Supplementary Materials for

The quantitative and condition-dependent *Escherichia coli* proteome

Alexander Schmidt[1], Karl Kochanowski[2], Silke Vedelaar[5], Erik Ahrné[1], Benjamin Volkmer[2], Luciano Callipo[2], Kèvin Knoops[4], Manuel Bauer[1], Ruedi Aebersold[2,3], Matthias Heinemann[2,5]

[1] Biozentrum, University of Basel, Basel, Switzerland
[2] Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
[3] Faculty of Science, University of Zurich, Zurich, Switzerland
[4] Molecular Cell Biology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands
[5] Molecular Systems Biology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

Corresponding authors:
Alexander Schmidt (alex.schmidt@unibas.ch) and Matthias Heinemann (m.heinemann@rug.nl)

This PDF file includes:

Supplementary Figures:

Supplementary Figure 1: Number of quantified peptides/proteins using different sample preparation schemes.

Supplementary Figure 2: Impact of LC gradient length on the number of peptides and proteins identified per LC-MS analysis.

Supplementary Figure 3: Impact on different peptide fractionation schemes on the number of identified proteins and the required running time per sample.

Supplementary Figure 4: Properties of the two different large-scale quantitative LC-MS datasets included in this manuscript.

Supplementary Figure 5: Fold error estimation of determined protein concentrations.

Supplementary Figure 6: Proteome coverage assessment.

Supplementary Figure 7: Evaluation of the technical and biological reproducibility of our absolute quantification approach.

Supplementary Figure 8: Correlation of our absolute protein abundance estimates with various published small datasets including only a few growth conditions.

Supplementary Figure 9: Hierarchical clustering of relative protein abundance changes (to glucose) of all samples.

Supplementary Figure 10: Quantitative proteome comparison of three *E. coli* strains (BW25113, MG1655, NCM3722) grown in minimal (glucose) and rich (LB) media.

Supplementary Figure 11: Correlation between the growth rate and the relative protein mass fractions for all proteins assigned to the different COG categories.

Nature Biotechnology: doi:10.1038/nbt.3418

Supplementary Figure 12: Cumulative distribution of the coefficient of variation (CV) for all COG classes (red line) compared to the whole detected proteome (black dashed lines).

Supplementary Figure 13: Periplasmic protein mass distribution geometrically corrected for increase in cell size with higher growth rates.

Supplementary Figure 14: Cryo-electron microscopy analysis of *E. coli* cells.

Supplementary Figure 15: Unrestricted open modification search of all unassigned MS/MS-spectra.

Supplementary Figure 16: Correlation of the growth rate and the relative abundance of all identified peptides carrying a specific modification, respectively.

Supplementary Figure 17: Relative change in abundance of the identified N^{α}-acetylation sites for wild type (WT) and mutant strains lacking the three known N-acetyltransferases annotated in the *E. coli* genome (ΔrimI, ΔrimJ and ΔrimL), respectively.

Supplementary Figure 18: Relative change in abundance of the identified N^{α}-acetylation sites according to the modified amino acid at the protein N-terminus.

Supplementary Notes:

Supplementary Note 1: Dataset quality assessment

Supplementary Note 2: Statistical data analysis

Supplementary Note 3: Cell volume considerations

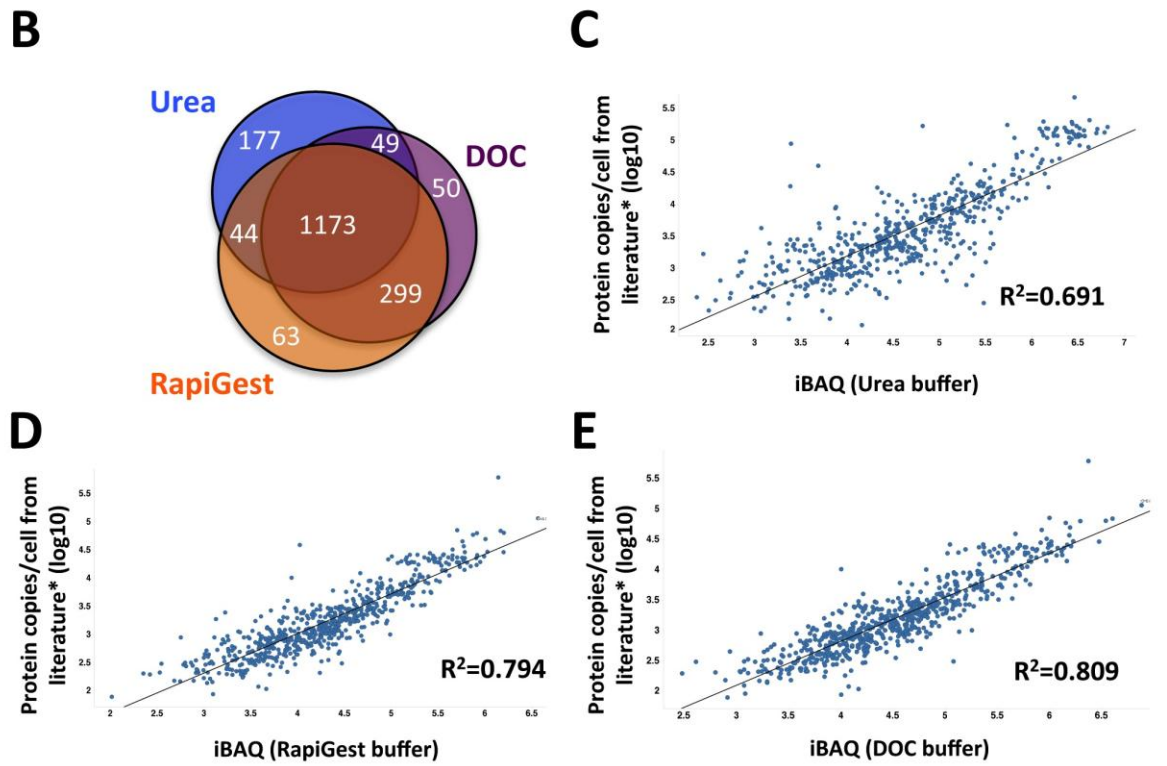
Supplementary Figure 1:

A

Number of identified proteins upon using different peptide fractionation schemes

Lysis Buffers Used[1]	Urea	DOC	RapiGest
# of Peptide Spectrum Matches[2]	32533	33550	30453
# of Peptides Quantified by LFQ[2]	11496	12269	11475
# of Proteins Quantified by LFQ[2]	1444	1571	1579

1) For detailed lysis buffer conditions see material and methods
2) FDR was set to 1% based on the number of decoy hits in the dataset



Supplementary Figure 1: Number of quantified peptides/proteins using different sample preparation schemes. (A) A few popular sample preparation methods based on three different lysis buffer conditions using urea as well as the detergents sodium deoxycholate (DOC) and RapiGest, respectively, were applied to *E. coli* cells grown in glucose media (see material and methods for details). For the RapiGest samples, DOC was replaced by 1% RapiGest in the lysis buffer. The number of identified spectra as well as the number of peptides and proteins suitable for quantification by our label-free quantification (LFQ) approach are shown. (B) Venn diagram of the protein set quantified by the three different sample preparation conditions. (C) Correlation of iBAQ values determined by LFQ[5] for the proteins extracted using the urea based buffer with recently published protein concentrations using ribosome profiling[6]. Likewise, the correlations of iBAQ values obtained from RapiGest (D) and DOC (E) based lysis buffers are illustrated. Squared Pearson correlation coefficients (R^2) are indicated for all plots. Based on the high number of quantified peptides and proteins and the high correlation with published absolute protein abundances, the DOC buffer was found to be most suited for system-wide proteome analysis and was applied to all *E. coli* samples in this study. FDR: false-discovery-rate.

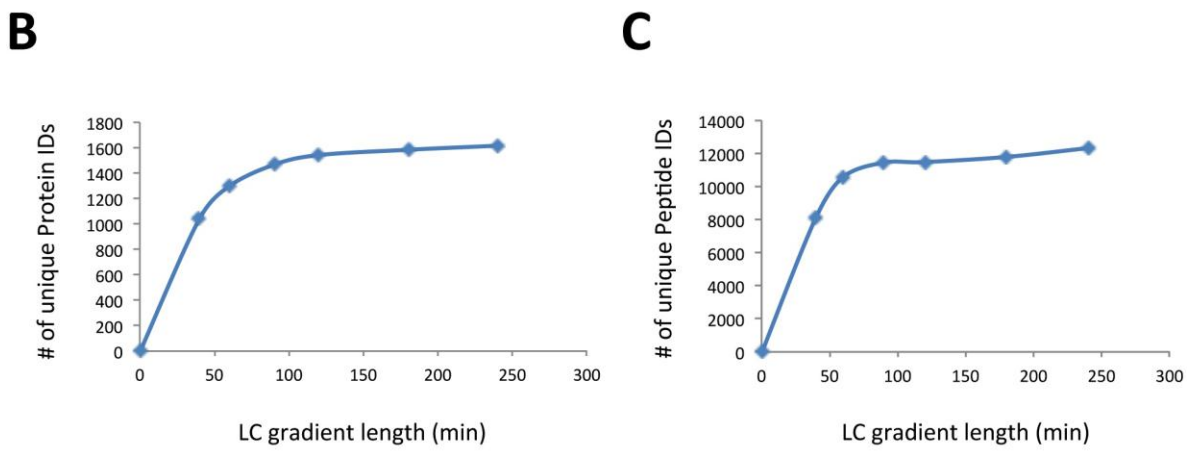
Supplementary Figure 2:

A

Number of unique peptide/protein IDs using different LC gradient lengths

LC-MS Experiment	1	2	3	4	5	6
Gradient Time (min)	40	60	90	120	180	240
# of Unique Protein IDs[1]	1039	1297	1466	1542	1584	1616
# of Unique Peptide IDs[2]	8133	10585	11456	11489	11799	12333
Protein IDs/h[3]	1558	1297	977	771	528	404

1) FDR for protein IDs was set to 1% based on the number of decoy hits in the dataset
2) FDR for peptide IDs was set to 1% based on the number of decoy hits in the dataset
3) Not including times for sample loading and column equilibration/washing steps



Supplementary Figure 2: Impact of LC gradient length on the number of peptides and proteins identified per LC-MS analysis. (A) Six different linear LC gradient length starting from 5 to 28% solvent B were applied to analyze sample aliquots containing 2 µg of peptides generated from *E. coli* cells grown in glucose minimal media. The number of unique protein and peptide identification together with the protein identification rate per hour of gradient time is given for all LC lengths, respectively. (B) Line chart illustrating the number of unique protein IDs for the different LC methods shown in (A). (C) Like (B) for peptide identifications. Based on the number of identified proteins per time, a 120 minutes gradient was found to be the best compromise with respect to proteome coverage and measurement time on the LC-MS/MS platform employed for this study. FDR: false-discovery-rate, ID: identification, LC: liquid chromatography.

Supplementary Figure 3:

A

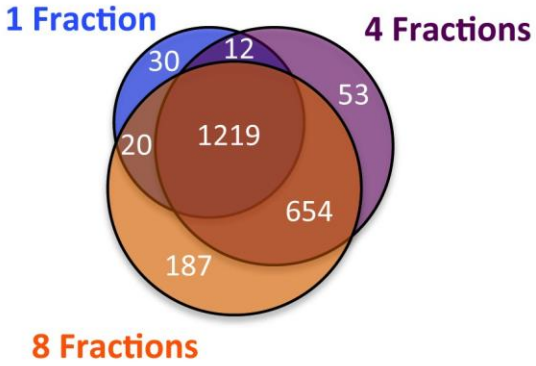
Number of identified proteins upon using different peptide fractionation schemes

# of OGE Fractions	# of Proteins Identified[1]	Measurement Time/Sample[2]	Protein IDs/h
-	1281	3	427
4	1938	12	161.5
8	2080	24	86.67

1) FDR=1% based on the number of decoy hits in the database

2) 3 hours per LC-MS anaylsis (2 hours gradient + 1 hour for sample loading/column washing)

B



Supplementary Figure 3: Impact on different peptide fractionation schemes on the number of identified proteins and the required running time per sample. (A) Two different pooling schemes were applied to the 12 peptide fractions obtained after Off-Gel electrophoresis (OGE) peptide separation resulting in 8 and 4 peptide fractions per sample. The number of protein identification together with the measurement time required per sample is given for an unfractionated sample as well as for the same sample fractionated into 4 and 8 OGE fractions, respectively. Based on the number of identifications and LC-MS measurement time, 4 fractions were considered to be the best compromise regarding comprehensiveness and analytical efforts, and therefore were used in this study. (B) Venn diagram showing the number of overlapping and exclusive protein identifications. FDR: false-discovery-rate.

Supplementary Figure 4:

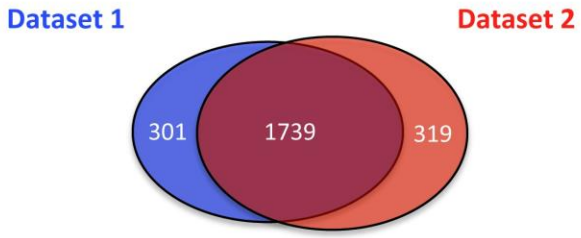
A

Properties of the two large LC-MS dataset generated

Dataset	1	2
MS/MS spectra acquisition mode	CID	HCD
OGE-Fractionation	Yes	No
Total Number of LC-MS runs	99	72
Number of Biological Replicates	1	3
# of Acquired MS/MS Spectra	1,997,474	3,244,959
# of Peptide Spectrum Matches[1]	918,544	1,105,784
# of Unique Identified Protein Clusters[1]	2,308	2,205
# of Unique Quantified Protein Clusters[2]	2,040	2,058
# of Unique Quantified Membrane Protein Clusters[3]	122	310
Quantified Membrane Protein Clusters of Total (%)[4]	6.0	15.1
# of Unique Identified Protein Clusters (combined)[5]	2,571	
# of Unique Quantified Protein Clusters (combined)[5]	**2,359**	

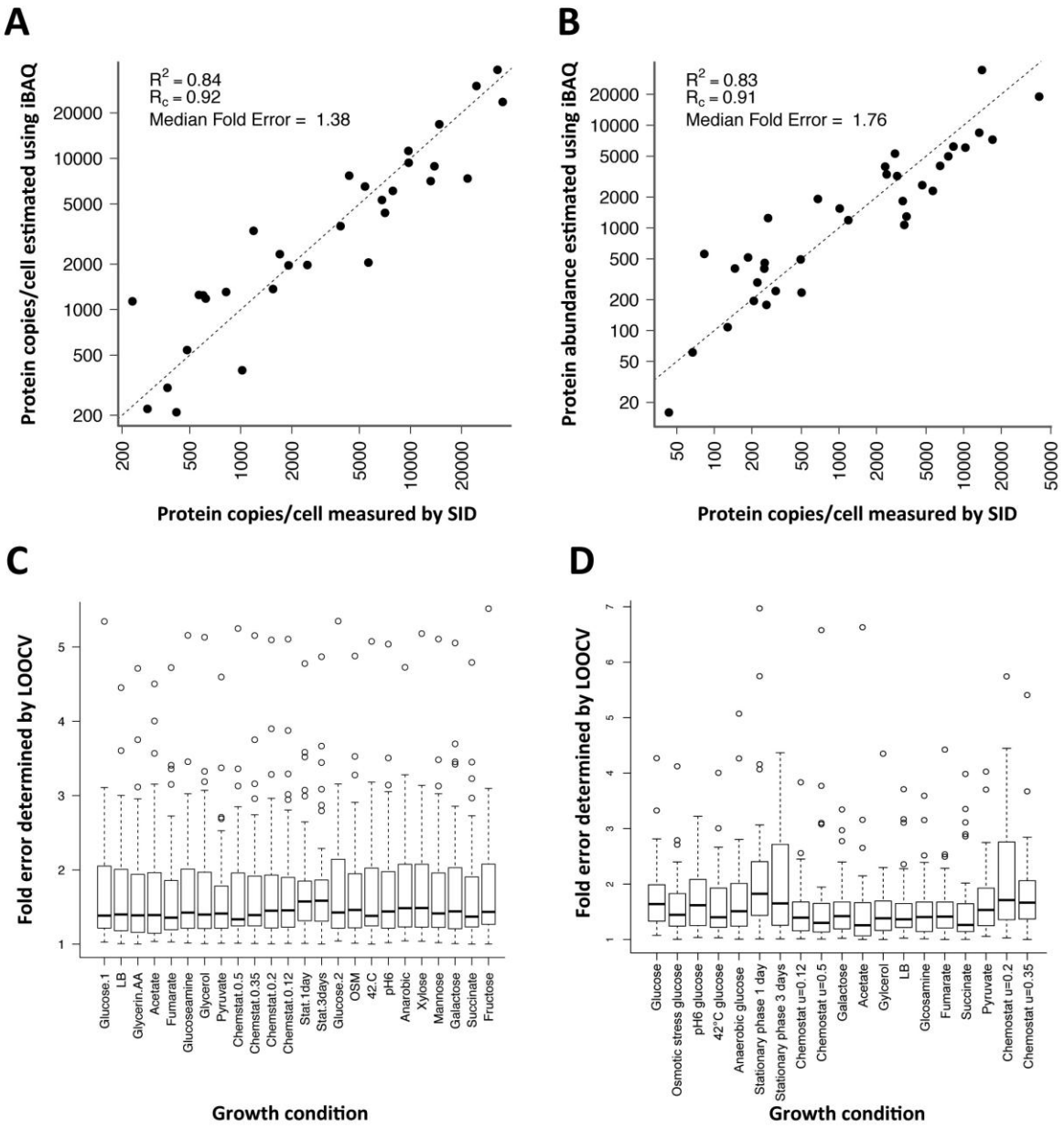
1) FDR was set to 1% based on the number of decoy hits in the dataset
2) Proteins with estimated absolute protein abundances using iBAQ
3) Membrane proteins classified as proteins with at least 1 predicted transmembrane helix using the TMHMM algorithm
4) Ratio of membrane proteins quantified to all quantified proteins
5) Total number of proteins identified and quantified in both datasets

B



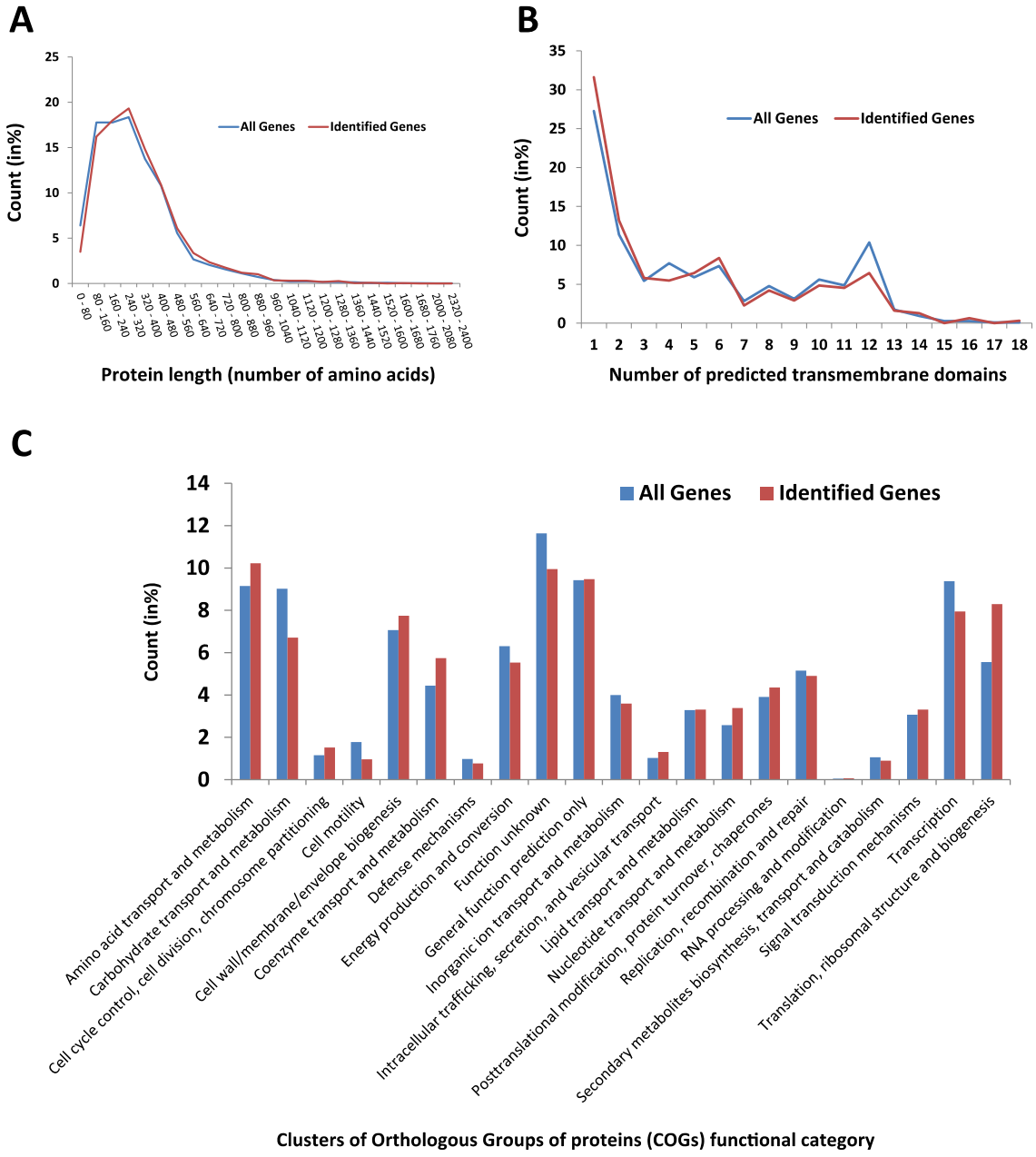
Supplementary Figure 4: Properties of the two different large-scale quantitative LC-MS datasets included in this manuscript. (A) Summary of the specific properties of the two LC-MS datasets acquired in this study. The main differences included peptide fragmentation mode (collision-induced dissociation (CID) in a linear ion trap or higher energy collisional dissociation (HCD) in a c-trap), sample fractionation using off-gel electrophoresis (OGE) and the number of biological replicates analyzed. Naturally, the high sample numbers generated by OGE fractionation impeded the analysis of biological replicates for 20+ conditions in a reasonable time frame. Additionally, due to the low number of quantified membrane proteins in the first acquired dataset (data set 1) the urea/RapiGest lysis buffer was exchanged by a sodium deoxycholate (DOC) containing buffer (see material and methods for details) that has been shown to be well suited for the analysis of membrane proteins, in particular located on the outer membrane[7]. Indeed, the proportion of quantified membrane proteins was 3 times higher in data set 2 using a DOC based lysis buffer. Therefore, and to provide statically controlled quantitative values, condition and growth rate-dependent quantitative results provided in this study are solemnly based on data set 2, where each condition was analyzed in biological triplicates. (B) Venn diagram showing the number of overlapping and exclusive quantified protein clusters for the two data sets. FDR: false-discovery-rate, iBAQ: intensity-based absolute quantification[1,5], TMHMM: transmembrane protein topology with a hidden Markov model[1,2].

Supplementary Figure 5:



Supplementary Figure 5: Fold error estimation of determined protein concentrations. The correlation of the actual cellular abundances of 41 selected proteins (in copies/cell, see Supplementary Tables 1-3 for details) determined by selected reaction monitoring and stable isotope dilution (SRM/SID)[3,8] and the intensity-based absolute quantification (iBAQ) values[2,5] determined by label-free quantification (both in logarithmic scale) from dataset 2 (A) and 1 (B) for cells grown in glucose minimal media are shown (see Supplementary Figure 4 for data set details). The squared Pearson correlation coefficients (R^2), the Lin's Concordance Correlation coefficient (R_c) and the median fold error are displayed (for details see supplemental text). (C) Fold errors determined by leave one out cross validation (LOOCV) as box plots for all growth conditions included in data set 2. The black bar indicates the median fold error. (D) like (C) for data set 1.

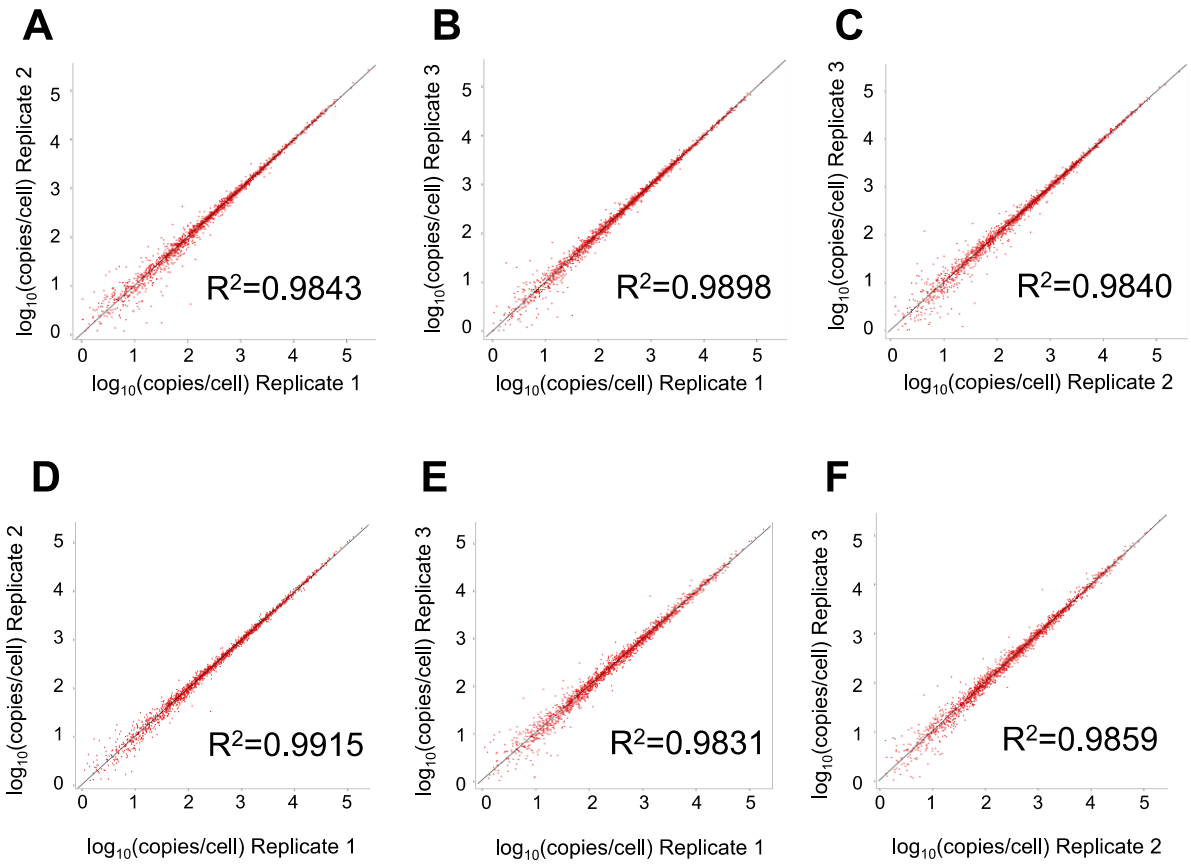
Supplementary Figure 6:



Supplementary Figure 6: Proteome coverage assessment. Relative distribution of all genes predicted by genome annotation and used in this study (shown in blue) and all genes which proteins could be identified from all predicted genes using this protein database (shown in red) according to (A) protein length, (B) number of transmembrane proteins predicted by TMHMM algorithm[1,4] and (C) COG functional categories[2,5].

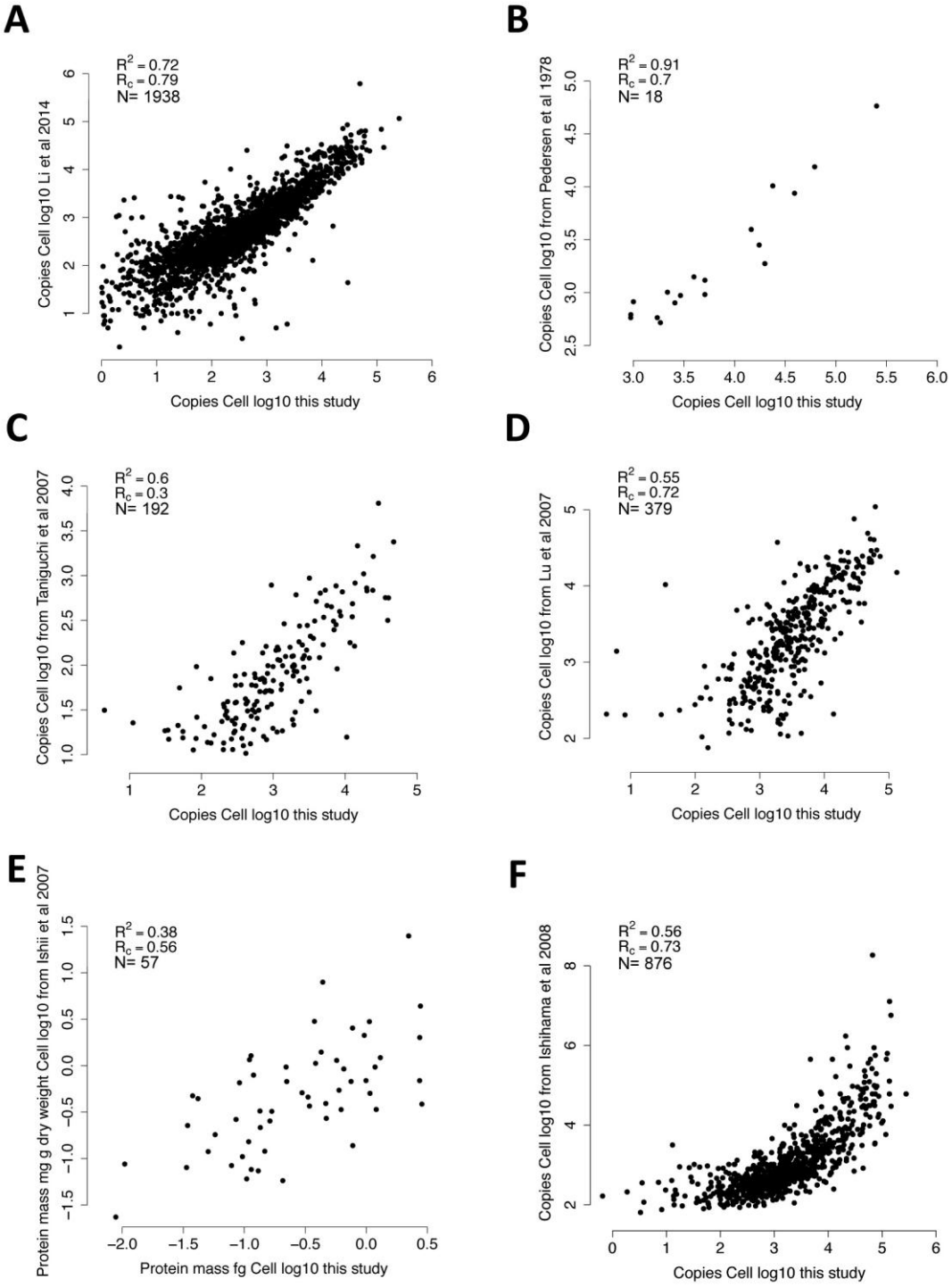
Nature Biotechnology: doi:10.1038/nbt.3418

Supplementary Figure 7:



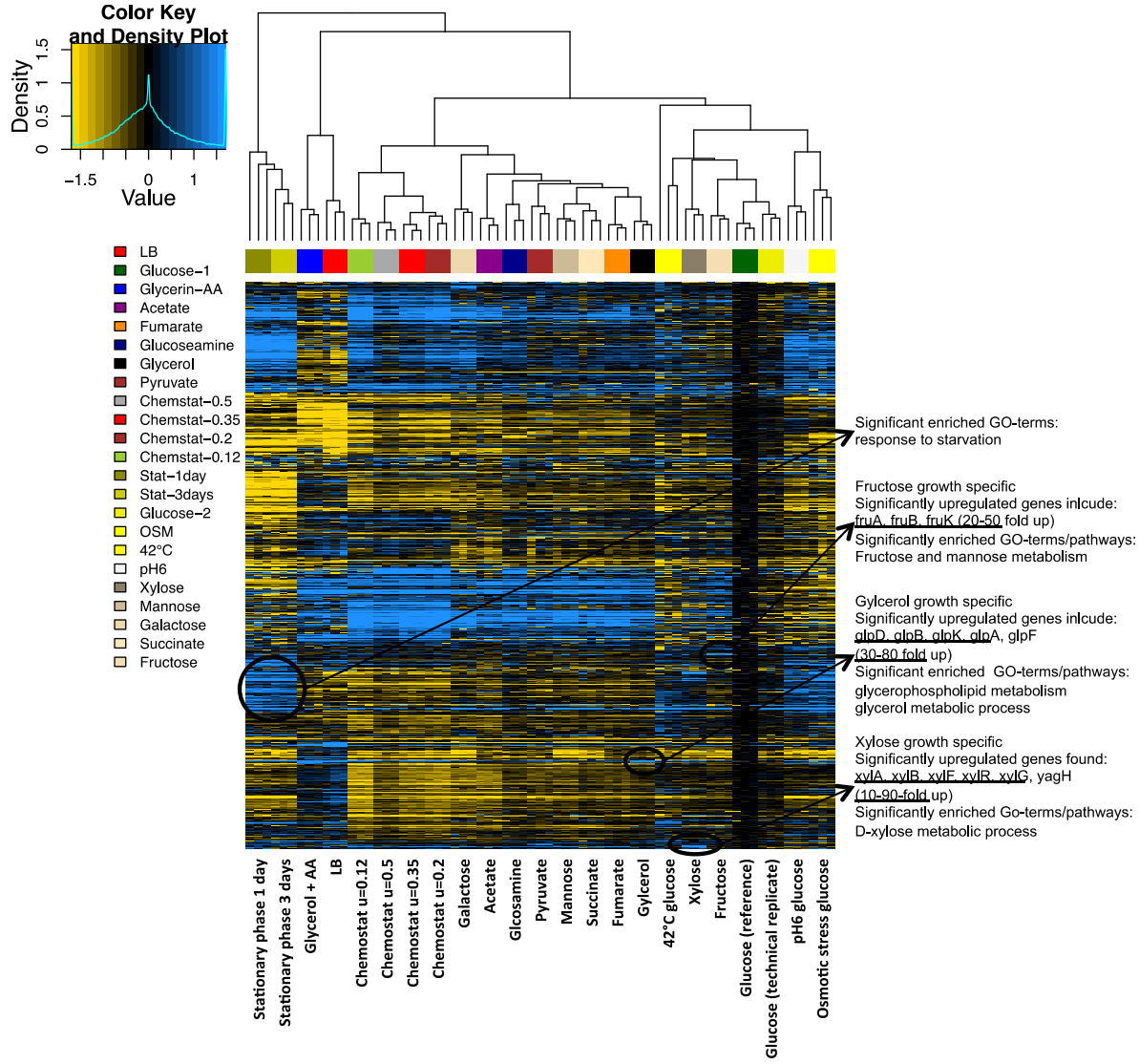
Supplementary Figure 7: Evaluation of the technical and biological reproducibility of our absolute quantification approach. Correlation of estimated absolute protein abundances (in copies per cell) for all proteins quantified by the workflow applied (see Figure 1) for three biological replicates of cells grown in glucose media (A-C) and chemostat conditions at a growth rate of 0.5 h^{-1} (D-F). The determined squared Pearson correlation coefficients (R^2) are shown for each plot.

Supplementary Figure 8:



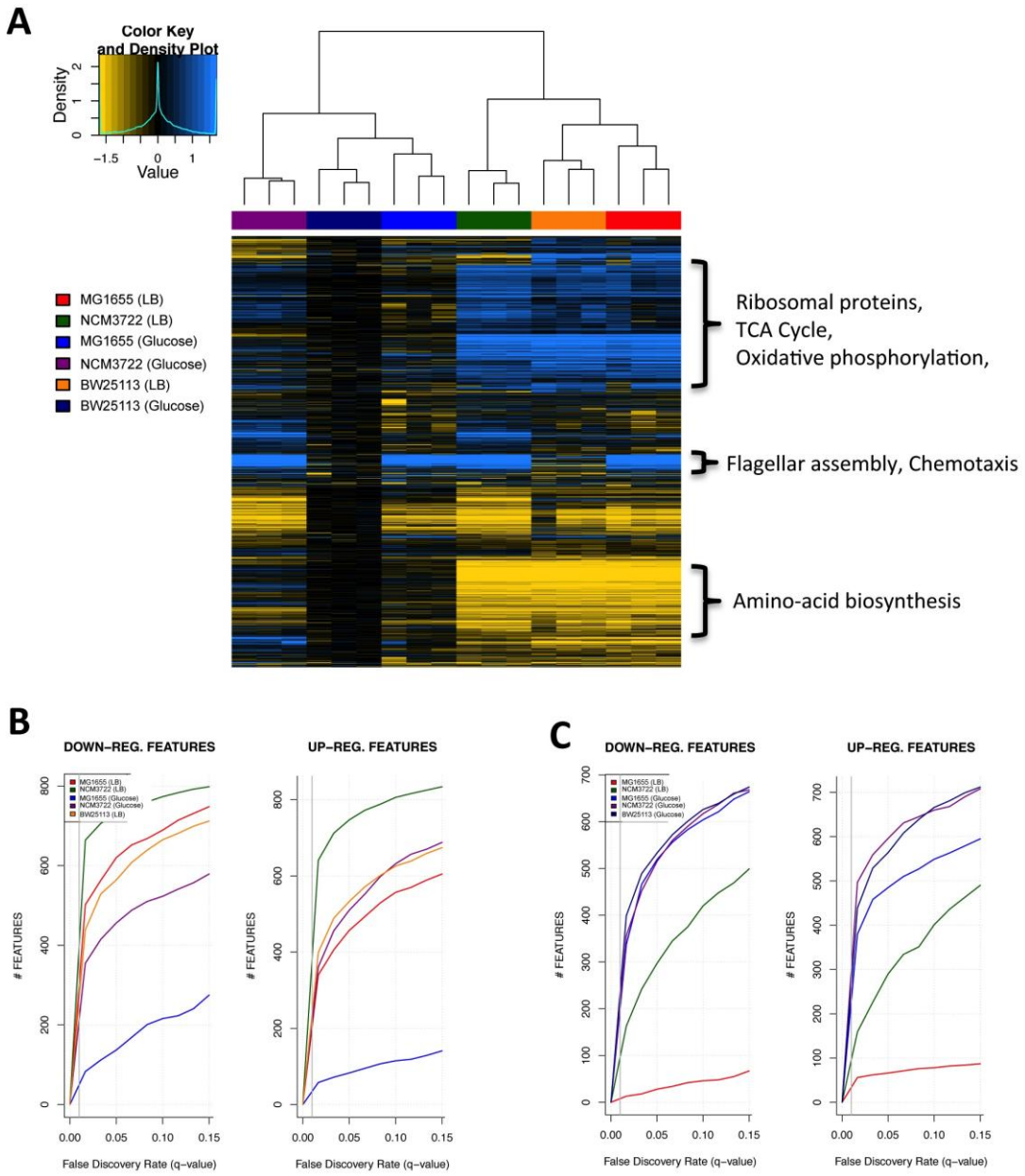
Supplementary Figure 8: Correlation of our absolute protein abundance estimates with various published small datasets including only a few growth conditions. Specifically, the quantitative data of this study was compared to Li *et al.*[6] (A), Pedersen *et al.*[7,9] (B), Taniguchi *et al.*[10] (C), Lu *et al.*[11] (D), Ishii *et al.*[12] (E) and Ishihama *et al.*[13] (F). For each plot, the squared Pearson correlation coefficients (R^2), the Lin's Concordance Correlation coefficient (R_c) and number of values (N) are displayed (for details see supplemental text).

Supplementary Figure 9:



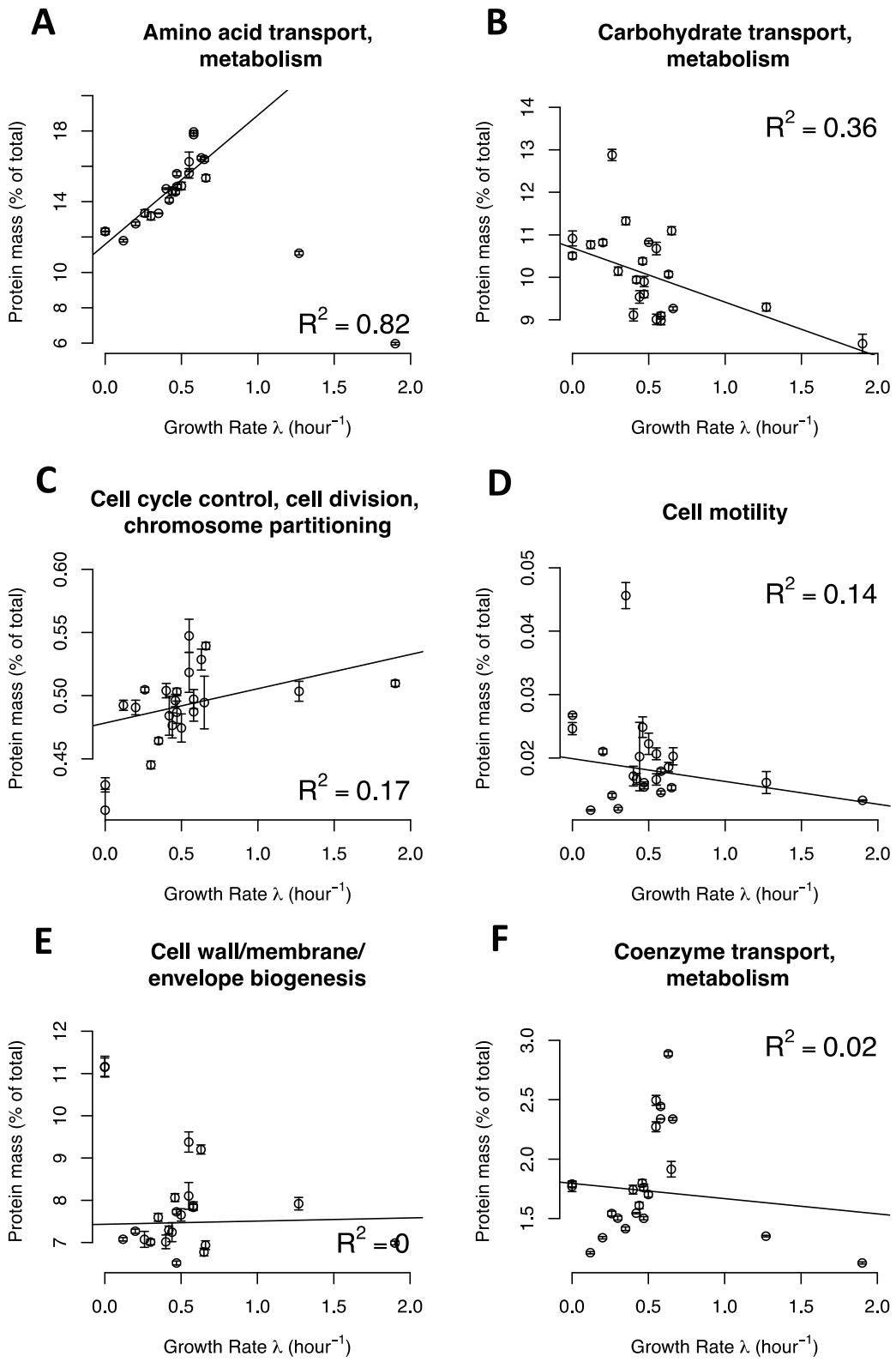
Supplementary Figure 9: Heatmap of relative protein abundance changes (to glucose) of all samples. In detail, hierarchical clustering of protein log2 abundance ratios determined using the SafeQuant software (v2)[3] (see Supplementary Table 8) was preformed using Ward's algorithm and the Pearson Correlation distance metric. Subsequently, a heatmap was created using the gplots R package (http://cran.r-project. org/package=gplots). The dendrogram illustrates similarity of protein abundance patterns. The genes of the most prominent condition-specific and significantly up-regulated protein groups are shown on the right together with their corresponding significantly enriched GO-terms/pathways (p-value (Benjamini corrected) <0.05) determined using the DAVID algorithm[14].

Supplementary Figure 10:



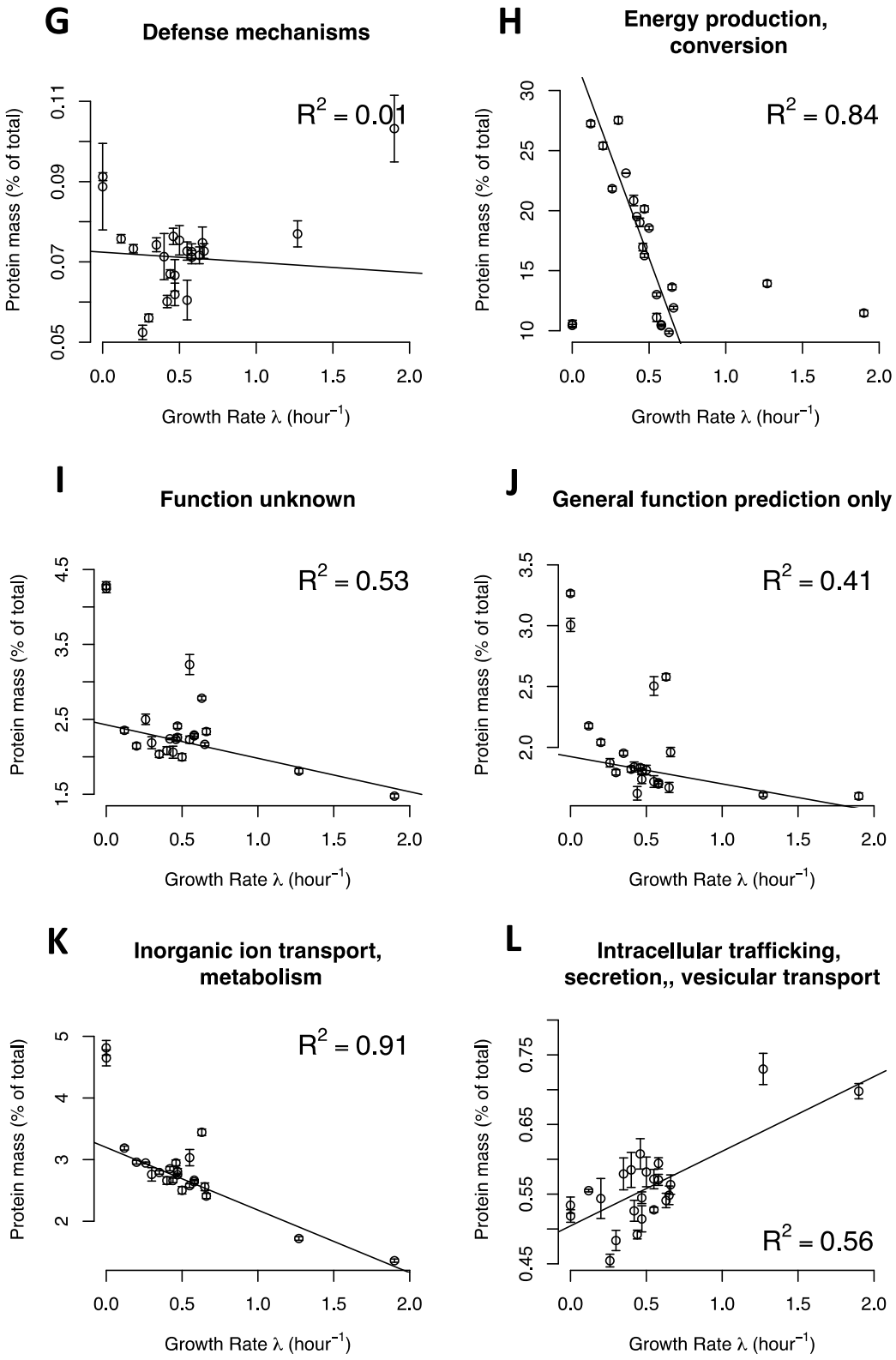
Supplementary Figure 10: Quantitative proteome comparison of three *E. coli* strains (BW25113, MG1655, NCM3722) grown in minimal (glucose) and rich (LB) media. (A) Heatmap of absolute protein levels obtained for each strain and growth condition analyzed in biological triplicates. In detail, hierarchical clustering of protein log2 abundance ratios (see Supplementary Table 9) was preformed using Ward's algorithm and the Pearson Correlation distance metric. Subsequently, a heatmap was created using the gplots R package (http://cran.r-project. org/package=gplots). The dendrogram illustrates similarity of protein abundance patterns. Protein groups with significant protein abundance differences are indicated together with their corresponding significantly enriched GO-terms (p-value < 0.05, Benjamini corrected) determined using the DAVID algorithm[14]. (B) The numbers of significantly up- and down-regulated proteins for each growth condition using BW25113 (glucose minimal media) as the control condition are shown as a function of q-values as determined by the SafeQuant software (v2)[3] (see statistical section below for details). (C) Like (B) using BW25113 grown in LB medium as the reference condition.

Supplementary Figure 11:

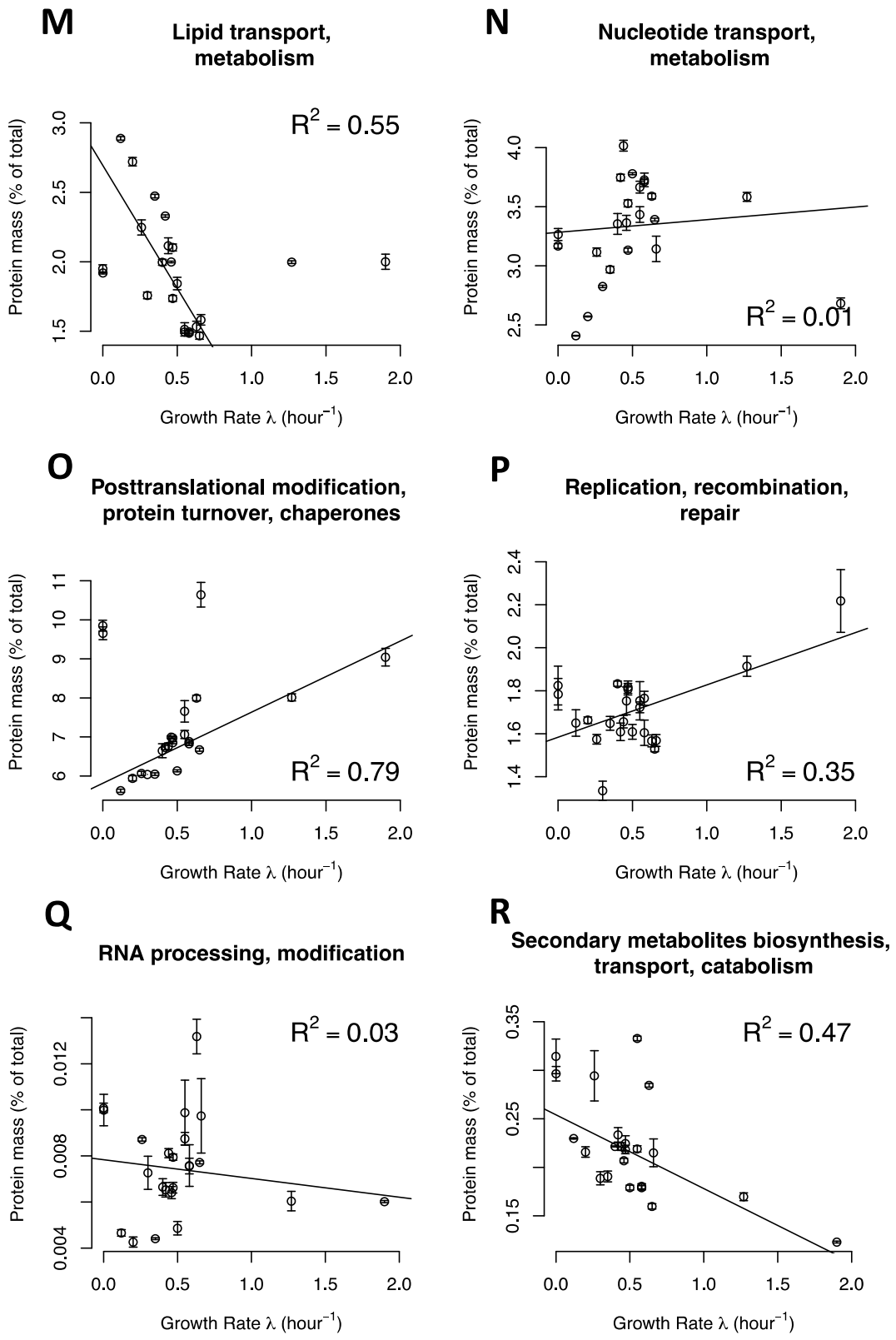


Nature Biotechnology: doi:10.1038/nbt.3418

Supplementary Figure 11:

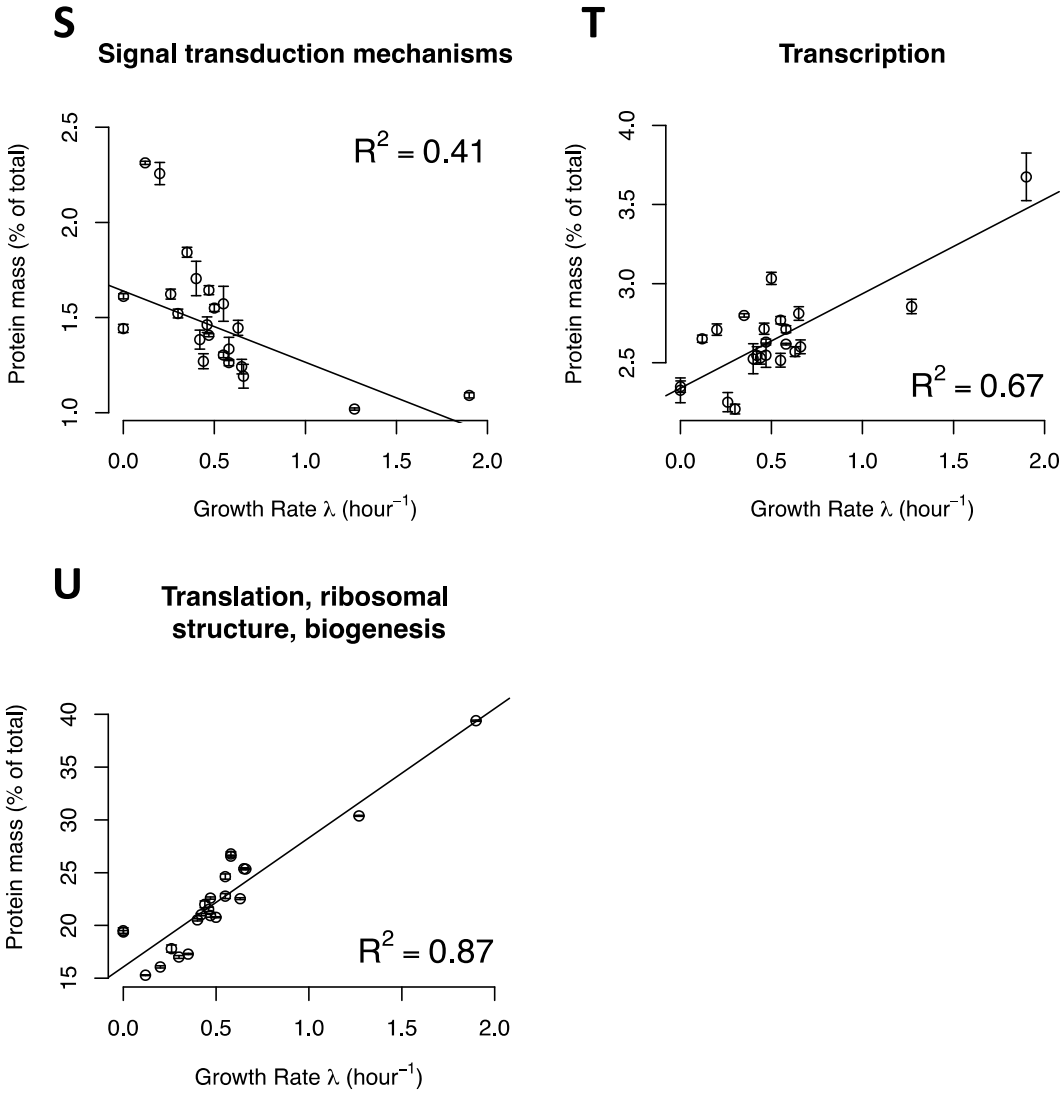


Supplementary Figure 11:



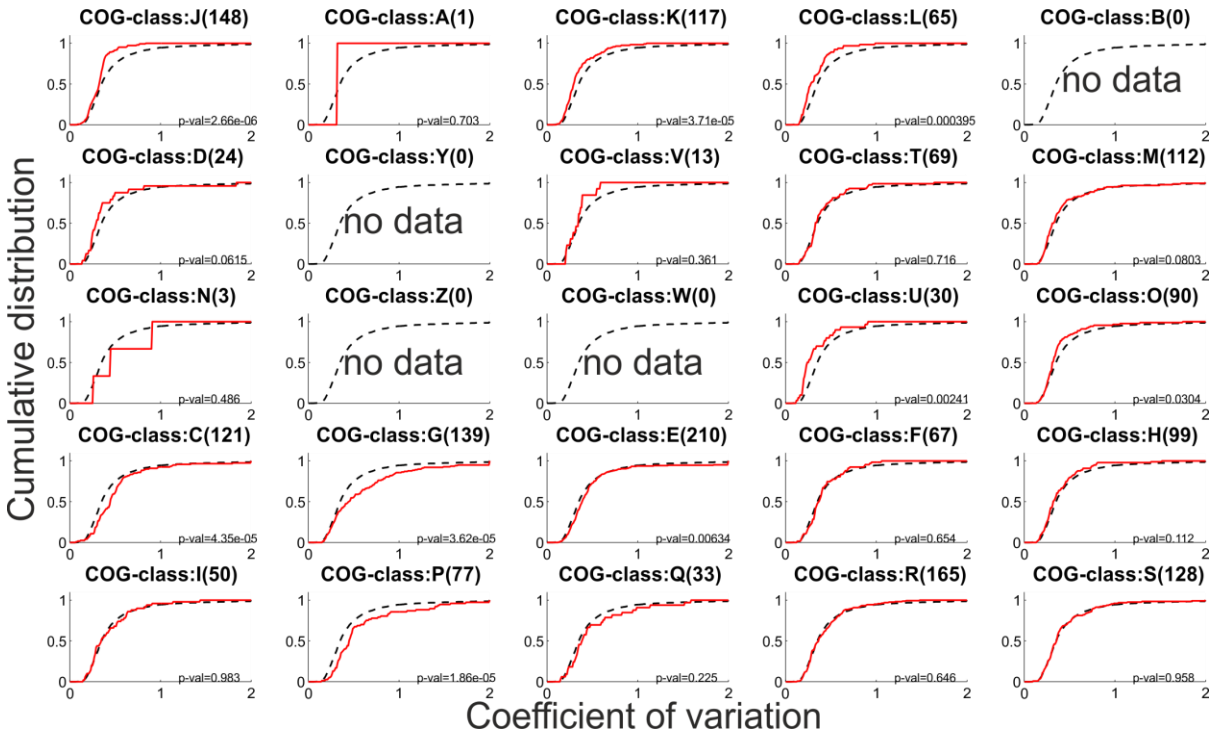
Nature Biotechnology: doi:10.1038/nbt.3418

Supplementary Figure 11:



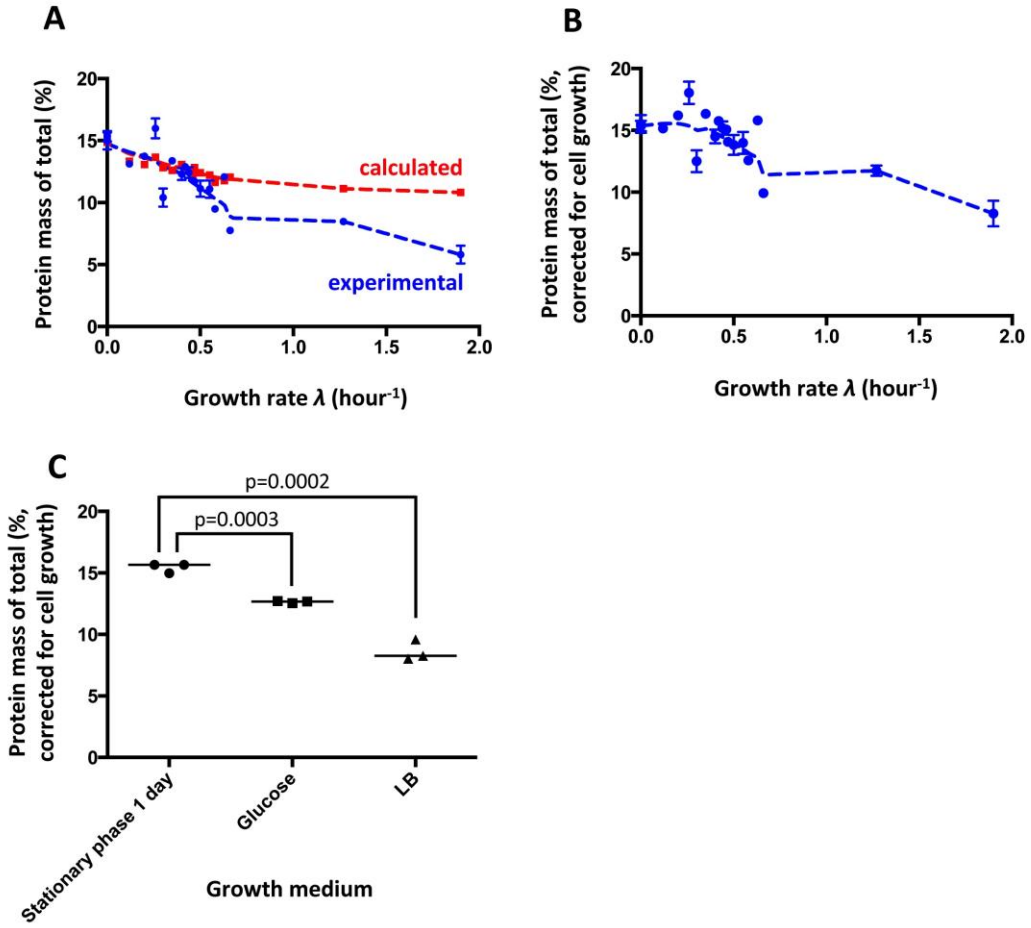
Supplementary Figure 11: Correlation between the growth rate and the relative protein mass fractions for all proteins assigned to the different COG categories. The correlation between the growth rate and the relative protein mass fractions for all proteins assigned to the different COG categories[2], respectively, are shown. The standard deviations are indicated for each data point and the determined squared Pearson correlation coefficients (R^2) are calculated for all conditions. For all plots, a robust linear regression was applied (see supplemental note 2 for details).

Supplementary Figure 12:



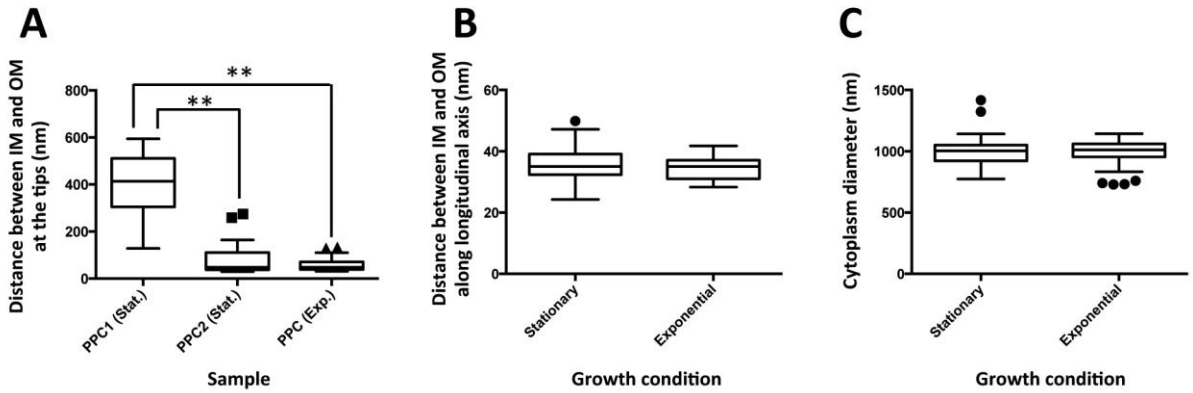
Supplementary Figure 12: **Cumulative distribution of the coefficient of variation for all COG classes.** Cumulative distribution of the coefficient of variation (CV) for all COG classes (red line) compared to the whole detected proteome (black dashed lines). Protein concentrations were calculated from protein copy numbers and cell volumes for all conditions. For each protein, the coefficient of variation was calculated as its relative standard deviation across conditions using only conditions, in which this protein was reliably quantified (relative error of quantification < 30%). Only proteins, for which more than 50% of the conditions yielded reliable quantification were used. The significance of the difference in median CV between each COG class and the whole proteome was assessed by a two-sided Wilcoxon rank sum test, and the respective p-value is reported for each class. Classes for which no protein was quantified are labeled as “no data”. The numbers of proteins associated to each COG class are illustrated in brackets. Abbreviations of COG classes: **J**: “Translation, ribosomal structure and biogenesis”; **A**: “RNA processing and modification”; **K**: “Transcription”; **L**: “Replication, recombination and repair”; **B**: “Chromatin structure and dynamics”; **D**: “Cell cycle control, cell division, chromosome partitioning”; **Y**: “Nuclear structure”; **V**: “Defense mechanisms”; **T**: “Signal transduction mechanisms”; **M**: “Cell wall/membrane/envelope biogenesis”; **N**: “Cell motility”; **Z**: “Cytoskeleton”; **W**: “Extracellular structures”; **U**: “Intracellular trafficking, secretion, and vesicular transport”; **O**: “Posttranslational modification, protein turnover, chaperones”; **C**: “Energy production and conversion”; **G**: “Carbohydrate transport and metabolism”; **E**: “Amino acid transport and metabolism”; **F**: “Nucleotide transport and metabolism”; **H**: “Coenzyme transport and metabolism”; **I**: “Lipid transport and metabolism”; **P**: “Inorganic ion transport and metabolism”; **Q**: “Secondary metabolites biosynthesis, transport and catabolism”; **R**: “General function prediction only”; **S**: “Function unknown”.

Supplementary Figure 13:



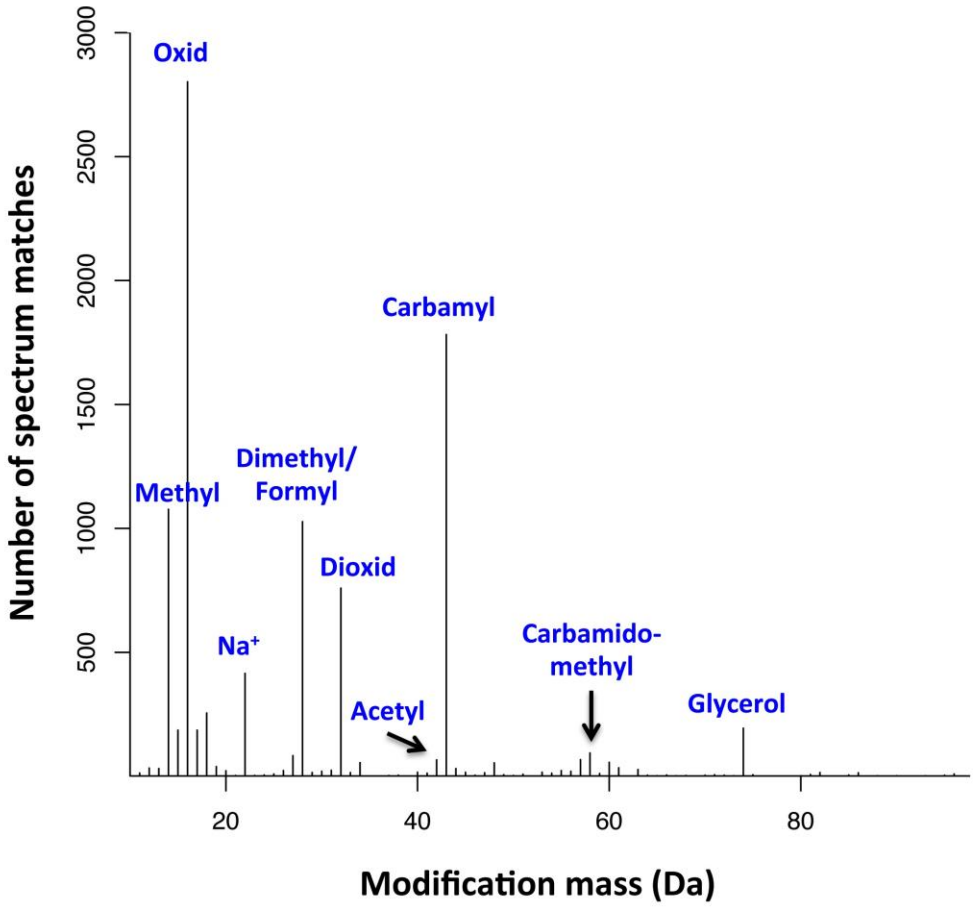
Supplementary Figure 13: Periplasmic protein mass distribution geometrically corrected for increase in cell size with growth rates. (A) Ratio of periplasmic protein mass to total protein mass (in %) for the different growth conditions (blue) as well as the calculated reduction of periplasmic space expected for fast growing cells due to the increase in cell size (red). Here, the relative proportion of periplasmic space was calculated using the measured cell dimensions (see Supplementary Table 23) and dividing the periplasmic (pi*(w-p)^2*((l-2*p)-(w-p)/3)/4) and the total (pi*w^2*(l-w/3)/4) cell volume (w=cell width, l=cell length, p=periplasmic width). The growth condition with the highest ratio of periplasmic protein and slowest growth (stationary phase 1) was selected as a reference and the periplasmic width was set to 54 nm on average to match the experimental periplasmic protein ratio (see Supplementary Table 28). Of note, the periplasmic width was only 28 nm on average for fast growing cells and was in good agreement with previous electron microscopy studies[15]. Error bars with standard deviation across biological triplicate measurements for the experimental data as well as Lowess curves are indicated. (B) Same as (A) for the observed decrease in relative periplasmic protein mass corrected for the geometric changes calculated. (C) Relative geometric-corrected periplasmic protein mass for three different growth conditions. Mean (black line) as well as individual values for each biological replicate are shown. The calculated significance (p-value, t-Test; two-tailed distribution assuming equal variance; homoscedastic) to the reference condition (stationary phase 1 day) are also indicated.

Supplementary Figure 14:



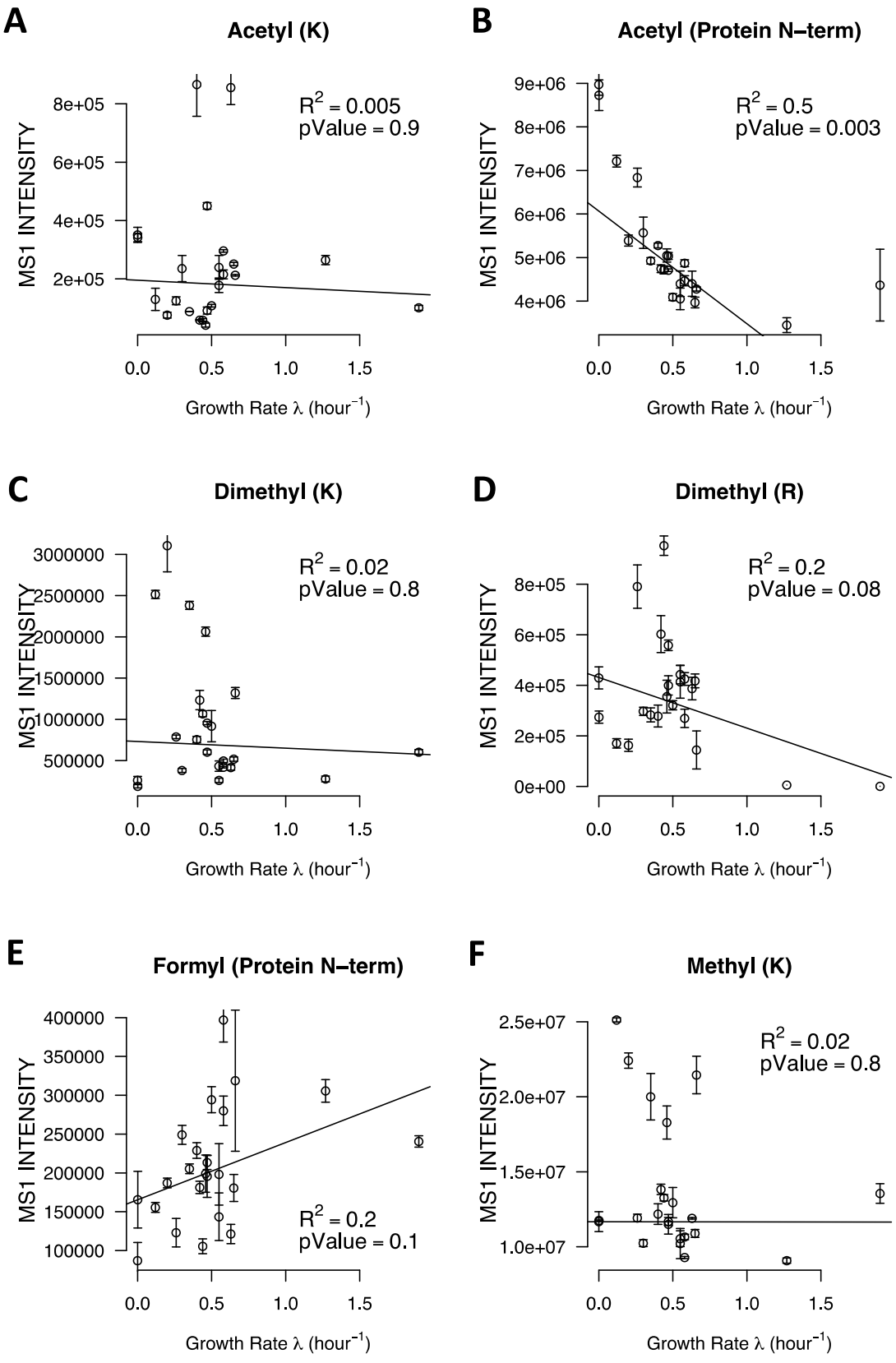
Supplementary Figure 14: Cryo-electron microscopy analysis of *E. coli* cells. *E. coli* cells were grown exponentially on LB medium ('Exp.' or 'Exponential'), or, to 3-day stationary phase ('Stats.' or 'Stationary') after a glucose culture. (A) Box plot showing the distribution of distances between the inner- (IM) and outer-membrane (OM) at the tip of the periplasmic cone (PPC). For the stationary phase cells, the largest cone was denoted PPC1 whereas the smallest cone was denoted PPC2. (B) Box plot showing the distribution of distances between the inner- and outer-membrane along the longitudinal axis of the cell. (C) Box plot showing the distribution of diameter of the cytoplasm. Data was determined from 25 cells. Each box spans the interquartile range. The notches extend to the most extreme data point, which is no more than 1.5 times the interquartile range from the box. The thick horizontal line in each box indicates the median. The calculated significance (t-Test; two-tailed distribution assuming equal variance; homoscedastic), p-values <0.01 (**)) are indicated. For further details see Supplementary Table 29.

Supplementary Figure 15:

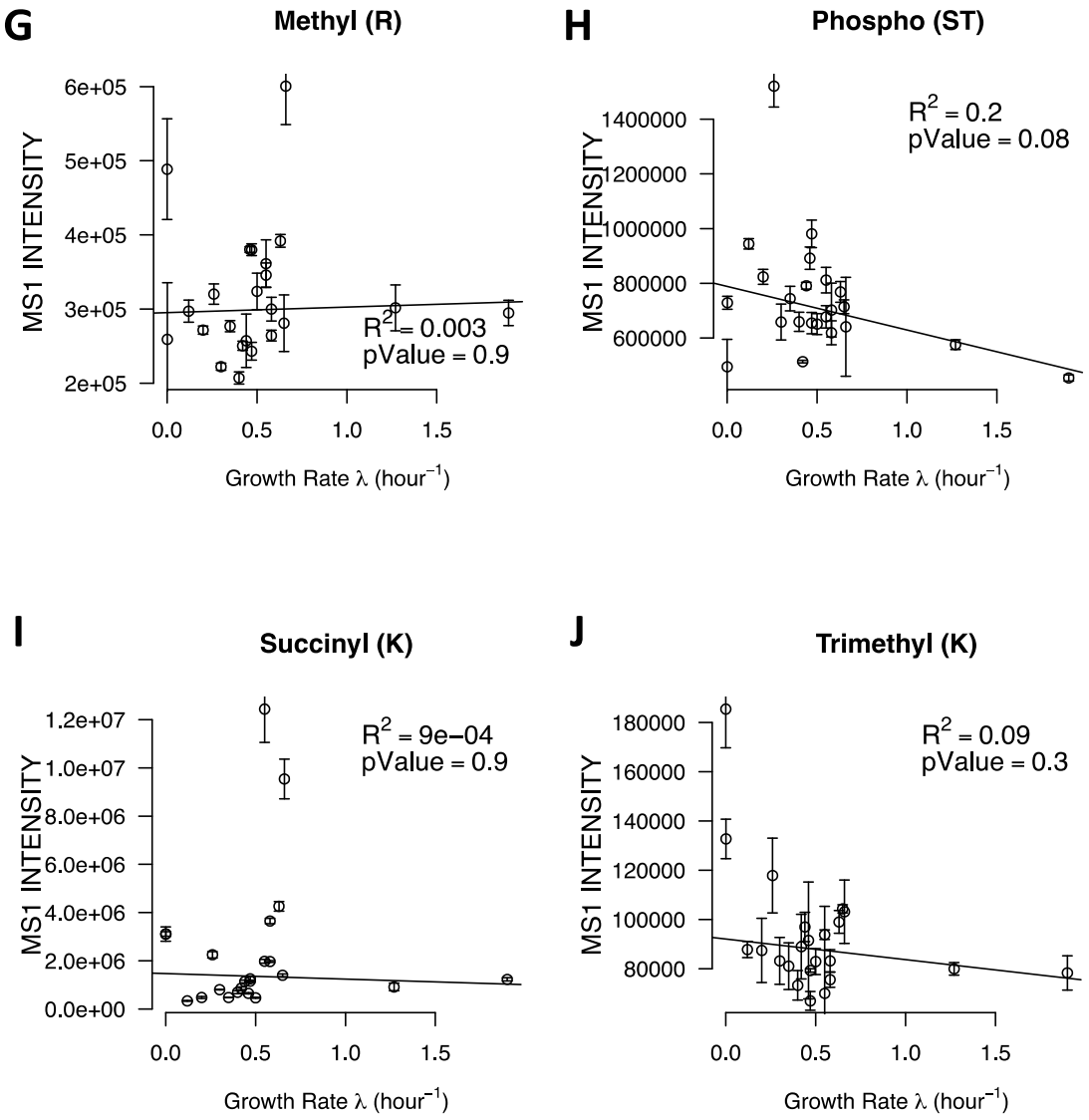


Supplementary Figure 15: Unrestricted open modification search of all unassigned MS/MS-spectra. Unrestricted open modification search of all unassigned MS/MS-spectra against a spectrum library database using the QuickMod search algorithm[4]. The mass difference and numbers of precursor ion masses between spectrum matches of unmodified and modified peptides are shown. It is important to note that sodium adducts, carbamidomethylation (found on C residues, which are due to double carbamidomethylation) and glycerol (at D/E residues introduced during OFFGEL electrophoresis[16]) can be considered *in vitro* artifacts (see www.uniprot.org for details). The other modifications can be either introduced posttranslationally (PTM) or artificially. In more detail, oxidation was mostly detected at methionine residues in this data set, which is considered an *in vitro* artifact, but oxidation can be a PTM when located at D, K, N, P, Y, R or C residues. Methylation is a PTM in most cases, including the mostly detected methylated K and R residues, but can also be artificially introduced at the peptide C-term. Dimethylation was only detected at K and R residues, which are considered PTMs. Formylation is considered a PTM at protein N-term, which was the exclusive modified position found. Acetylation was mostly detected at K residues, which can be introduced posttranslationally or artificially and at protein N-term, which is considered a PTM. Carbamylation was mostly detected at K residues and peptide N-term, which was found to be either an artificial modification, formed by a non-enzymatic reaction with breakdown products of urea, or a PTM[17].

Supplementary Figure 16:

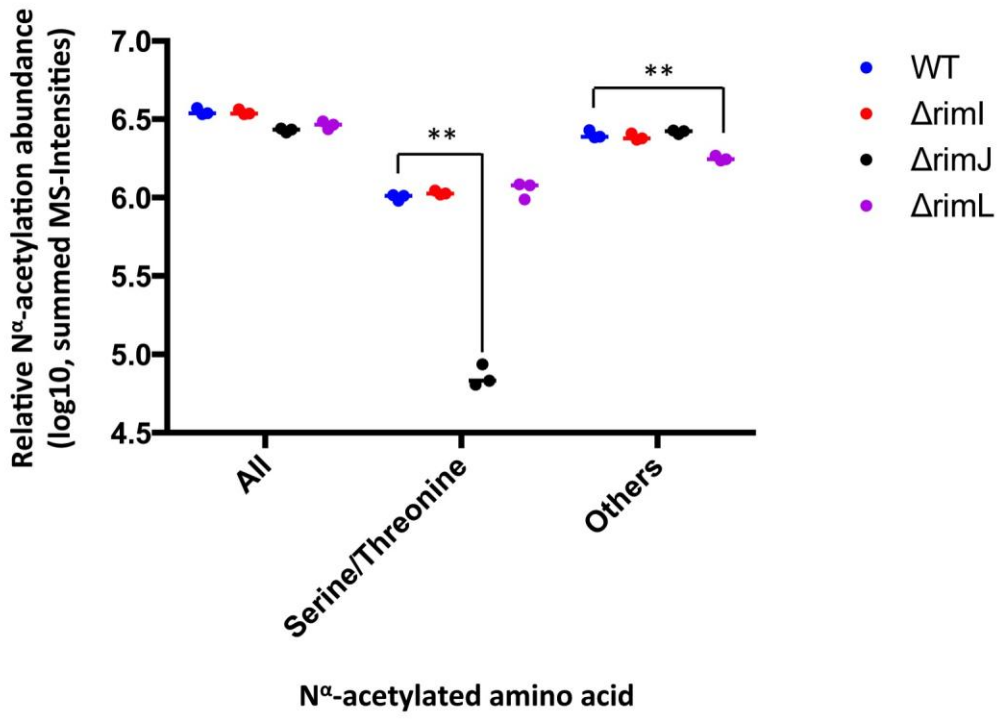


Supplementary Figure 16:



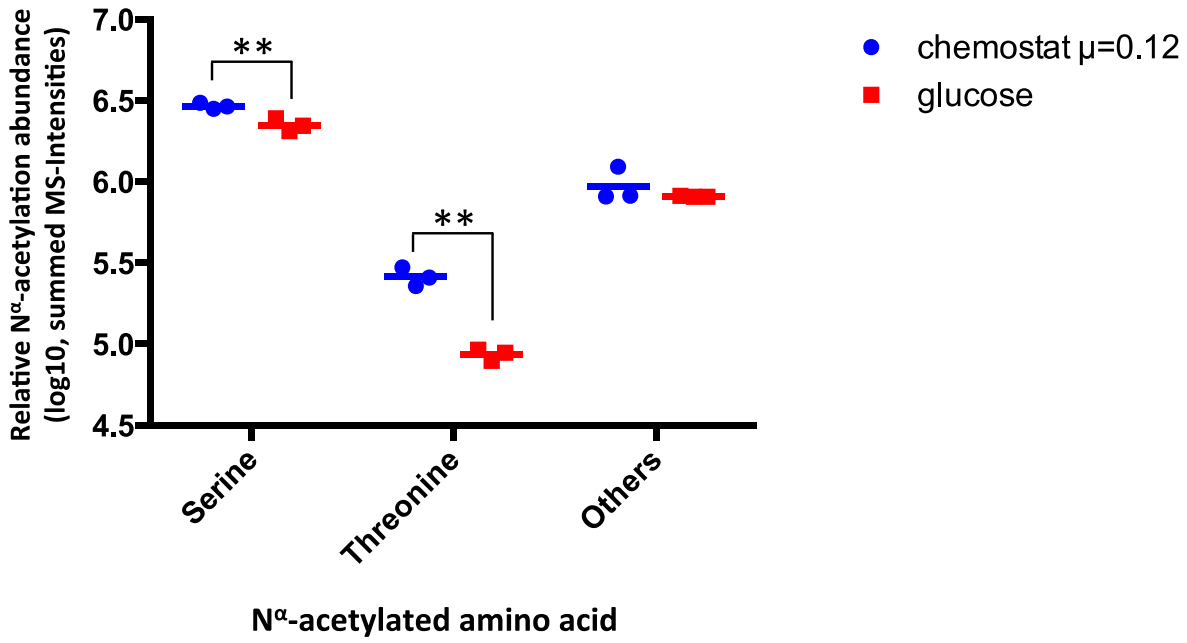
Supplementary Figure 16: Correlation of the growth rate and the relative abundance of all identified peptides carrying a specific modification, respectively. The standard deviations are indicated for each data point and the determined squared Pearson correlation coefficients (R^2) are calculated for all conditions. A robust linear regression was applied in all plots and corresponding p-values (Benjamini-Hochberg-corrected) are shown (see supplemental text for details).

Supplementary Figure 17:



Supplementary Figure 17: Relative change in abundance of the identified N^{α}-acetylation sites for wild type (WT) and mutant strains lacking the three known N-acetyltransferases annotated in the *E. coli* genome (ΔrimI, ΔrimJ and ΔrimL), respectively. All MS intensities of acetylated peptides of all N-termini (All), serine and threonine or other amino acids were summed for the different mutant strains grown on glucose. Relative abundances as sum of all MS intensities of acetylated peptides located at all, serine and threonine or any other amino acid are illustrated as log10 values. The MS-intensities of all N^{α}-acetylated peptides were normalized using the corresponding protein MS-intensities and are illustrated in Supplementary Table 21. Interestingly, we grew the wildtype and the deletion mutants of the known N-acetyltransferases (rimL, rimJ, rimI) on glucose and acetate – two conditions, where we found different degrees of N-acetylations. Here, we found that the rimI deletion had basically no effect on the growth rate, the rimL deletion resulted in a severe growth defect on acetate, and the rimJ deletion on both carbon sources (Supplementary Table 24). The median value and the calculated significance of mutant strains to WT (t-Test; two-tailed distribution, two-sample assuming equal variance, performed on log transformed abundances); p-values <0.01 (**) are indicated.

Supplementary Figure 18:



Supplementary Figure 18: Relative change in abundance of the identified N^{α}-acetylation sites according to the modified amino acid at the protein N-terminus. All MS intensities of acetylated peptides located at a specific amino acid were summed for glucose (red) and the growth condition that supported slowest growth (chemostat μ=0.12, blue), where we observed the highest abundance of N^{α}-acetylation sites (Figure 5C). Relative abundances as sum of all MS intensities of N^{α}-acetylated peptides located at a specific amino acid are illustrated as log10 values for the most frequently observed N-terminal amino acids (cf. Figure 5B), serine and threonine as well as the combined values for the other acetylated peptides identified including N-terminal methionine and alanine residues. The mean value and the calculated significance (t-Test; two-tailed distribution, two-sample assuming equal variance, performed on log transformed abundances); p-values <0.01 (**) are indicated.

Nature Biotechnology: doi:10.1038/nbt.3418

Supplementary Notes

Supplementary Note 1:

Dataset quality assessment

We evaluated the achieved coverage with regard to potential biases. Although, very short and hydrophobic proteins are notoriously difficult to identify by MS-based proteomics because of the limited number of tryptic peptides amenable for LC-MS/MS analysis[18,19], we found no bias against small (Supplementary Figure 6A) or hydrophobic proteins even for proteins with multiple predicted transmembrane domains (Supplementary Figure 6B). Also on a functional level no bias against any functional protein class was observed (Supplementary Figure 6C). Thus, the dataset represents a comprehensive, unbiased representation of the *E. coli* proteome.

Next, we evaluated the accuracy and reproducibility of the protein quantities determined in our quantitative data. Therefore, we first assessed the degree of technical and biological variability and applied the protein abundance estimation strategy to three independent batches of *E. coli* cells. When plotting estimated protein levels in copies/cell obtained from three independent biological replicate samples grown on glucose excess (Supplementary Figure 7A-C) and chemostat (μ=0.5) conditions (Supplementary Figure 7D-F) against each other, we found very high squared Pearson correlation coefficients R^2 (>0.98) indicating nearly linear relationship between all replicates. As can further be noted from these scatter plots, the variability of signal intensities among replicates is very low across the whole abundance range.

Towards gaining evidence for the accuracy of the determined absolute protein abundances, we compared the obtained protein concentrations with those of six different published quantitative datasets[9-12,20]. Good correlations were observed, in particular with protein quantities determined not with LC-MS approaches but with classical radioactivity-[9], ribosomal profiling[6] and fluorescence-based[10] assays (Supplementary Figure 8A-C). The somewhat lower agreement with previous LC-MS datasets (Supplementary Figure 8D-F) is likely a result of the semi-quantitative spectral count-based quantification strategy used therein, which has shown to be of lower accuracy and linear range than the intensity-based quantification we used to generate our dataset[21-23]. In particular, the quantities for some high abundant proteins reported by Ishihama *et al.*[20] are several orders higher compared to our and the other five studies (Supplementary Figure 8F). These proteins comprise mostly ribosomal proteins that are, with an average abundance of 4.18×10^6 copies per cell, much higher than reported numbers of ribosomes in *E. coli* from other sources ranging between 6,800 and 72,000 molecules per cell in low and fast growing cells, respectively[24]. The most abundant ribosomal protein (rpmG) alone would have a protein weight of 1.96 pg per cell that is several fold higher than the reported total protein mass per *E. coli* cell being between 100 and 450 fg[25].

As further evidence for the quality of our quantification, we used stoichiometries of protein complexes. When comparing abundances of proteins being present in stable protein complexes, we found that the protein abundances matched the expected values within calculated error rates for most proteins, including hydrophobic membrane (e.g. outer membrane protein assembly complex) and low abundant (e.g. ethanolamine ammonia-lyase) protein complexes (Supplementary Table 27). We also observed highly similar levels for all ribosomal proteins (Supplementary Table 6) and

provide the stoichiometries of the 30S and 50S ribosomal subunit that matched the expected 1:1 ratio across all 22 growth conditions, which also indicates systematic co-regulation of the single complex components (Supplementary Table 27). It is important to note that albeit proteins are synthesized with specific stoichiometries[6], different protein degradation rates can alter cellular levels for some proteins and explain some of the mismatching stoichiometries observed.

Finally, a hierarchical cluster analysis of all measured protein abundances revealed actually expected biologically meaningful clusters (Supplementary Figure 9); for instance, the most significant induced proteins in glycerol medium are glpD, glpB, glpK and glpQ, (see also Supplementary Table 8) which are all key enzymes for glycerol metabolism. We made similar consistent observations for other conditions (e.g. xylose, fructose, stationary phase). Further, the 3 glucose growth condition samples were analyzed at the beginning and end of the quantitative LC-MS experiment to check if the data was impaired by any systematic performance changes of the LC-MS platform employed throughout the whole analysis time of the high number of >60 samples. The conjoined clustering of these 6 samples indicated that the data quality was high and consistent throughout the whole LC-MS experiment (Supplementary Figure 9).

Along these lines, the identification of consistent numbers of proteins across all samples (see Supplementary Table 23) indicates that sample preparation, LC-MS and data analysis were very robust throughout the whole LC-MS experiment. It also showed that most quantified proteins were expressed in all conditions and across the whole range of growth rate including stationary phase samples.

Supplementary Note 2:

Statistical data analysis

SafeQuant Description

The Progenesis analysis results were further processed using the SafeQuant R package v.2.1 (https://github.com/eahrne/SafeQuant/) to obtain protein relative abundances. This analysis included global data normalization by equalizing the total MS1 peak areas across all LC-MS runs, summation of MS1 peak areas per protein and LC-MS/MS run, followed by calculation of protein abundance ratios. The summarized protein expression values were used for statistical testing of between condition differentially abundant proteins. Here, empirical Bayes moderated t-Tests were applied, as implemented in the R/Bioconductor limma package[26]. The resulting per protein and condition comparison p-values were adjusted for multiple testing using the Benjamini-Hochberg method.

Underlying Statistical Assumptions

All LC-MS analysis runs are acquired from independent biological samples. To meet additional assumptions (normality and homoscedasticity) underlying the use of linear regression models and Student t-Test MS-intensity signals are transformed from the linear to the log-scale.

Linear Regression

Unless stated otherwise linear regression was performed using the ordinary least square (OLS) method as implemented in *base* package of R v.3.1.2 (R Core Team (2014). R: A language and

environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.)

Power

The sample size of three biological replicates was chosen assuming a within-group MS-signal Coefficient of Variation of 10%. When applying a two-sample, two-sided Student t-test this gives adequate power (80%) to detect protein abundance fold changes higher than 1.65, per statistical test. Note that the statistical package used to assess protein abundance changes, SafeQuant, employs a moderated t-Test, which has been shown to provide higher power than the Student t-test[27]. We did not do any simulations to assess power, upon correction for multiple testing (Benjamini-Hochberg correction), as a function of different effect sizes and assumed proportions of differentially abundant proteins.

Additional Clarifications

In Figure 2 and Supplemental Figures 11 and 16 we applied a robust linear regression, which is less sensitive to outliers than OLS regression. For this purpose, we used the algorithm implemented in the R package *MASS* (Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0,) selecting the *MM-estimation* method.

In Supplemental Figures 9 and 10 the hierarchical clustering of protein log2 abundance ratios was preformed using Ward's algorithm and the Correlation distance metric. Subsequently, a heatmap was created using the *gplots* R package (Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B: gplots: Various R programming tools for plotting data. 2009).

To assess the agreement between measured and estimated protein abundance values in Supplemental Figures 5 and 8, we calculate Lin's Concordance Correlation coefficient (R_c, Lin L (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics 45: 255 - 268.), as implemented in the R package *epiR*.

Supplementary Note 3:

Cell volume considerations

In this manuscript, we used condition-dependent cell volumes determined by Volkmer and Heinemann[28]. Using these volumes, the determined cell counts of the samples, and assuming that the cellular protein concentration is constant across the different growth conditions, we used the glucose condition as reference to compensate for any protein extraction biases for the other growth conditions. Cell grown on glucose were chosen as a reference, since we always obtained were consistent protein masses per cell of 280 fg / cell. We determined the masses independently by BCA assays and by mass spectrometry.

Note, that very recently we have re-determined – using superresolution microscopy – cell volumes for three growth conditions (Radzikoski *et al.*, 2015, submitted). A comparison between these volumes and the volumes determined in the PLoS ONE study showed that the PLoS ONE volumes were likely slightly overestimated: On fumarate: 1.11±0.58 fL (PLoS ONE 2.4±1.2 fL), on glucose 2.15±0.84 fL (PLoS ONE 3.2±1.2 fL), cells starved for 8h 0.92±0.40 (PLoS ONE 1.5±1.2 fL starved for 24 hours). If similar factors (i.e. 0.46; 0.67; 0.61) between the likely more correct volumes (as determined by superresolution microscopy) and the here used volumes (as determined in the PLoS

ONE study) are the case for the other growth conditions, then this would mean that protein concentrations would slight chance. Note that for consistency reasons, for this work here, we have solely used the PLoS ONE volumes. We suggest that researchers who need most accurate cellular protein concentrations apply a respective correction factor, i.e. determined on the basis of the volume measurements mentioned above.

References:

1. Krogh, A., Larsson, B., Heijne, Von, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305,** 567–580 (2001).
2. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4,** 41 (2003).
3. Glatter, T. *et al.* Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. *J. Proteome Res.* **11,** 5145–5156 (2012).
4. Ahrné, E., Nikitin, F., Lisacek, F. & Müller, M. QuickMod: A tool for open modification spectrum library searches. *J. Proteome Res.* **10,** 2913–2921 (2011).
5. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
6. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157,** 624–635 (2014).
7. Masuda, T., Saito, N., Tomita, M. & Ishihama, Y. Unbiased quantitation of Escherichia coli membrane proteome using phase transfer surfactants. *Mol Cell Proteomics* **8,** 2770–2777 (2009).
8. Brun, V., Masselon, C., Garin, J. & Dupuis, A. Isotope dilution strategies for absolute quantitative proteomics. *J Proteomics* **72,** 740–749 (2009).
9. Pedersen, S., Bloch, P. L., Reeh, S. & Neidhardt, F. C. Patterns of protein synthesis in E. coli: a catalog of the amount of 140 individual proteins at different growth rates. *Cell* **14,** 179–190 (1978).
10. Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329,** 533–538 (2010).
11. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25,** 117–124 (2006).
12. Ishii, N. *et al.* Multiple High-Throughput Analyses Monitor the Response of E. coli to Perturbations. *Science* **316,** 593–597 (2007).
13. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **4,** 1265–1272 (2005).
14. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* **8,** R183 (2007).
15. Matias, V. R. F., Al-Amoudi, A., Dubochet, J. & Beveridge, T. J. Cryo-transmission electron microscopy of frozen-hydrated sections of Escherichia coli and Pseudomonas aeruginosa. *J Bacteriol* **185,** 6112–6118 (2003).
16. Xing, G., Zhang, J., Chen, Y. & Zhao, Y. Identification of four novel types of in vitro protein modifications. *J. Proteome Res.* **7,** 4603–4608 (2008).
17. Kollipara, L. & Zahedi, R. P. Protein carbamylation: in vivo modification or in vitro artefact? *PROTEOMICS* **13,** 941–944 (2013).
18. Corbin, R. W. *et al.* Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc Natl Acad Sci USA* **100,** 9232–9237 (2003).
19. Wu, C. C. & Yates, J. R. The application of mass spectrometry to membrane proteomics. *Nat Biotechnol* **21,** 262–267 (2003).
20. Ishihama, Y. *et al.* Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* **9,** 102 (2008).
21. Arike, L. *et al.* Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. *J Proteomics* **75,** 5437–5448 (2012).
22. Ahrné, E., Molzahn, L., Glatter, T. & Schmidt, A. Critical assessment of proteome-wide label-

free absolute abundance estimation strategies. *PROTEOMICS* **13,** 2567–2578 (2013).
23. Grossmann, J. *et al.* Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *J Proteomics* **73,** 1740–1746 (2010).
24. Nilsson, M., Bülow, L. & Wahlund, K. G. Use of flow field-flow fractionation for the rapid quantitation of ribosome and ribosomal subunits in Escherichia coli at different protein production conditions. *Biotechnol. Bioeng.* **54,** 461–467 (1997).
25. Dennis, P. P. & Bremer, H. Macromolecular composition during steady-state growth of Escherichia coli B-r. *J Bacteriol* **119,** 270–281 (1974).
26. Rossini, A. J., Sawitzki, G., Smith, C. & Smyth, G. Bioconductor: open software development for computational biology and bioinformatics. *Genome …* (2004).
27. Yang, D., Parrish, R. S. & Brock, G. N. Empirical evaluation of consistency and accuracy of methods to detect differentially expressed genes based on microarray data. *Comput. Biol. Med.* **46,** 1–10 (2014).
28. Volkmer, B. & Heinemann, M. Condition-dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. *PLoS ONE* **6,** e23126 (2011).