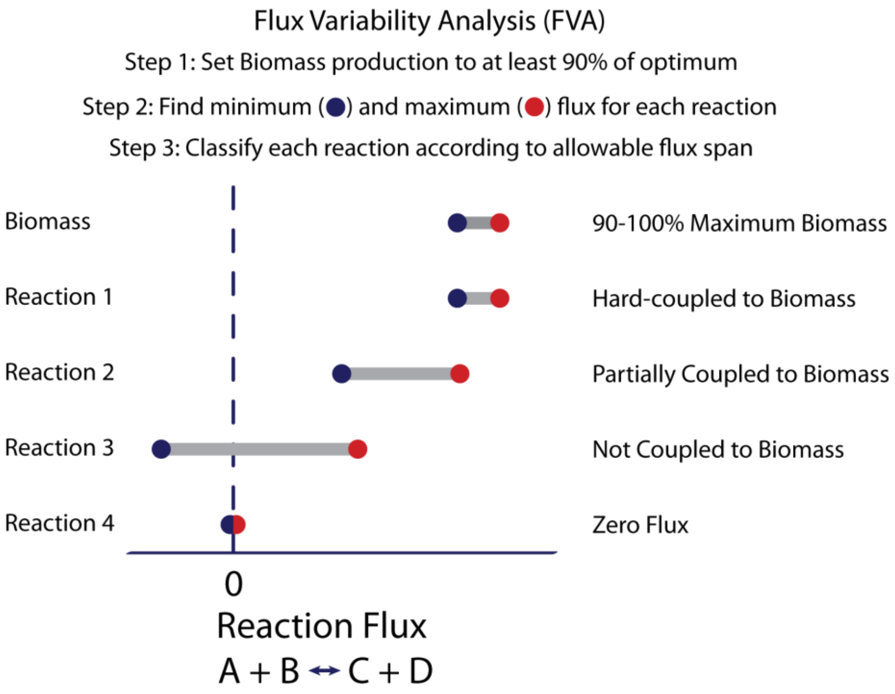
±5ppm and a normalized elution time tolerance of ±0.03. In order to obtain the peptide identifications of these LC-MS clusters, their average masses and average normalized elution times were matched against the AMT tag database created earlier, with a mass tolerance of ±5ppm and a normalized elution time tolerance of ±0.03. The abundance values of these peptides were obtained as the maximum ion current intensity from all MS scans in which it elutes. These steps were carried out using the in-house developed software tool MultiAlign (http://omics.pnl.gov/software/).

Flux Variability Analysis

Flux Variability Analysis (FVA) is a variant of flux balance analysis (FBA) in which the range of allowable flux for each reaction is computed given a range for the allowable predicted growth rate. Using a previously published genome-scale model of *E. coli* K-12 metabolism(Feist et al, 2007), FVA was applied to each environmental condition corresponding to the data provided here. The model was set to 90-100% of the optimal growth rate, and then the maximum and minimum flux for each reaction was computed. Reactions that cannot carry a flux in any condition (i.e., blocked reactions) were removed from the model for all analyses in this study.

Each gene associated reaction was then classified (see Supplementary Table 8). The classifications used in this work are as follow (see Supplementary Figure 6). “Zero-flux” reactions cannot carry a flux while maintaining at least 90% of the maximum growth rate. “Hard-coupled to biomass” reactions need to maintain an exact, specific flux to maintain the optimal growth rate. If the flux through hard-coupled reactions decreases, the growth rate will also decrease in a linear fashion. “Partially coupled to biomass” reactions require a non-zero flux, but can vary their flux while maintaining at least 90% of the maximum flux distribution. All reactions that can maintain a zero or non-zero flux are classified as “Not coupled to biomass.” A small number of remaining reactions that contributed to thermodynamically infeasible loops were removed from the analysis since the flux levels are not constrained and therefore inaccurate.

Reactions associated with each protein were determined using the gene-protein-reaction associations (GPR) in the iAF1263 genome-scale model of *E. coli* metabolism (Feist et al, 2007). These were then reduced to the unique reaction classification-GPR pairs to avoid bias from proteins that can catalyze many reactions (see Supplementary Table 9). Enrichment of each reaction class was computed using the hypergeometric test and p-values are provided in Supplementary Tables 3 and 10.


Flux Variability Analysis (FVA)
Step 1: Set Biomass production to at least 90% of optimum
Step 2: Find minimum (●) and maximum (●) flux for each reaction
Step 3: Classify each reaction according to allowable flux span
Biomass
Reaction 1
Reaction 2
Reaction 3
Reaction 4
90-100% Maximum Biomass
Hard-coupled to Biomass
Partially Coupled to Biomass
Not Coupled to Biomass
Zero Flux
0
Reaction Flux
A + B ↔ C + D

Supplementary Figure 6. Flux Variability Analysis. FVA was used to compute the range of allowable steady state fluxes for all reactions in the metabolic network for each strain, assuming a biomass production that is at least 90% of the optimal growth rate. A reaction was classified as “Hard-coupled to biomass” if the flux varied exactly with biomass production. “Partially coupled to biomass” included reactions that were required to have a non-zero flux, but were more flexible in the range. Reactions were classified as “Not coupled to biomass” if they could have a zero or non-zero flux while maintaining 90% biomass. Reactions were considered “zero flux” if they could maintain a flux in other conditions, but could not in the growth conditions for the strains tested here.

Parsimonious Enzyme Usage FBA

pFBA is a method used to classify genes based on condition-specific pathway usage as predicted in silico (Supplementary Figure 1). It uses a bilevel optimization in which the growth rate (biomass) is optimized using FBA, followed by the minimization of total flux through all gene-associated reactions. The metabolic network is represented by a stoichiometric matrix (Palsson, 2006), S_{irrev}, in which all reversible reactions are split into two irreversible reactions. Each reaction is constrained to carry a non-negative, steady-state flux, v_{irrev}. Thus the net flux is minimized through gene associated reaction subject to optimal biomass:

$$\begin{aligned} &\min \sum_{j=1}^{m} v_{irrev,j} \\ &\quad s.t. \max v_{biomass} = v_{biomass,lb} \ , \\ &\quad\quad s.t.\ S_{irrev} \cdot v_{irrev} = 0 \\ &\quad\quad\quad 0 \le v_{irrev,j} \le v_{\max} \end{aligned}$$

where m is the number of gene-associated irreversible reactions in the network, $v_{biomass}$ approximates the growth rate and $v_{biomass,lb}$ is the lower bound for the biomass rate. A flowchart for the entire process is provided in Supplementary Figure 5.

The underlying assumption here is that the growth selection pressure at exponential growth (as seen here with the adaptively evolved strains) will select for the fastest growing strains. This is followed by the assumption that cells with more efficient enzyme use will have an increased growth advantage. These assumptions have been implemented, as reported previously (Schuetz et al, 2007). This implementation, called “max biomass per unit flux” optimizes the ratio of biomass to the square of the total network flux:

$$\begin{aligned} &\max \frac{v_{biomass}}{\sum_{i=1}^{n} v_i^2} \\ &\quad s.t.\ S \cdot v = 0 \qquad . \\ &\quad\quad v_{\min} \le v_i \le v_{\max} \end{aligned}$$

The “max biomass per unit flux” method is non-linear and non-convex. Moreover, it optimizes for network states that use a higher number of low-flux reactions as opposed to increasing flux through a smaller number of high-flux reactions. In addition, may allow for sub-optimal biomass, if it greatly decreases the flux through the network. In pFBA, on the other hand, the biomass function is maximized, and then the sum of the magnitude of all fluxes is minimized, thereby not biasing the results against the use of higher flux pathways.

Singular Value Decomposition (SVD)

In Singular Value Decomposition (SVD), a matrix, M, is decomposed into three matrices:

$$M = U \Sigma V^T$$

The matrix ∑ is a diagonal matrix that contains rank ordered weightings (singular values) that are indicative the importance of the corresponding modes, as represented by columns in U and V, in reconstituting the data set. More specifically, the amount by which a mode explains variation in the data matrix M is found by squaring the singular value and dividing by the sum of the squares of all singular values (Wall et al, 2003). Sometimes the number of significant singular values is interpreted as being associated with the number of biological processes that produce the variation in the data (Wall et al, 2003).

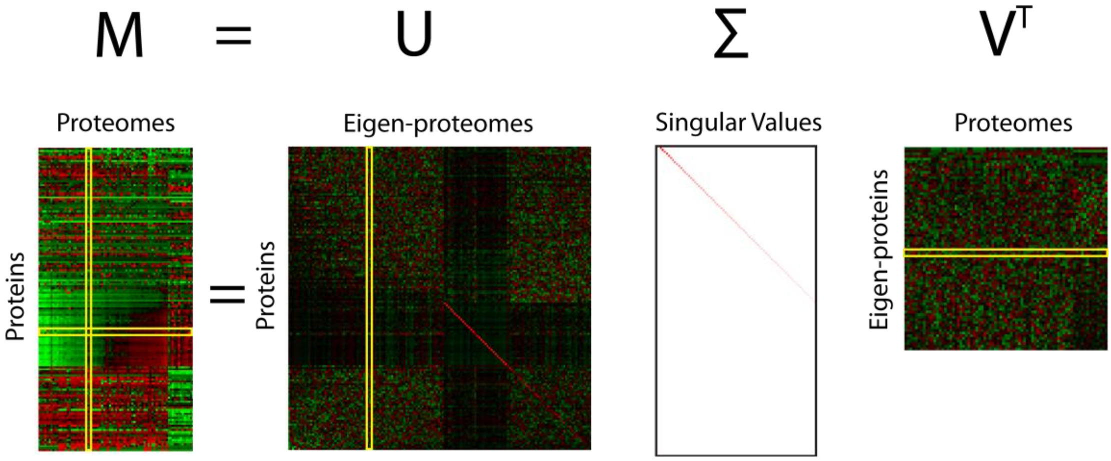
Previously the SVD has been described for microarray data for use in characterizing transcriptional programs and in classification (Alter et al, 2000;Alter, 2006;Challacombe et al, 2004;Liu et al, 2003;Wall et al, 2003;Yeung et al, 2002), and a more detailed description of its significance in gene expression profile analysis has been published (Wall et al, 2003). SVD, however, has been employed less with proteomic data (Bowers et al, 2005;McLaughlin et al, 2007;Vohradsky et al, 2007). In the sense of proteomic data, as depicted in Supplementary Figure 7, each column in the matrix U represents an “eigen-proteome”, which is an orthonormal superposition of the measured proteomes of all experiments, providing a unique combination of proteins expressed at various measures in the mode. These vectors span the space of experiments or measured proteomes.

Each column in the matrix V represents an “eigen-protein”, an orthonormal superposition of proteins which provides a combination of experiment or proteome loadings that represent a unique pattern of globally, uncorrelated, and decoupled proteins. The eigen-proteins span the space of proteomic states. While these eigen-proteins and eigen-proteomes may not have direct biological meaning in detailing mechanisms in transcription, they can provide insights into biological properties and guide further studies, as has been done in DNA microarrays (Wall et al, 2003).

Herein the proteomic (and microarray) data was logarithmically transformed, missing measurements were imputed using an Iterated Local Least Squares Imputation method (Cai et al, 2006), and the mean expression level for each protein across all experiments was subtracted prior to SVD.

The data were subdivided based on gene ontology classification and the dimensionality of these subsets was assessed. The SVDs of expression levels for proteins associated with all GO classes (for protein coding ORFs) were computed in order to probe their relative information content. The distribution of singular values for each GO annotation category was then compared to the distributions of 1000 randomly chosen sets of proteins from the data (each random set contained the same number of proteins as the GO annotation class with which each GO class was compared). Significance of low-dimensional GO classes were determined by t-test on all modes with an explained variance larger than 0.7/n (FDR = 0.05).

In addition, first two eigen-proteins (left singular vectors) were queried to find GO classes in which the data significantly separated the evolved and non-evolved strains.



Supplementary Figure 7. The Singular Value Decomposition in the Proteome Context.

PathWave analysis

Reactions from the genome-scale metabolic network iAF1260 were divided into subsystems (pathways in which they contribute), thereby allowing an elucidation of regions of the metabolic network that underwent significant expression changes in the adaptive evolution process. Each individual pathway was represented by its adjacency-matrix. Pathways that consisted of more than one connected component were further partitioned into sub-pathways. Every sub-pathway was represented by its corresponding adjacency matrix.

We then calculated an embedding for every pathway into a 2-dimensional, regular square lattice grid. To preserve neighborhood characteristics of the nodes, we were looking for embeddings in which adjacent nodes of the network were placed onto the grid as close to each other as possible. As a measure of distance in the lattice, we used the Manhattan distance. We wanted to determine an optimal neighborhood in which the total edge length of the graph on the lattice was minimized while conserving the network topology. This resulted in an NP-hard combinatorial optimization problem. We stated this problem as an integral linear program (IP). The basic model was enhanced by a number of graph dependent, additional constraints on the distance variables. They provided lower bounds for the distance sums of well-known sub graph motifs. Subsequently, the expression data was mapped onto the optimally ordered grid representations of all pathways and sub graphs.

In order to explore every possible expression pattern of neighboring reactions and groups of reactions within a pathway that showed significant differences between samples of different conditions, we calculated features performing a Haar wavelet transform for each optimized grid representation of the pathways. To rank the pathways according to the enrichment of differentially expressed features, the distance between the resulting features of the different conditions was calculated. The maximal value of these distances was taken and compared with an extreme value distribution fitted on the underlying null distribution. The null distribution was estimated by 1000 fold permutation resampling of the expression data samples. Resulting p-values were corrected for multiple testing (Gordi & Khamis, 2004). Furthermore, each single feature was statistically ranked to identify locally differentially regulated patterns. For details see Schramm et al. (Schramm et al, 2010).

C. Supplementary Analysis

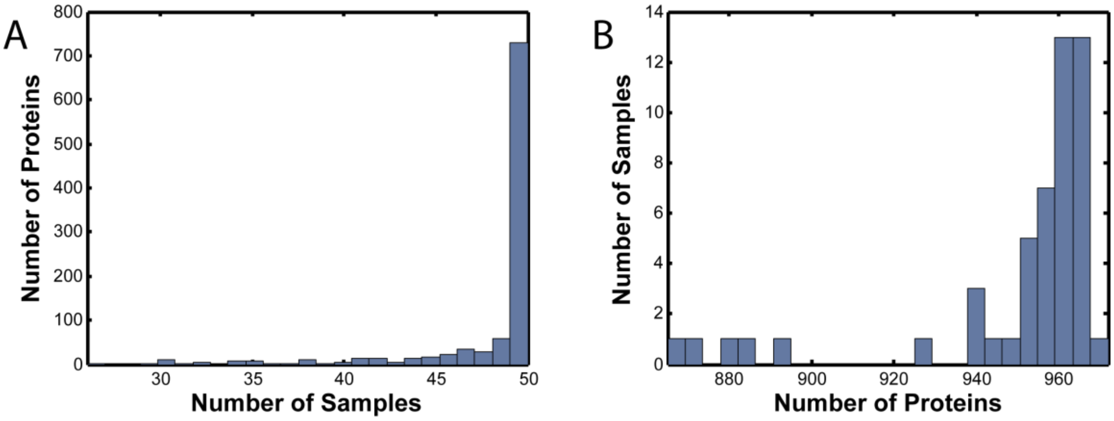
Evaluation of protein species identified in that data set

Gene set enrichment of the proteomic data

Genome scale gene expression evaluation through microarrays and deep sequencing purports to find all RNA transcripts in the biological sample. However, due to the diverse range of chemistry exhibited by proteins, the large range of modifications, and the large dynamic range of protein concentrations, proteomics can only identify a portion of the true proteome in a sample. However, each year brings novel technologies that increase the identification of true positives. In this study, more than 1000 proteins were identified (983 with more than one peptide). 731 of these were found with high confidence in all 50 different samples. Supplementary Figure 8.a shows the overlap of protein identification across all datasets. However, despite this overlap, the datasets had varying numbers of protein identifications (see Supplementary Figure 8.b).

In order to assess this variation in the quantity of identified proteins, the datasets were each tested for enriched (Supplementary Table 11) and depleted (Supplementary Table 12) Gene Ontology classes (Ashburner et al, 2000) with an FDR of 0.05. In this analysis it clear that certain types of proteins were consistently missing from the datasets. The most significantly depleted GO classes (listed in Supplementary Table 12) were dominated by classes of membrane-spanning proteins, and proteins that are intrinsic to the membrane (e.g., “integral to membrane”, “transporter activity”, “flagellum”). For example, the significantly depleted ($p = 5 \times 10^{-12}$) GO class “Integral to Membrane” has 894 genes in it. However, across all datasets 833 of those genes were missing, even when all one-hit-wonders were included in the analysis. Even though novel methods have reduced the difficulty of identifying membrane-spanning and other highly hydrophobic proteins (Ferguson & Smith, 2003), many such proteins were missing from our proteomic datasets.

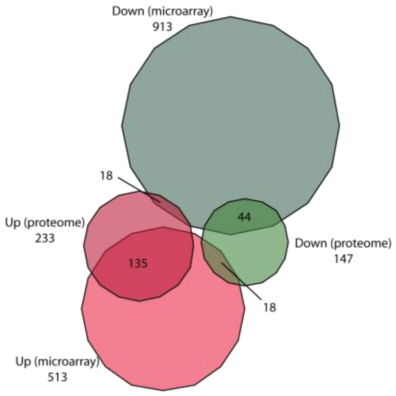
To assess whether this was due to an inherent low expression of membrane-spanning proteins or due to losses in the sample preparation, a similar analysis was done on a series of microarrays that corresponded to about 2/3 of the samples that were proteomically profiled here (Lewis et al, 2009) (http://systemsbiology.ucsd.edu/In_Silico_Organisms/E_coli/E_coli_expression2). In this analysis, presence/absence calls were made from the microarray data using the Wilcoxon Rank sum test with ~20 negative controls on each array (FDR = 0.05). Enriched and depleted GO classes were determined for the microarray data (Supplementary Tables 13-14). Here, similar GO classes were found to be depleted. For example, the GO class "Integral to Membrane" was also depleted in the microarray data ($p << 6 \times 10^{-13}$), though only 332 genes were consistently considered "off" in all conditions considered. Other GO classes that were significantly depleted in both the proteomic and microarray data also include "membrane", "transport", "transporter activity", "uniporter activity", "No GO annotation / non-protein coding", "membrane", "transposase activity", and "transposition, DNA-mediated". In fact, at an FDR of 0.05, the depletion of expressed gene sets in the gene expression data includes almost all sets significantly depleted in the proteomic data. While the quantity of missing genes/proteins differed between the two different data sources, this concordance of depletion of GO classes between the proteomic data and gene expression data lends support to the reliability of both the proteomic methodology and the data sets themselves.



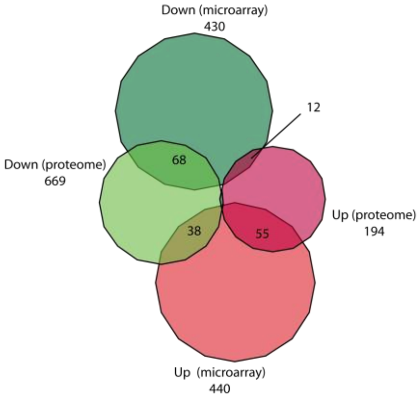
Supplementary Figure 8. Coverage of the proteomic data. The overlap of identified proteins in the various data samples is relatively high. A) Most proteins are found in all samples, and only a small fraction is found in fewer than 45 of the 50 samples. B) In like manner, most samples contain more than 950 proteins that are identified with more than one unique peptide.

Comparison of differential expression in proteomic and microarray data

In the adaptation process hundreds proteins are differentially expressed (Supplementary Table 1), representing 32%, 45%, and 59% of the identified proteins in the glycerol, lactate and Δ*pgi* strains, respectively. In the microarray data, 52% and 35% of the expressed genes are differentially expressed in the glycerol and lactate strains, respectively. However, various studies have shown only a moderate correlation between proteomic and microarray data (Ansong et al, 2009). Here we tested to see if there was agreement between the sets of differentially expressed genes and proteins. For this we compared the overlap of up and down regulated genes and proteins in glycerol (Supplementary Figure 9) and lactate (Supplementary Figure 10) evolved strains. Out of the subset of genes and proteins found both in the transcriptomic and proteomic data, 83% and 71% of the differentially expressed species change their expression in the same direction in the glycerol and lactate strains respectively, which is far more than expected by chance ($p << 1 \times 10^{-18}$).



Supplementary Figure 9. Overlap of differentially expressed genes and proteins in glycerol evolved strains.



Supplementary Figure 10. Overlap of differentially expressed genes and proteins in lactate evolved strains.

Differential expression is associated with central carbon and amino acid metabolism

The COGs demonstrate that many expression changes are associated with specific metabolic processes; however, the effects on specific pathways are not clear from such analyses. Therefore, PathWave, a method based on the Haar wavelet of metabolic network structure (see Supplementary methods and (Schramm et al, 2010)) was used to identify metabolic network subsystems that significantly change in the evolved strain proteomes and transcriptomes. In this analysis, all growth conditions (except the lactate microarray) show significant changes in either the proteomic data and/or the microarrays in central carbon metabolism (i.e., oxidative phosphorylation, pyruvate metabolism, citric acid cycle, anaplerotic reactions, and/or pentose phosphate pathway), tRNA charging, and/or the metabolism of specific amino acids (see Supplementary Table 2). Thus, regional changes in the metabolic network correlate with subsystems, thereby allowing for improved oxidative growth and protein synthesis.

A comparison between FBA and pFBA

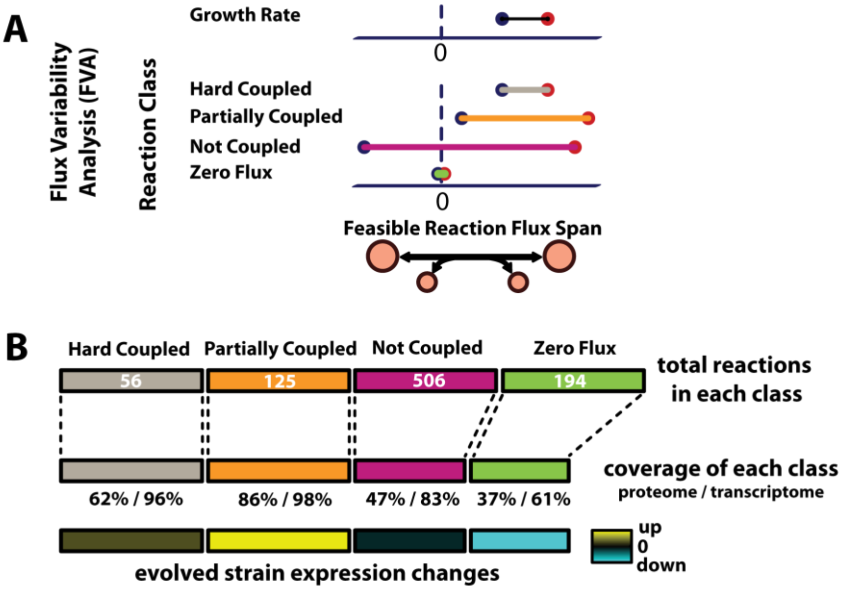
pFBA is a variant of Flux Balance Analysis (FBA) in which growth rate is maximized as in FBA; however, in pFBA, flux through the metabolic network is also minimized. Therefore, a comparison between FBA and pFBA is warranted here. A key part of pFBA, however, is the assessment of all alternate optima and subsequent classification of all genes based on the simulation results. Therefore in this comparison, we have subjected the FBA results to the same search of the alternate optima (using Flux Variability Analysis) and gene classification of the resulting simulations. In the end, the only differences in gene classes are found in the pFBA optima and ELE classes, which are combined in FBA into one "Non-essential FBA optima" class, since flux is not minimized. Percent coverage of this class is slightly lower than the "pFBA optima" class (Supplementary Figure 2.A-B). In like manner, over the adaptive evolution time course, up and down-regulation of genes are slightly more consistent with the pFBA optima than the Non-essential FBA optima (Supplementary Figure 2.C).

Omic Data Supports the Use of the in silico Flux Variability Analysis Optimal Growth States

The optimal solutions computed from FBA are not unique (Lee et al, 2000). Therefore, Flux Variability Analysis (FVA) computes the range of flux values for every reaction that is consistent with the optimal solution. Thus, each reaction in the network can be classified with respect to its possible contribution to the optimal growth state as follows (Supplementary Figures 6 and 11.a):

1. Not coupled to biomass formation,
2. Partially coupled to biomass formation,
3. Hard-coupled to biomass formation, or
4. Unable to carry a flux (zero-flux).

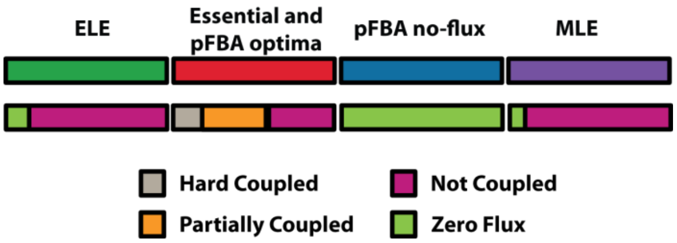
These classes are consistent with the group of expressed and differentially expressed genes and proteins. That is, the more genes and proteins that are more coupled to biomass production show high coverage in the data, and in the evolution time course, these genes and proteins are significantly up-regulated (Supplementary Figure 11.b).



Supplementary Figure 11. FVA shows consistency with proteomic and transcriptomic data. Flux Variability Analysis (FVA) was used to label all metabolic reactions, based on simulation results. (A) FVA classifies each reaction based on its growth rate coupling. (B) Omic data support predicted coupling of reactions to growth, where growth-coupled reactions show higher coverage from omic data than reactions that cannot carry a flux. Moreover, in adaptive evolution, reactions that are partially-coupled to growth are up-regulated, and non-functional reactions are down-regulated. Thus the data support predicted optimal growth states, and the results suggest that laboratory evolved strains further enhance these optimal growth states.

Comparison of pFBA genes and FVA reactions

The pFBA genes were mapped to the FVA reaction classes. From this it is clear that the hard-coupled and partially-coupled reactions were all associated with the essential and pFBA optima genes, and the pFBA no-flux genes were all within the FVA zero-flux reactions. However, FVA Zero Flux reactions were identified in the other pFBA classes since some Zero-Flux reactions are catalyzed by genes which may be active for alternative, functional reactions.



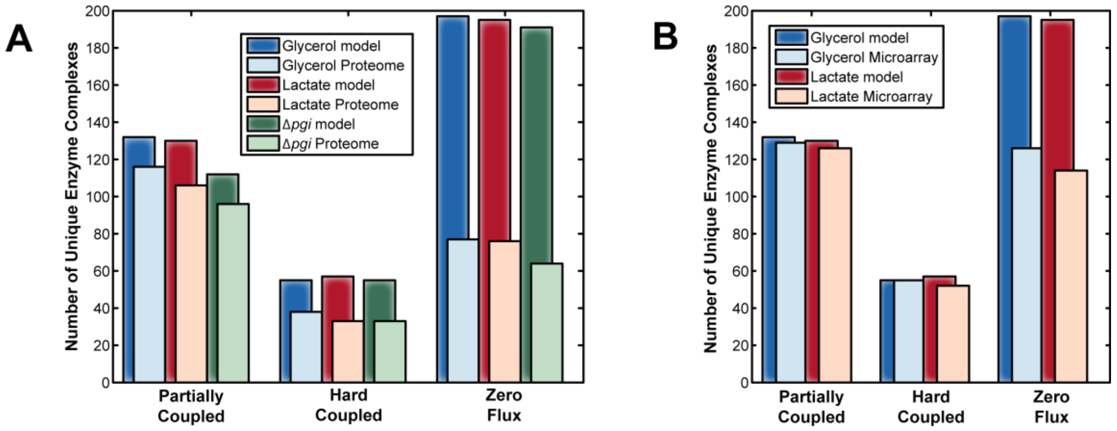
Supplementary Figure 12. A comparison of pFBA genes and FVA reactions.

Expression supports pathway usage from FVA optimal growth predictions

Does FVA show that omics data support optimal growth states? Reactions necessary for optimal growth (i.e., reactions in the partially-coupled and hard-coupled classes) show higher coverage in the proteomic and transcriptomic data than reactions that are predicted to be unnecessary (i.e., not-coupled and zero-flux reactions).

Reactions in the partially-coupled class show the most complete coverage (Supplementary Figure 13). This coverage is much higher than expected by chance for all strains and data types (Supplementary Table 10). Hard-coupled reactions also show good coverage by the proteomic data, and near complete coverage by the transcriptomic data. Missing hard-coupled reactions from the proteomic data may be below the level of detection since they maintain a very small maximum flux (three orders of magnitude lower than partially-coupled reactions; see Supplementary Figure 14); therefore, most hard-coupled reactions will only require a miniscule amount of protein.

Reactions that are not coupled to biomass (NC) show moderate coverage (Supplementary Figure 13). This result is expected since NC reactions may be used, but can represent unnecessary reactions, less efficient pathways, or redundant pathways. Reactions that are predicted to have a zero-flux under the respective growth conditions are significantly depleted in all data sets (Supplementary Table 10). The coverage of these FVA reaction classes suggests that for the given environmental conditions, the transcriptional regulatory network already suppresses many unnecessary genes, and expresses those needed for a high predicted growth rate.



Supplementary Figure 13. FVA classifications are consistent with omic data. Simulations for each growth condition were used to classify each reaction, followed by a comparison to all (A) identified proteins and (B) expressed transcripts. Reaction classes that require flux (Partially and Hard Coupled) consistently have higher coverage from the data than classes that cannot carry flux (Zero Flux).

Adaptation suppresses inactive FVA pathways

Excess unused enzyme mass creates a large maintenance demand on cells (Kurland & Dong, 1996); therefore, cells under selective pressure for growth are expected to modulate expression levels of enzymes as needed for growth (Dekel & Alon, 2005). While we showed an up-regulation of optimal pathways using pFBA, it is expected that genes and proteins associated with non-functional reactions should be down-regulated, thereby saving resources.

Are conditionally nonfunctional genes down-regulated? Genes and proteins associated with the FVA reactions that must have a zero flux *in silico* are significantly down-regulated (Supplementary Table 3). Thus, during the process of adaptive evolution, computationally-predicted nonfunctional pathways are suppressed through a concerted down-regulation of genes associated with such pathways.

Adaptation induces more flexible metabolic reactions

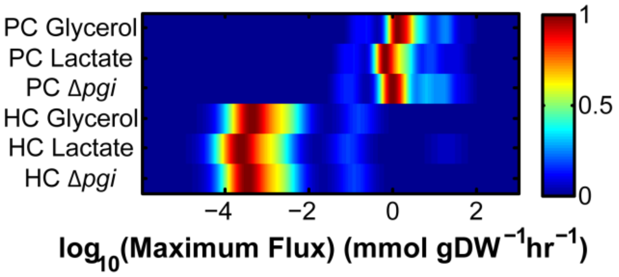
The suppression of unused metabolic pathways frees resources for use in growth-coupled processes. pFBA supports this hypothesis with the emergence of the pFBA optima and the lack of up-regulation among metabolically inefficient enzymes. However, the pFBA optima are associated with reactions with all levels of coupling to growth, including all hard and partially-coupled reactions (Supplementary Figure 12). Therefore, it may be expected that in adaptive evolution, the more flexible partially-coupled reactions would be less up-regulated than the rigid hard-coupled reactions, since the hard coupled reactions are directly essential for growth.

Does adaptive evolution shift expression within the more flexible or more rigid pathways? Surprisingly, reactions that are rigid (hard coupled to growth) are not more frequently up-regulated than expected by chance for almost any of the datasets (Supplementary Table 3). However, in almost all data sets, the more flexible partially-coupled reactions are significantly up-regulated (Supplementary Table 3). Thus, reactions that have a less-direct effect on growth are the most significantly changed. However, this is not problematic if WT strains buffer the rigid reactions by over-expressing the hard-coupled enzymes to allow for a more robust phenotype.

Are the more rigid reactions buffered with over-expression in WT? Simulations predict that the median flux of partially-coupled reactions is more than 3 orders of magnitude higher than hard-coupled reactions (Supplementary Figure 14). However, the protein and transcript abundance medians for partially-coupled reactions are only 1.8-3.7 fold higher in the WT strains. This suggests that hard-coupled reactions may be buffered in WT strains with excess enzyme, thus protecting against momentary shifts in environmental conditions. The buffer would be less necessary for partially-coupled enzymes, since they can maintain a lower flux with only having a weak effect on biomass production. These results suggest that during adaptive evolution, there is a shift of expression towards pathways that are flexible, since hard-coupled pathways are already over-expressed. Thus, metabolically, the evolved strains will be slightly less robust against shifts to significantly different growth conditions.

Hard-coupled reactions carry a much lower flux than partially-coupled reactions

It is puzzling as to why there is much less coverage of reactions that are hard-coupled (HC) to biomass in the proteomic data. While membrane-bound proteins are depleted in the data (Supplementary Table 12), only a few HC reactions that are associated with membrane proteins were not identified in the data. However, when the maximum flux of each reaction is computed, most partially-coupled (PC) reactions have a higher flux than HC reactions (Supplementary Figure 14). In fact, the median flux for HC reactions is more than three orders of magnitude smaller than the median PC flux values.



Supplementary Figure 14. Hard-coupled reactions carry a much lower flux than partially-coupled reactions. For each growth condition, the fluxes for all hard-coupled and partially-coupled reactions were determined. The distribution of maximum fluxes for partially-coupled reactions is significantly higher than the distribution of allowed fluxes for hard-coupled reactions. HC = hard-coupled, PC = partially-coupled.

Contribution of up-regulated less-efficient (ELE and MLE) genes

A small fraction of metabolically less-efficient (MLE) and enzymatically less efficient (ELE) genes were up-regulated in the evolved strains. Potentially, less efficient genes can increase growth rate despite their decreased efficiency, if they allow for the usage of kinetically faster enzymes and shorter pathways. The results we have presented, however, do not support this hypothesis, since the strains presented here evolved to a higher biomass yield (Supplementary Table 7), thereby suggesting that the metabolic networks of the evolved strains are more efficient than that of the unevolved strains. However, to further answer this question, an analysis is presented here to test if the less efficient genes are up-regulated within functional pathways or if their up-regulation is for other processes beyond biomass production.

It has been shown by our study and others that the end points of adaptive evolution can accurately be described using *in silico* models (Fong & Palsson, 2004;Ibarra et al, 2002). Here, we aimed to determine the most likely condition-specific metabolic model given high-throughput data obtained from cells exhibiting the various optimal phenotypes. This allowed for a systematic evaluation of pathways that were up- and down-regulated.

To do this, an optimization was performed subject to all existing constraints from the metabolic model, including the previously reported substrate uptake rates (Charusanti et al, submitted;Fong et al, 2005). Additionally, the model was constrained to meet at least 90% of WT growth rate. The goal was to retain all genes supported by up-regulated expression data, but maintain model functionality towards biomass formulation. This was carried out by minimizing the sum of all fold-changes for all genes with a fold change less than 2 (increased expression in the evolved strains):

$$min \sum_i w_i g_i$$

Here, w_i is a gene-specific weight for each gene added to the model, g_i. These weights are computed using the fold-change data in conjunction with a $\log_2$ fold-change cut-off of 1. More specifically, it is computed by taking the distance from the fold-change for each gene to the fold change cut-off. As an example, if the $\log_2$ fold change distance is 2.5, and the $\log_2$ fold change cut-off is 1, the weight for this gene would be 1.5 (and the model will include this gene in the solution returned since $g_i = 1$). If the $\log_2$ fold change was 0.5, and the $\log_2$ fold change cut-off is 1, the weight for this gene would be -0.5. The model will only include this gene in the solution returned if it is critical for meeting 90% of the biomass objective.

We recognize that this formulation likely leads to the inclusion of genes that cannot be functional (towards biomass formation) in the final returned model, if they have a fold change higher than the cutoff. To address this concern, we subsequently performed flux variability analysis (FVA) and removed all genes that cannot possibly have a role in the objective. More specifically, we removed genes for which all reactions they can participate in must carry a 0 flux. However, this method retains all isozymes.

A few MLE and ELE genes consistently showed up in the final returned models. These genes were of considerable interest as their expression leads to an optimal phenotype but violates the pFBA objective. Most of these genes are relevant in amino acid and nucleic acid metabolism. The most prominent examples are as follow.

As mentioned in the main text, only one MLE gene was up-regulated in all data sets. Deoxyuridinetriphosphatase (DnaS; 3.6.1.23), which dephosphorylates dUTP, is up-regulated in all datasets, presumably to maintain genome integrity at the increased growth rates by decreasing the dUTP concentration in the cell (Hochhauser & Weiss, 1978).

Another enzyme that is up-regulated in the glycerol and lactate strain microarray data and the *pgi* deletion strain proteomic data is Formyltetrahydrofolate Hydrolase (PurU; 3.5.1.10). This enzyme converts 10-Formyltetrahydrofolate to tetrahydrofolate, thereby producing extra formate for purine synthesis in aerobic conditions (Nagy et al, 1995), thereby increasing the concentration of purines for the increased demands for DNA and RNA precursors at higher growth rates. Moreover, PurU is known to regulate glycine biosynthesis, thereby safeguarding it under high purine concentrations (Nagy et al, 1995).

The glycerol strain microarray data and the *pgi* deletion strain proteomic data also show an up-regulation of the NAD transhydrogenase in conjunction with the up-regulation of the oxidative pentose phosphate pathway (which is necessary for the pgi deletion strain, but MLE in the glycerol strains). The transhydrogenase expression level has been shown to correlate with growth rate previously (Canonaco et al, 2001). In the glycerol and *pgi* deletion strains, the up-regulation of the NAD transhydrogenase is likely important for maintaining decreasing the buildup of NADPH from the up regulation of the oxidative pentose phosphate pathway.

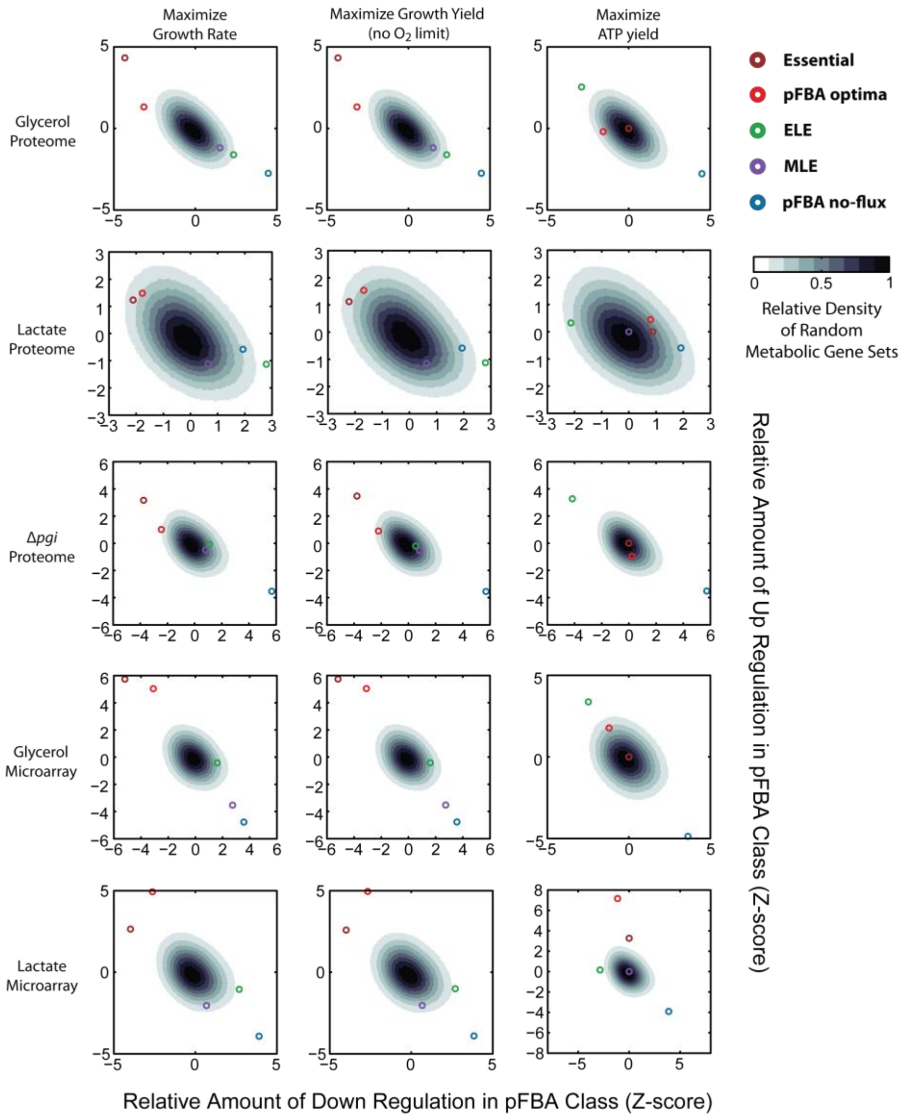
The remaining functional MLE and ELE genes contribute primarily to amino acid transport and nucleotide metabolism and oxidation, likely due to the minimal media and high growth rates, respectively. Overall, of all MLE and ELE genes only a small fraction is up-regulated and can contribute to functional pathways (Supplementary Figure 3).

Comparison between pFBA "growth rate" objective vs. other common objective functions for adaptive evolution

A wide number of different cellular objectives have been used to describe cell decisions in phenotypic prediction. Researchers have used optimization methods to describe cellular growth, such as Flux Balance Analysis, with objectives such as optimizing growth yield (mol biomass per mol substrate), "growth rate" (biomass for a given substrate uptake rate, with constraints for cellular maintenance), or ATP production. A recent work (Schuetz et al, 2007) conducted a large analysis of various objective functions and additional constraints to assess which combinations predicted a flux distribution most similar to the experimentally measured flux through the network. In that work, no combination was found to best describe the six experimental conditions tested. However in batch culture, it was shown that ATP yield per unit flux was the most accurate. For nutrient scarce continuous conditions (similar to those in our study) the maximization of ATP or biomass yield was best.

Here, we have assessed the accuracy of a few different objectives to identify which, when combined with flux minimization, is most consistent with the expression changes seen in adaptive evolution. As shown in Supplementary Figure 15, very little difference was found between maximizing "growth rate" and "growth yield" (including the removal non-growth associated maintenance and the measured oxygen uptake rate constraints). On the other hand, maximizing ATP yield led to poor results. This, however, was expected. In the previous study that showed positive results for ATP yield maximization, only a model of central metabolism was used (Schuetz et al, 2007). However, in a genome-scale model, the maximization of ATP selects against the usage of biosynthetic pathways, since the end products are not specified in the objective function. Thus, these pathways will either decrease the ATP yield or increase flux through the metabolic network. In fact, few genes contribute to the Essential and pFBA optima classes under the ATP yield objective, and most genes are included in the ELE class.

Across all of the data sets only the pgi deletion proteomic data showed an up-regulation of essential and pFBA optima genes when optimizing ATP. Most other sets only showed either up-regulation or a lack of down-regulation (Supplementary Figure 15). This inconsistency of results between data sets when optimizing ATP yield for a genome-scale model, suggests that optimizing "growth rate" is a better in silico objective (though "growth yield" is just about the same).



Supplementary Figure 15. A comparison of different objective functions. pFBA with a maximization of growth rate, is not significantly different from pFBA with an maximization of the growth yield (without O_2 constraints). However, pFBA with the maximization of ATP yield performs worse since the Essential genes and pFBA optima are not up-regulated for any data sets except for the pgi deletion strain proteomic data. The cloud represents the normalized distribution of the summed up and down regulated genes or proteins of randomly chosen differentially expressed genes, with x and y values representing the Z-score for the sum of down- and up- regulated gene fold-change, respectively.

Significant changes in translation and metabolism in evolved strains through SVD

To discover processes that are affected in the evolved strains, singular value decomposition (SVD) was employed. SVD is commonly used to determine dimensionality of a data set, and the dimensionality is thought to be correlated with the number of biological processes that produce the variation in the data (Wall et al, 2003). When sets of protein/gene expression data have a significantly low dimensionality (as determined through SVD), this suggests that there is a mechanism that allows such molecules to co-vary in their expression, possibly due to similarities in function, genomic location, and/or regulation. The SVDs were computed for all Gene Ontology (GO) classes. At an FDR of 0.05, there are 11 GO annotation classes in which the proteomic data have a significantly lower dimensionality than randomly chosen proteins, and 20 significant GO classes in the gene expression profiles (see Supplementary Tables 15 and 16).

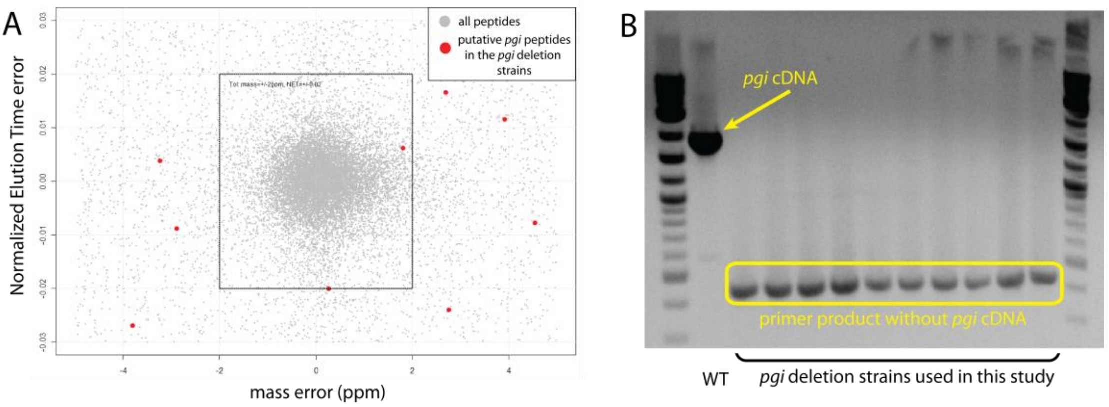
The global analysis of dimensionality of GO classes in the proteomic and transcriptomic data demonstrates the dominance in the changes in translation and metabolism (see Supplementary Tables 15 and 16). In the SVD of all datasets, ribosomal and translation-associated protein classes dominate the low dimensional classes. In addition, when the SVD is conducted on data from the individual experimental conditions, strains grown on lactate or glycerol are also dominated by GO classes for several metabolic processes (see Supplementary Tables 17-18). For the SVD of the data from strains in which *pgi* was deleted, however, more than 200 GO classes significantly separate the evolved and non-evolved strains in only two modes (see Supplementary Table 19). Most of these classes are also involved with transcription, translation, and metabolism. Thus, the SVD of data from all three evolution conditions supports results presented in Figure 4 of the main text, thereby demonstrating the importance of changes in translation and metabolism in the process of adaptive evolution.

Effect of down regulated regulons on metabolism

Several metabolism-associated regulons are enriched among each growth condition. Therefore it is desirable to quantify the effect each regulon has on the different pFBA classes. If the pFBA classes are reasonable, one would expect that for each regulon, down regulation would have only a small effect on the essential genes and pFBA optima, while a higher fraction of pFBA no-flux genes should be down-regulated. For each growth condition, metabolism-related regulons that were significantly down-regulated were selected and tested to assess their coverage for all expressed genes or identified proteins. As predicted, a higher fraction of identified proteins and expressed genes are down-regulated in the pFBA no-flux for the most highly enriched metabolic-associated regulons (Supplementary Figure 4).

Concern of Pgi identified in pgi-deletion proteomic datasets

Upon inspection of the proteomic data, it was surprising to find that Pgi was identified in all of the *pgi*-deletion strain proteomic data sets. Across all data sets, 47 different peptides for Pgi had been identified. However, *pgi*-deletion strains only had on average 4 of these 47 peptides in each sample. In addition, all but one peptide was identified outside of an FDR of 0.06 (Supplementary Figure 16.a). Further RTPCR and analysis of microarray data from the *pgi* deletion strains verified zero expression of the *pgi* gene (Supplementary Figure 16.b). For RTPCR, the flanking lanes were loaded with 2-log ladder from New England Biolabs (#N3200S) in order to provide an estimate of DNA fragment sizes. The second lane from the left is *pgi* amplified from wild-type *E. coli*. The remaining lanes are Δ*pgi* amplified from the evolved strains. The primers used for wild-type and the evolved strains were all the same, and surround the *pgi* gene but do not include any actual base pairs associated with the gene. Subsequently, fresh samples were subjected to a second round of proteomic profiling, and the same Pgi peptides were identified (data not shown). Therefore, we are able to conclude that the abundance values reported for Pgi in proteomic dataset for the Δ*pgi* strains represent false positive assignments.



Supplementary Figure 16. Verification that pgi removal was successful. Removal of *pgi* from Δ*pgi* strains was successful. A) A handful of the 47 distinct identified Pgi peptides are found in the proteomic data for the *pgi* deletion strains. However, all such peptides, except one, had high mass and/or normalized elution time errors and thus were identified at an FDR greater than 6% (black box). B) To further verify that *pgi* was successfully removed and that evolved strains were lacking *pgi*, all Δ*pgi* strains were subjected to RTPCR. While *pgi* is highly expressed in wild type *E. coli*, it is clearly not expressed in any of the Δ*pgi* evolved strains. Together these results show that *pgi* was successfully deleted and that the identified peptides were false positives.

References

Adkins JN, Mottaz HM, Norbeck AD, Gustin JK, Rue J, Clauss TR, Purvine SO, Rodland KD, Heffron F, Smith RD (2006) Analysis of the Salmonella typhimurium proteome through environmental response toward infectious conditions. *Mol Cell Proteomics* **5:**1450-1461. doi: 10.1074/mcp.M600139-MCP200

Alter O (2006) Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc Natl Acad Sci U S A* **103:**16063-16064. doi: 10.1073/pnas.0607650103

Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* **97:**10101-10106

Ansong C, Yoon H, Porwollik S, Mottaz-Brewer H, Petritis BO, Jaitly N, Adkins JN, McClelland M, Heffron F, Smith RD (2009) Global systems-level analysis of Hfq and SmpB deletion mutants in Salmonella: implications for virulence and global protein translation. *PLoS One* **4:**e4809. doi: 10.1371/journal.pone.0004809

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:**25-29. doi: 10.1038/75556

Bowers PM, O'Connor BD, Cokus SJ, Sprinzak E, Yeates TO, Eisenberg D (2005) Utilizing logical relationships in genomic data to decipher cellular processes. *FEBS J* **272:**5110-5118. doi: 10.1111/j.1742-4658.2005.04946.x

Cai Z, Heydari M, Lin G (2006) Iterated local least squares microarray missing value imputation. *J Bioinform Comput Biol* **4:**935-957

Canonaco F, Hess TA, Heri S, Wang T, Szyperski T, Sauer U (2001) Metabolic flux response to phosphoglucose isomerase knock-out in Escherichia coli and impact of overexpression of the soluble transhydrogenase UdhA. *FEMS Microbiol Lett* **204:**247-252

Challacombe JF, Rechtsteiner A, Gottardo R, Rocha LM, Browne EP, Shenk T, Altherr MR, Brettin TS (2004) Evaluation of the host transcriptional response to human cytomegalovirus infection. *Physiol Genomics* **18:**51-62. doi: 10.1152/physiolgenomics.00155.2003

Charusanti P, Knight EM, Conrad T, Venkataraman K, Chew C, Xie B, Gao Y, Palsson BØ Genetic basis of growth adaptation of Esherichia coli after deletion of pgi, a major metabolic gene. *Submitted*

Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436:**588-592. doi: 10.1038/nature03842

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3:**121. doi: 10.1038/msb4100155

Ferguson PL, Smith RD (2003) Proteome analysis by mass spectrometry. *Annu Rev Biophys Biomol Struct* **32:**399-424. doi: 10.1146/annurev.biophys.32.110601.141854

Fong SS, Palsson BO (2004) Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet* **36:**1056-1058. doi: 10.1038/ng1432

Fong SS, Joyce AR, Palsson BO (2005) Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. *Genome Res* **15:**1365-1372. doi: 10.1101/gr.3832305

Gordi T, Khamis H (2004) Simple solution to a common statistical problem: interpreting multiple tests. *Clin Ther* **26:**780-786

Hochhauser SJ, Weiss B (1978) Escherichia coli mutants deficient in deoxyuridine triphosphatase. *J Bacteriol* **134:**157-166

Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420:**186-189. doi: 10.1038/nature01149

Jaitly N, Mayampurath A, Littlefield K, Adkins JN, Anderson GA, Smith RD (2009) Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics* **10:**87. doi: 10.1186/1471-2105-10-87

Jaitly N, Monroe ME, Petyuk VA, Clauss TR, Adkins JN, Smith RD (2006) Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal Chem* **78:**7397-7409. doi: 10.1021/ac052197p

Kelly RT, Page JS, Tang K, Smith RD (2007) Array of chemically etched fused-silica emitters for improving the sensitivity and quantitation of electrospray ionization mass spectrometry. *Anal Chem* **79:**4192-4198. doi: 10.1021/ac062417e

Kiebel GR, Auberry KJ, Jaitly N, Clark DA, Monroe ME, Peterson ES, Tolic N, Anderson GA, Smith RD (2006) PRISM: a data management system for high-throughput proteomics. *Proteomics* **6:**1783-1790. doi: 10.1002/pmic.200500500

Kurland CG, Dong H (1996) Bacterial growth inhibition by overproduction of protein. *Mol Microbiol* **21:**1-4

Lee S, Palakornkule C, Domach MM, Grossmann IE (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Computers & Chemical Engineering,* **24:**711-716

Lewis NE, Cho BK, Knight EM, Palsson BO (2009) Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. *J Bacteriol* **191:**3437-3444. doi: 10.1128/JB.00034-09

Liu L, Hawkins DM, Ghosh S, Young SS (2003) Robust singular value decomposition analysis of microarray data. *Proc Natl Acad Sci U S A* **100:**13167-13172. doi: 10.1073/pnas.1733249100

Livesay EA, Tang K, Taylor BK, Buschbach MA, Hopkins DF, LaMarche BL, Zhao R, Shen Y, Orton DJ, Moore RJ, Kelly RT, Udseth HR, Smith RD (2008) Fully automated four-column capillary LC-MS system for maximizing throughput in proteomic analyses. *Anal Chem* **80:**294-302. doi: 10.1021/ac701727r

McLaughlin WA, Chen K, Hou T, Wang W (2007) On the detection of functionally coherent groups of protein domains with an extension to protein annotation. *BMC Bioinformatics* **8:**390. doi: 10.1186/1471-2105-8-390

Monroe ME, Shaw JL, Daly DS, Adkins JN, Smith RD (2008) MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Comput Biol Chem* **32:**215-217. doi: 10.1016/j.compbiolchem.2008.02.006

Monroe ME, Tolic N, Jaitly N, Shaw JL, Adkins JN, Smith RD (2007) VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* **23:**2021-2023. doi: 10.1093/bioinformatics/btm281

Nagy PL, Marolewski A, Benkovic SJ, Zalkin H (1995) Formyltetrahydrofolate hydrolase, a regulatory enzyme that functions to balance pools of tetrahydrofolate and one-carbon tetrahydrofolate adducts in Escherichia coli. *J Bacteriol* **177:**1292-1298

Page JS, Tolmachev AV, Tang K, Smith RD (2006) Theoretical and experimental evaluation of the low m/z transmission of an electrodynamic ion funnel. *J Am Soc Mass Spectrom* **17:**586-592. doi: 10.1016/j.jasms.2005.12.013

Palsson B (2006) *Systems biology : properties of reconstructed networks*. Cambridge University Press, Cambridge ; New York

Petritis K, Kangas LJ, Yan B, Monroe ME, Strittmatter EF, Qian WJ, Adkins JN, Moore RJ, Xu Y, Lipton MS, Camp DG,2nd, Smith RD (2006) Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal Chem* **78:**5026-5039. doi: 10.1021/ac060143p

Schramm G, Wiesberg S, Diessl N, Kranz AL, Sagulenko V, Oswald M, Reinelt G, Westermann F, Eils R, Konig R (2010) PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics* **26:**1225-1231. doi: 10.1093/bioinformatics/btq113

Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol* **3:**119. doi: 10.1038/msb4100162

Vohradsky J, Branny P, Thompson CJ (2007) Comparative analysis of gene expression on mRNA and protein level during development of Streptomyces cultures by using singular value decomposition. *Proteomics* **7:**3853-3866. doi: 10.1002/pmic.200700005

Wall ME, Rechtsteiner A, Rocha LM (2003) Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis,* Berrar DP, Dubitzky W, Granzow M (eds) pp. 91-109. Kluwer: Norwell, MA

Yeung MK, Tegner J, Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A* **99:**6163-6168. doi: 10.1073/pnas.092576199

Zimmer JS, Monroe ME, Qian WJ, Smith RD (2006) Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom Rev* **25:**450-482. doi: 10.1002/mas.20071