A systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data: Supporting Information

Laurence Yang[a,*], Justin Tan[a,*], Edward J. O'Brien[a], Jonathan Monk[a], Donghyuk Kim[a], Howard J. Li[a], Pep Charusanti[a], Ali Ebrahim[a], Colton J. Lloyd[a], James T. Yurkovich[a], Bin Du[a], Andreas Dräger[a,b], Alex Thomas[a,c], Yuekai Sun[d], Michael A. Saunders[e], Bernhard O. Palsson[a,c,†]

*These authors contributed equally [†]Corresponding author

[a]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, [b]Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, 72076 Tübingen, Germany, [c]Novo Nordisk Foundation Center for Biosustainability, 2970 Hørsholm, Denmark, [d]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, [e]Department of Management Science and Engineering, Stanford University, Stanford, CA 94305

Corresponding author: Bernhard O. Palsson, palsson@ucsd.edu

SI Figures

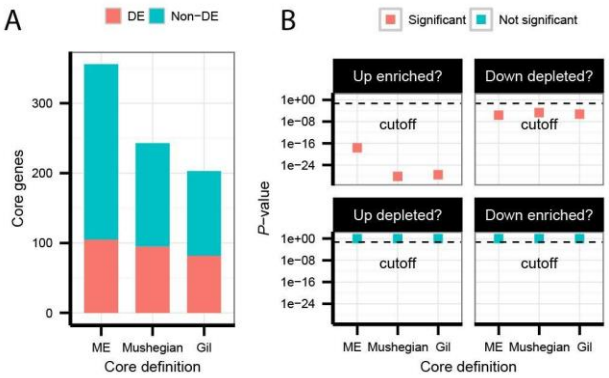


Figure S1 Enrichment and depletion analysis of differentially expressed (DE) genes in eight adaptively evolved strains (1). (A) Proportion of core genes that are differentially and non-differentially expressed. Genes were considered differentially expressed if they were commonly up- or down-regulated in six or more strains, as in (1). (B) A *p*-value cutoff of 0.05 (below the horizontal lines) was used to determine significant depletion.

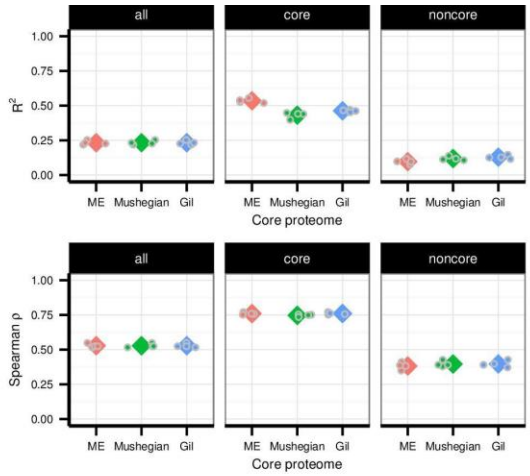


Figure S2 Correlation coefficients of proteomics data between MG1655 (2) and four strains of BW25113 (3). Correlation coefficients were calculated using *Z* scores of the $\log_2$-transformed proteomics data sets. Diamonds correspond to the mean coefficient across strains, while circles correspond to the correlation coefficient for each strain.

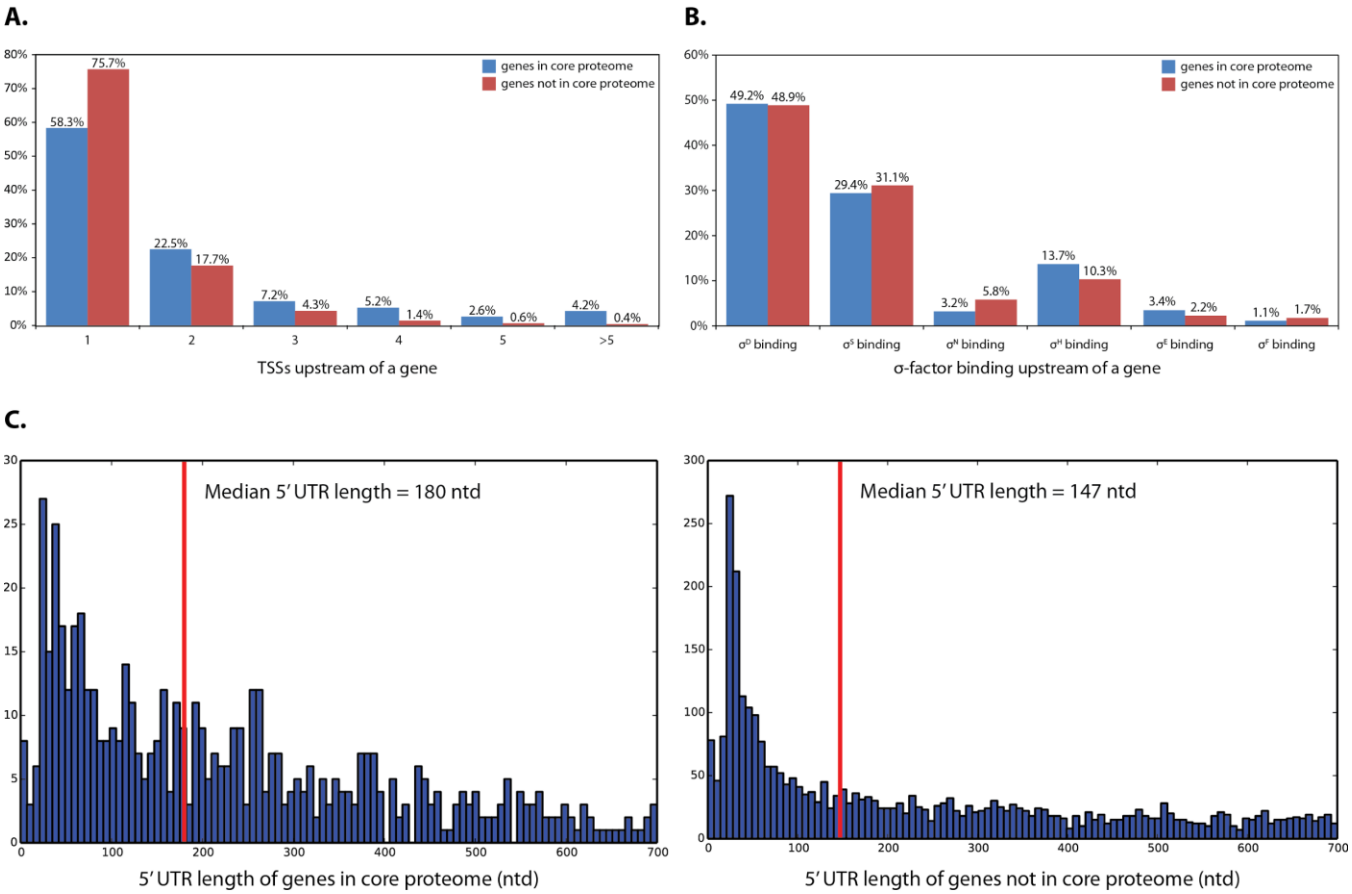


Figure S3 Transcriptional and post-transcriptional regulation of core vs. non-core genes of *E. coli*. (A) Transcription start sites (TSSs) per gene in core (blue) and non-core (red) proteomes. (B) Sigma factor binding patterns of core (blue) and non-core (red) proteomes. (C) 5' UTR length of genes in the core (left) and non-core (right) proteomes. We used published TSS (4) and sigma factor binding (5) datasets for the analyses.

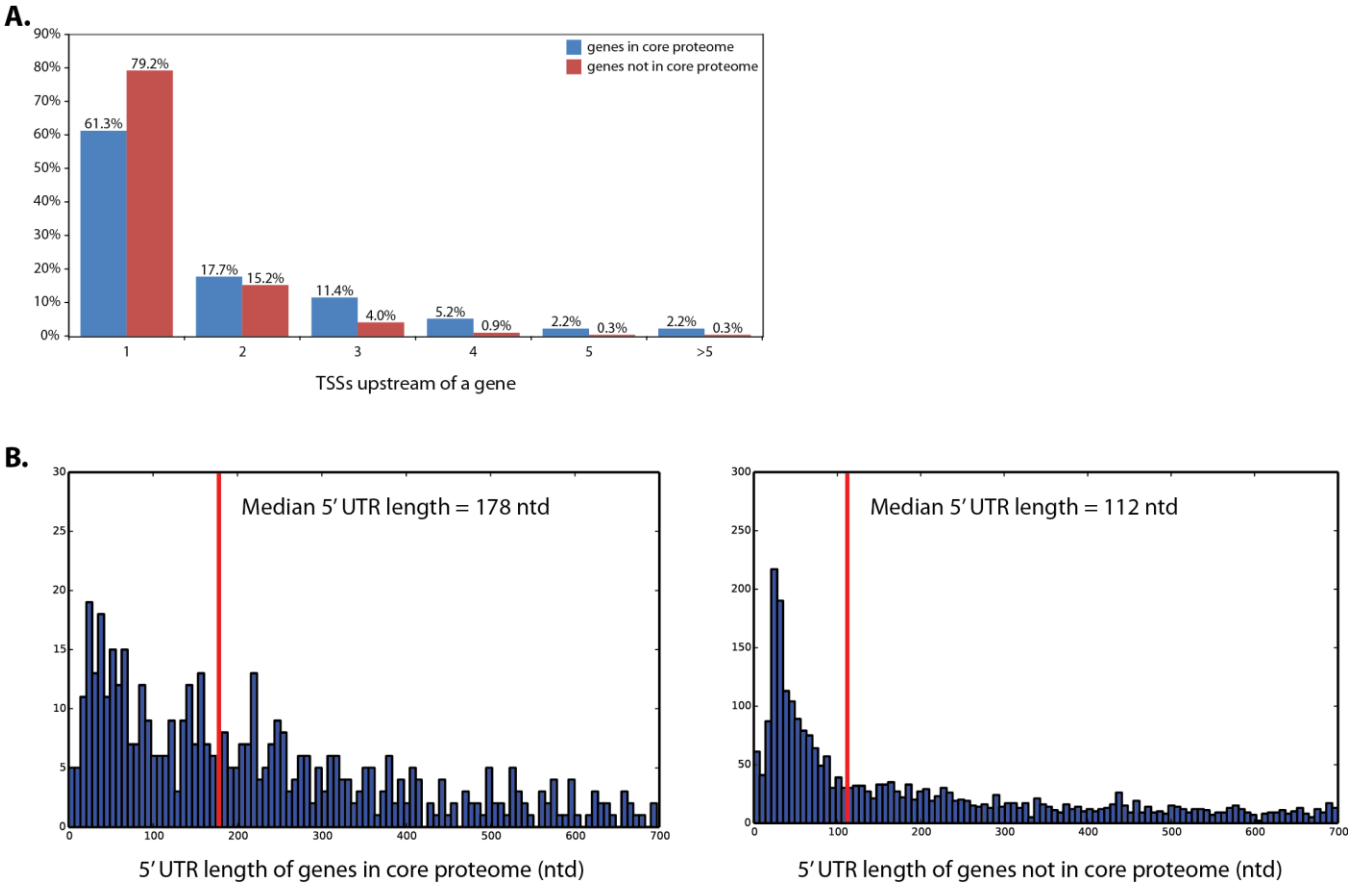


Figure S4 Transcriptional and post-transcriptional regulation of core vs. non-core genes of *K. pneumoniae*. (A) Transcription start sites (TSSs) per gene in core (blue) and non-core (red) proteomes. (B) 5' UTR length of genes in the core (left) and non-core (right) proteomes. We used published TSS datasets (4) for this analysis.

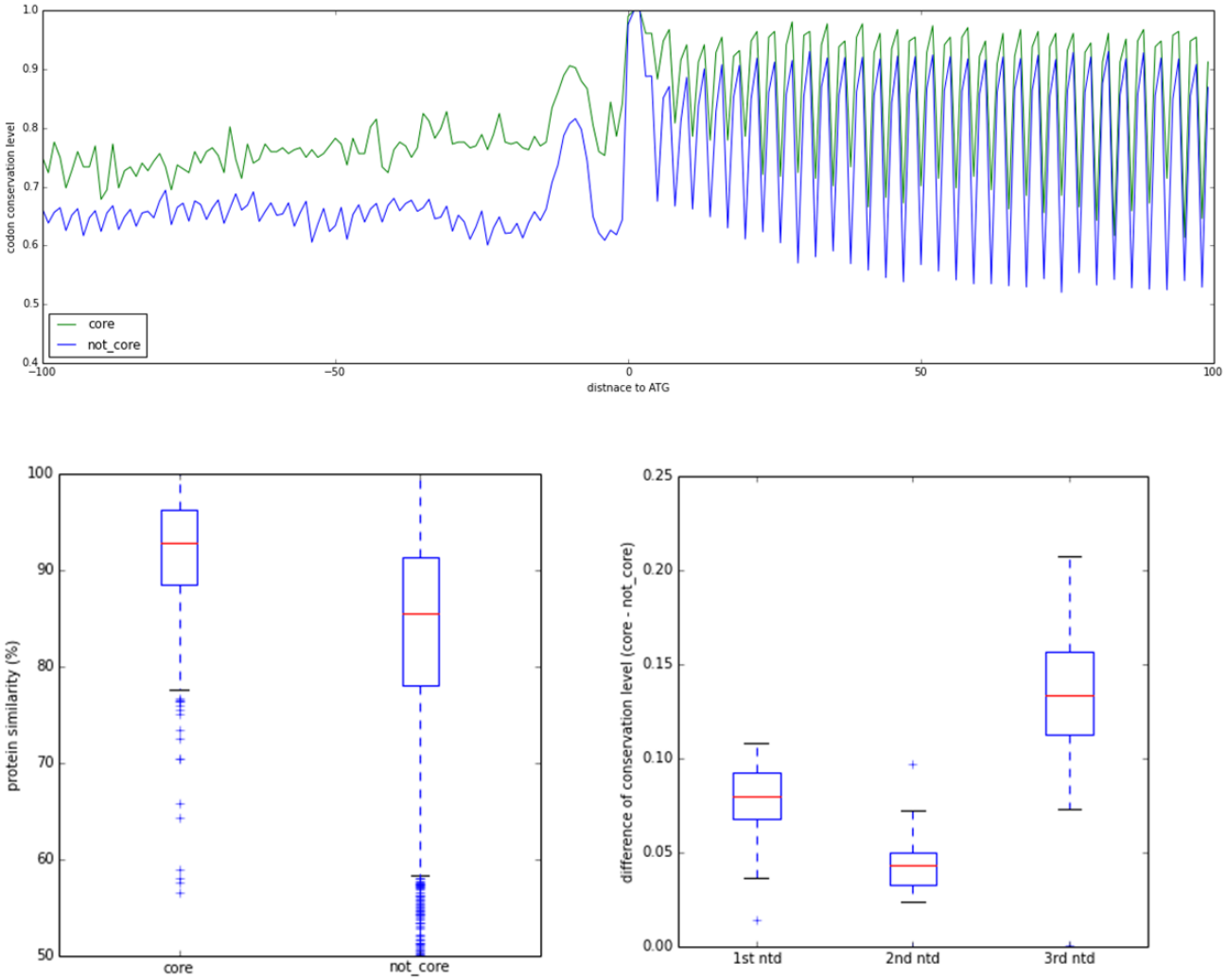


Figure S5 Sequence conservation of genomic regions surrounding translation start sites between *E. coli* and *K. pneumoniae*. Overall, genes in core proteome have higher level of sequence conservation than genes not in core proteome in all nucleotides for which nucleotide conservation level was measured. This difference in conservation level is implicated in the level of protein similarity of genes in the core proteome and genes not in the core proteome. When conservation level of 1st, 2nd, and 3rd nucleotides in each codon of coding region was compared, the 2nd nucleotide showed the least difference between genes in core proteome and genes not in core proteome, and the 3rd nucleotide showed the largest difference in comparison between two groups of genes. Published TSS datasets (4) were used for the analysis.

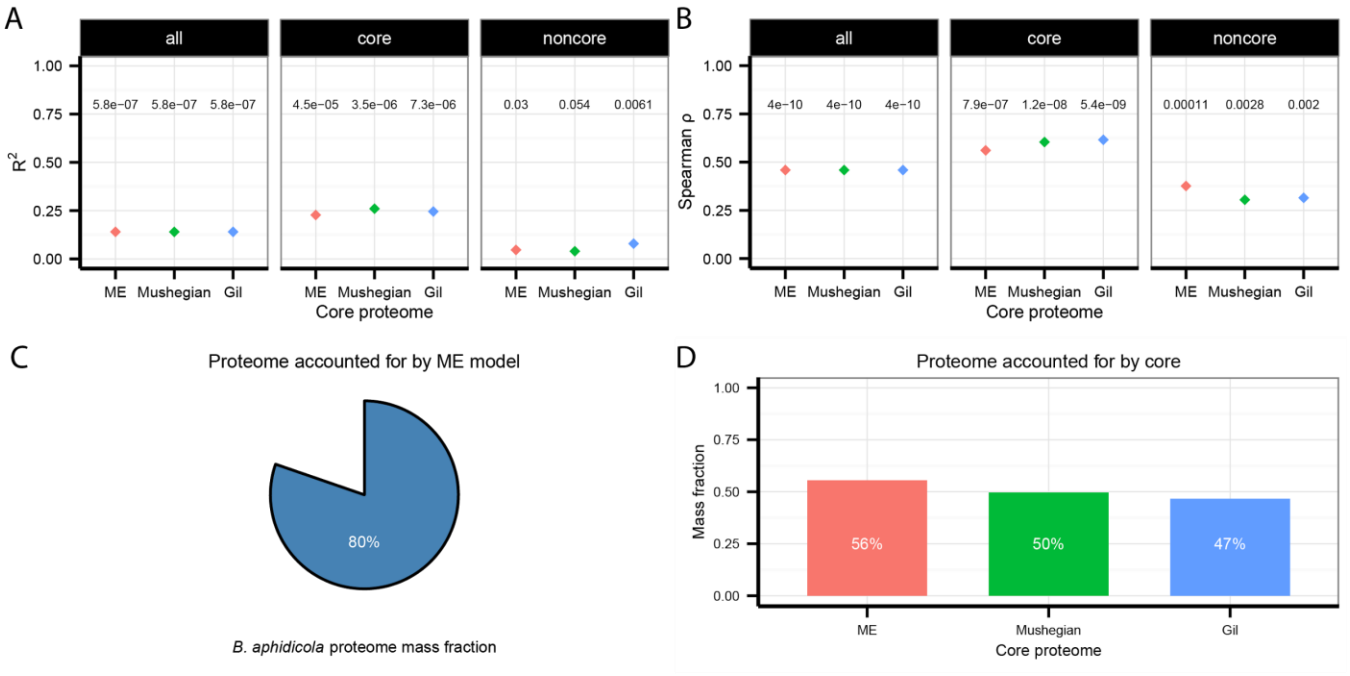


Figure S6 Characteristics of the orthologs of the *E. coli* core proteome in *B. aphidicola*. (A) Pearson coefficient of determination for core and non-core genes. (B) Spearman rank correlation for core and non-core genes. (C) Based on proteomics data (6), the ME model accounts for 80% of the orthologous genes in *B. aphidicola*. (D) Based on the same proteomics dataset, the model-based core proteome and two minimal genomes account for 56%, 50%, and 47% of the *B. aphidicola* orthologous proteome, respectively. Core genes showed higher Pearson and Spearman rank correlations than non-core genes, although these correlations were not statistically significant based on Fisher's *Z*-transform and Zou's confidence interval methods (7) and a permutation test.

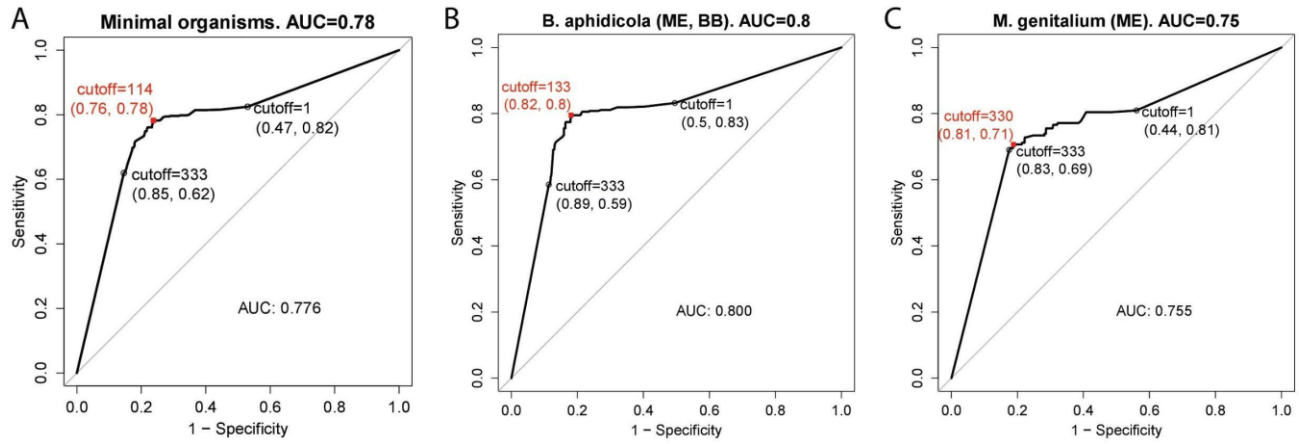


Figure S7 Overlap of the core proteome genes with two minimal organisms. (A) ROC plot for both *B. aphidicola* and *M. genitalium*. (B) ROC curve for *B. aphidicola* only. (C) ROC curve for *M. genitalium* only. The 'cutoff' was chosen to maximize the sum of sensitivity and specificity.

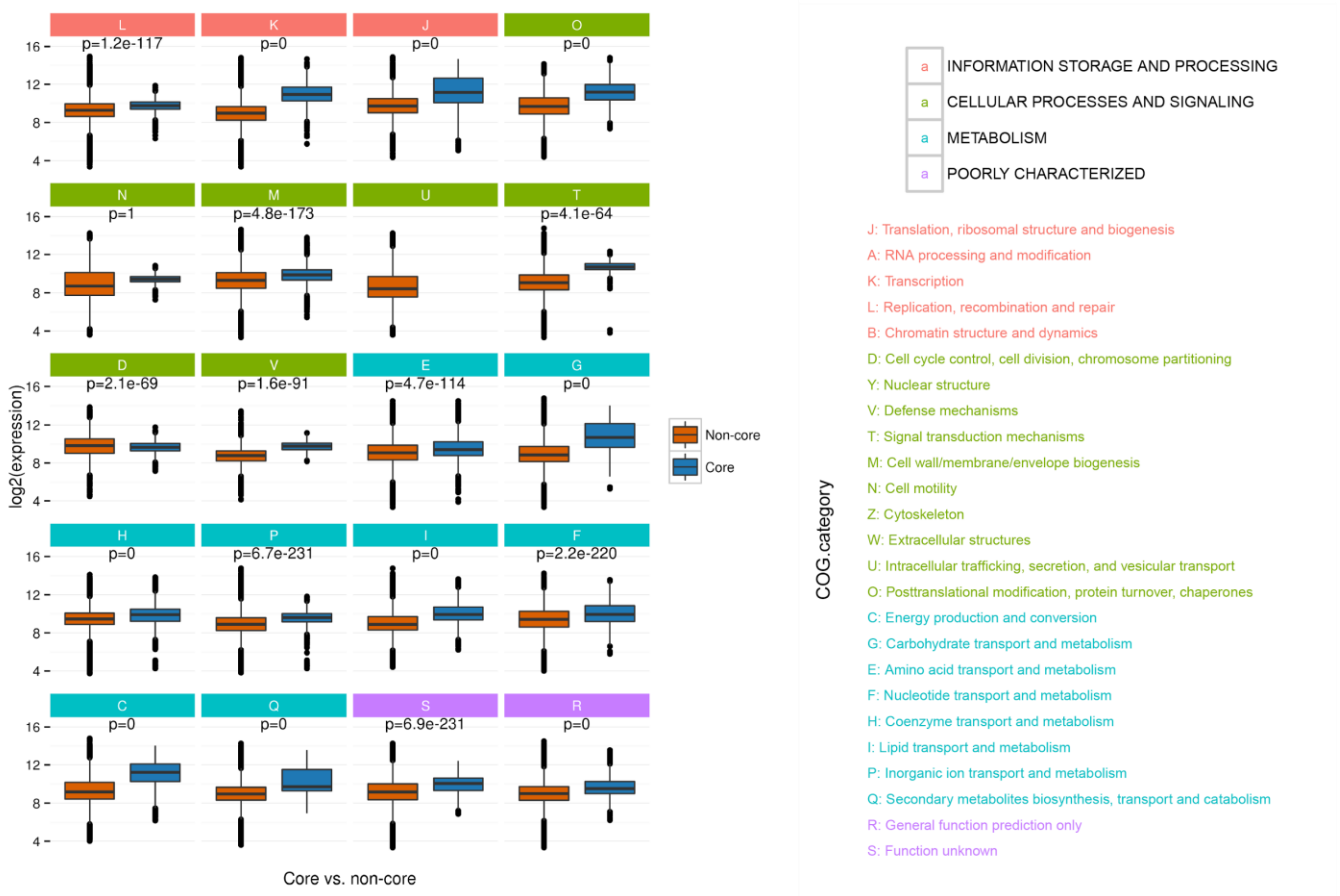


Figure S8 Expression of core and non-core genes, grouped by COG across 444 transcriptomics profiles (see Methods). The *p*-values are calculated from one-sided Wilcoxon rank-sum tests, with the alternate hypothesis that core gene have higher expression

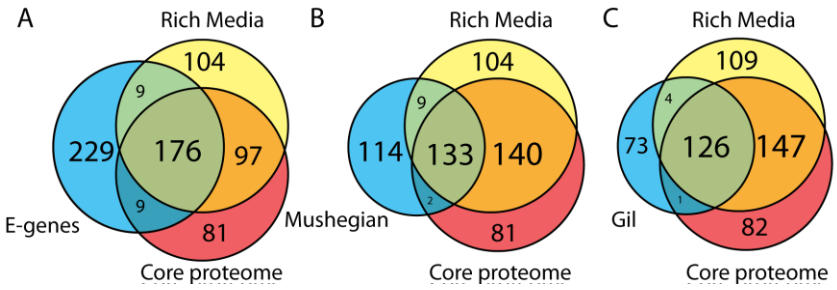
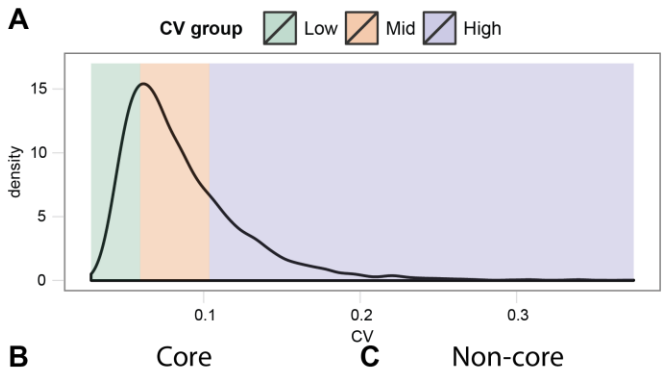
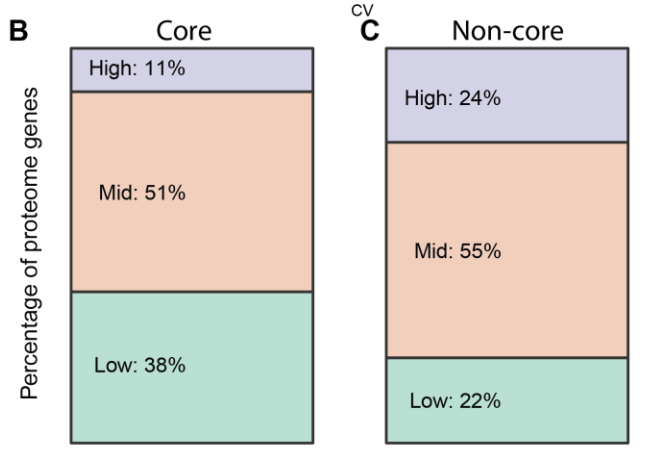


Figure S9 Comparisons of rich media ME model simulations with the computationally derived core proteome. A: The rich media simulation shows a large overlap with the core proteome, and around 50% of both the core and rich media simulations are composed of E (expression) genes(8). Most of the discrepancy between the core proteome and rich media genes are due to differences in M (metabolic) genes. This indicates that the expression machinery is consistent even under rich media environments where the cell can uptake all necessary metabolites. B and C: Comparison between rich media simulation, core proteome, and the paleomes defined in Gil *et al.* and Mushegian *et al.* The rich media simulations only marginally improve overlap with either of the paleomes. This indicates that many of the paleome genes are already accounted for by the core proteome, while those unaccounted for by the core proteome are currently outside the scope of the ME model (e.g. replication)





D

CV group	*P* (enriched)	*P* (depleted)
Low	**4.7e−09**	1.0e+00
Mid	4.1e−01	6.3e−01
High	1.0e+00	**5.1e−12**

E

CV group	*P* (enriched)	*P* (depleted)
Low	1.00000	**0.00078**
Mid	**0.00013**	1.00000
High	0.87000	0.15000

Figure S10 Enrichment and depletion of core genes in low-, mid-, and high-variation gene sets identified in the EcoMAC microarray compendium. (A) Distribution of the coefficient of variation (CV) of genes across arrays in EcoMAC. (B-C) The fraction of core genes that are in the low (<25th percentile), mid (between 25th and 75th percentiles) and high (above 75th percentile) variation gene sets for core and non-core genes. (D-E) Enrichment and depletion *P*-values of core proteome genes in low-, mid-, and high-variation gene sets. Significant enrichment or depletion (*P*-value < 0.01) is indicated in bold. Low, mid, and high variation sets were defined as genes with CV below the 25th, between 25th and 75th, and above the 75th percentiles, respectively, similar to Mar et al. (2011). Core genes showed higher fraction of low-CV, and lower fraction of high-CV genes than non-core; both results were statistically significant (permutation test, *P*-value < 0.0001).

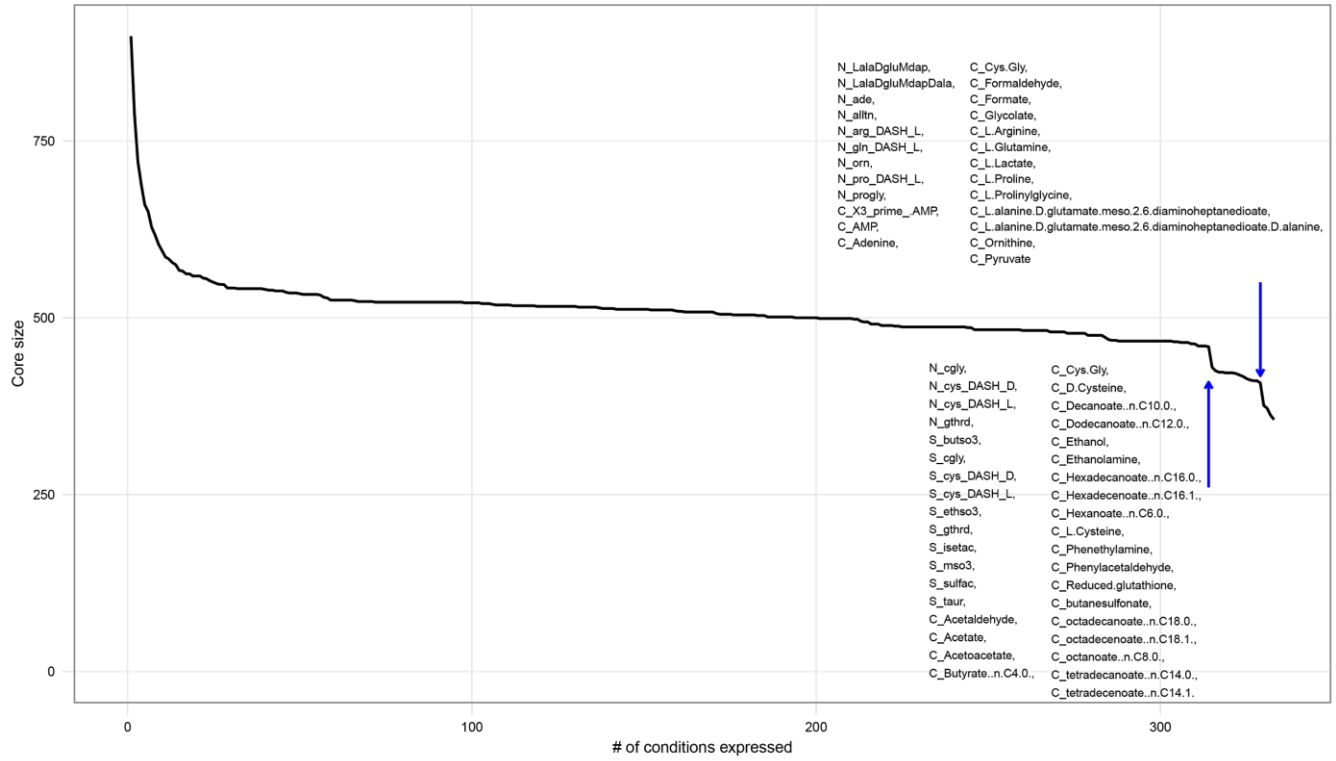


Figure S11 Variation in core proteome size (i.e., number of genes in the core set) with the number of conditions required for a gene to be expressed in ME simulations. The arrows point to notable reductions in core proteome size where # of conditions expressed are 314 and 329. At the first arrow (x=314), 29 genes are dropped from the core because they are expressed in 313 conditions but not in 314. Accordingly, the 37 conditions in which these dropped genes are not expressed are listed in the figure. Similarly, at x=329, 32 genes are dropped--the 25 conditions in which these genes are not expressed are listed in the figure. Prefixes C_, N_, P_, S_ indicate that the listed nutrient is an alternate carbon, nitrogen, phosphorous, or sulfur source, respectively.

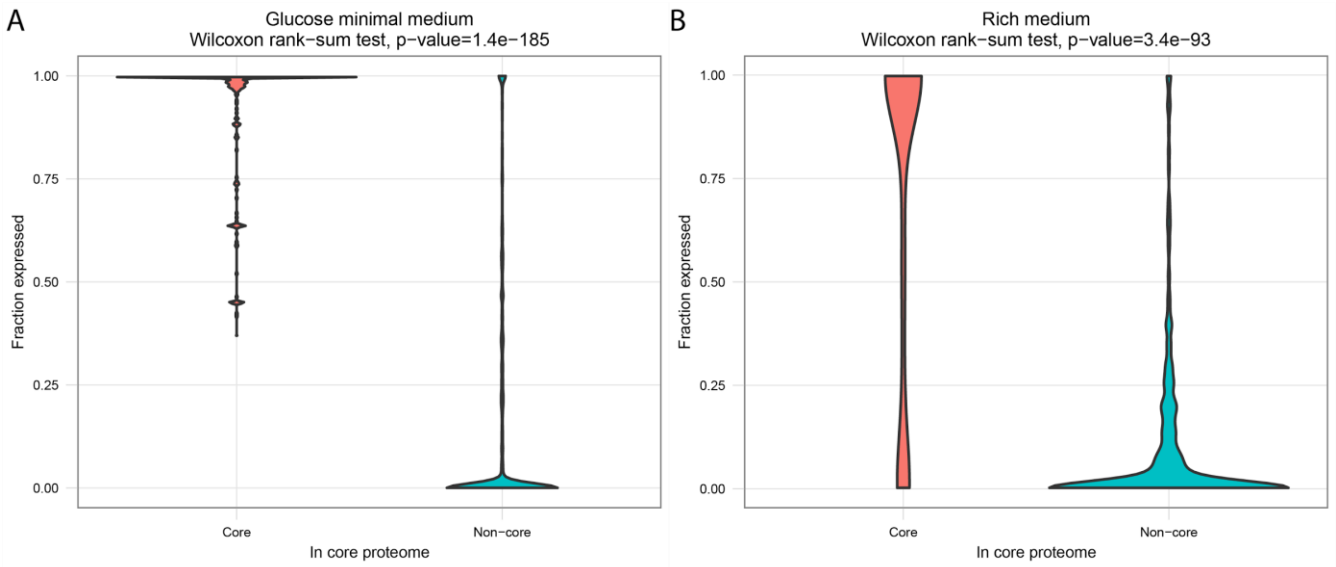


Figure S12 Fraction of simulations in which genes were expressed, for 300 random perturbations. Each randomly perturbed simulation corresponds to a ME simulation where effective rate constants were randomly perturbed. under (A) glucose minimal medium and (B) rich medium conditions. For glucose minimal media, the median percentage of samples where a core gene was expressed 100%, and the mean was 95%; non-core genes had median 0% and mean of 13%. For rich media, the percent expressed for core genes was 100% (median) and 72% (mean); non-core genes were expressed 0.67% (median) and 10% (mean).

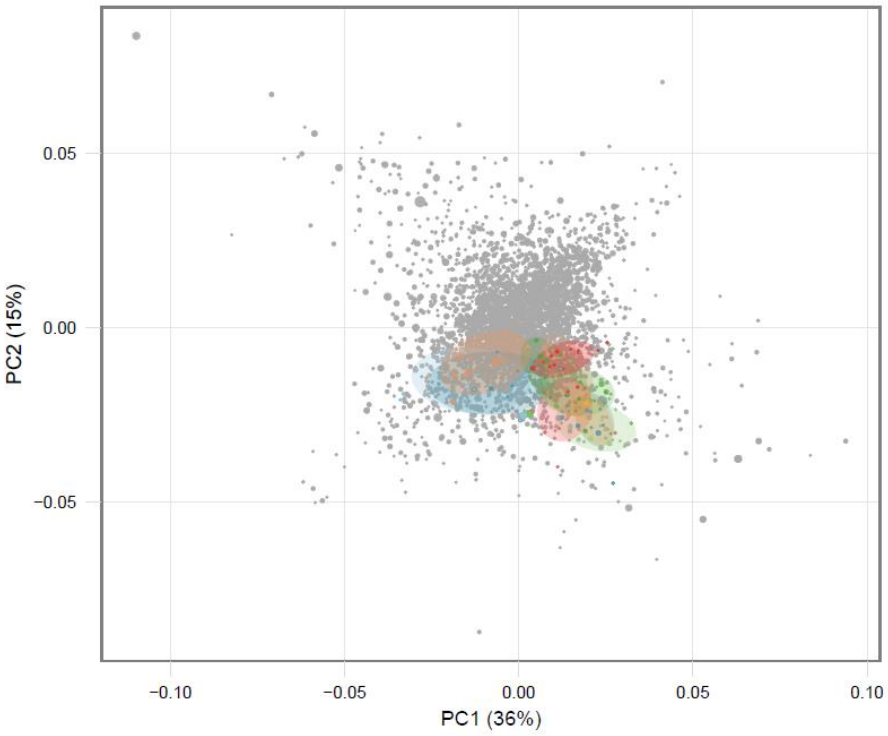


Figure S13 PCA plot of the EcoMAC compendium, transformed to log2-fold-change, relative to three reference samples (glucose M9 aerobic-grown MG1655). Colors correspond to biclusters that were significantly enriched for both KEGG pathways and the core proteome. Grey colored dots indicate genes that were not members of these doubly-enriched biclusters.

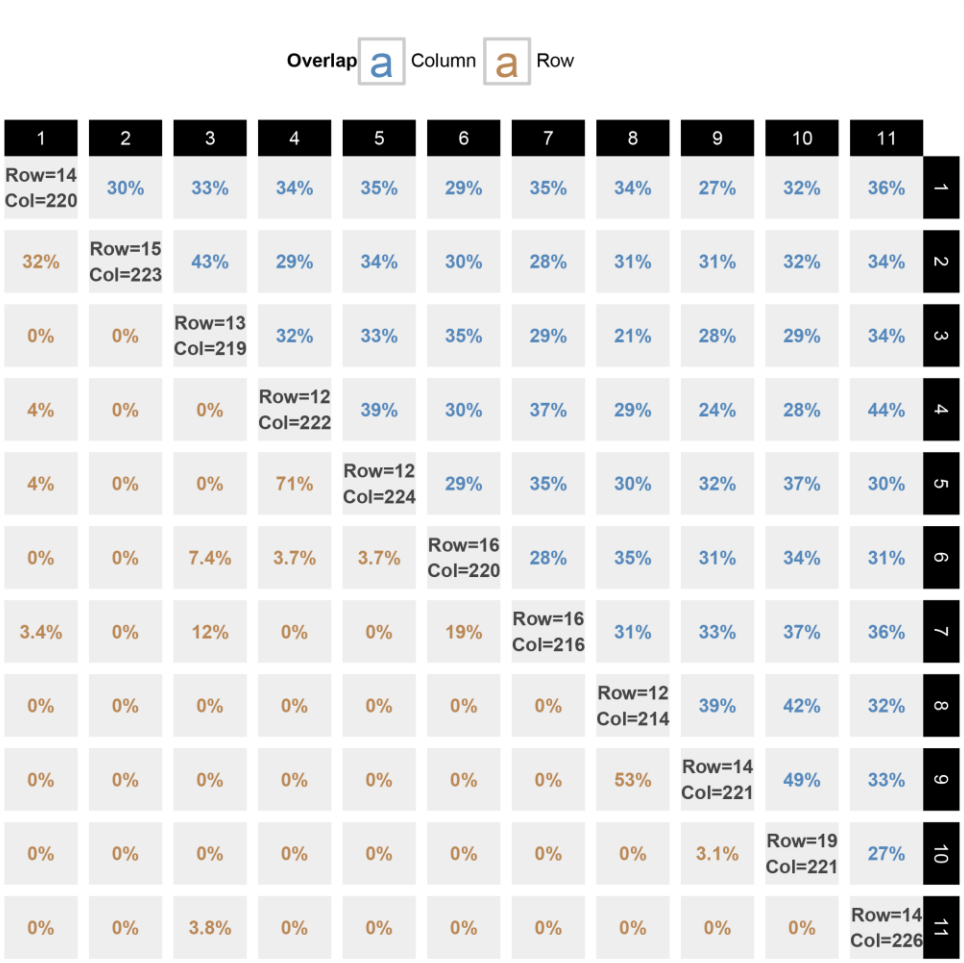


Figure S14 Size and overlap of biclusters that were significantly enriched for both KEGG pathways and the core proteome. Diagonals show the number of genes (rows) and conditions (columns) of a bicluster. Lower triangular values show row overlap, while upper triangular values show column overlap. Biclusters numbers are consistent with those in Figure 3 in the main manuscript.

SI Tables

Dataset S1 List of genes in the *E. coli* core proteomes (ME model, Gil, Mushegian)

(Dataset_S1_core_rich_genes_supp_1.xlsx)

Dataset S2 Simulated translation fluxes using the ME model under 333 minimal media, rich medium, glucose minimal with 300 randomly sampled effective rate constants, and rich medium with 300 randomly sampled rate constants

(Dataset_S2_sims_and_sens_analysis.xlsx)

Dataset S3 Core proteome enrichment and depletion analysis of commonly up/down-regulated genes in adaptively evolved strains.

(Dataset_S3_ALE_enrichment.csv)

Dataset S4 Clustering results for 333 simulated growth conditions

(Dataset_S4_clsutering.csv)

Dataset S5 Tn-seq gene essentiality results

(Dataset_S5_tnseq.csv)

Dataset S6 Core proteome genes that were not expressed in the Lewis et al. (2009) microarray compendium. The rich media gene set included an additional 10 non-expressed genes (Dataset S6): 9 transport activities and 1 phosphate acetyltransferase activity (eutD).

(Dataset_S6_SI_Table_core_off_in_ecoli2_clean.csv)

Dataset S7 Biclusters identified using cMonkey

(Dataset_S7_cmonkey_no_motifs1.xlsx)

Dataset S8 DAVID functional annotation clustering and enrichment results

(Dataset_S8_DAVID_clustering_supp_8.xlsx)

SI Methods

RNA-seq data preparation For RNA-seq measurements in three carbon sources (glucose, fructose, and acetate), HTSeq (9) was used to count reads, while DESeq2 (10) was used to obtain normalized counts and to identify differentially expressed genes. Wald significance tests were used in DESeq2 to test for differential expression. The *p*-values across multiple conditions were combined using Stouffer's method (11). We finally applied Benjamini-Hochberg correction for multiple testing (12), resulting in a single, combined and adjusted *p*-value per gene. In preparing RNA-Seq data for hierarchical clustering, we converted RNA-Seq data into *Z* scores after $\log_2$-transformation.

Correlation between transcriptomics or proteomics data Spearman rank and Pearson correlation coefficients were calculated between pairs of transcriptomics or pairs of proteomics data sets using the R function `cor.test() (13)`. Genes not expressed in either experiment were excluded to avoid biasing the correlation toward unexpressed genes. Statistical tests of difference between Pearson correlation coefficients was performed using Fisher's *Z*-transform method and Zou's confidence interval method (14). Both methods were performed using the cocor R package (7). If both methods were consistent, only the *Z*-transform *p*-value was reported in the text. Statistical test of difference between Spearman rank correlations was performed using a permutation test with 10,000 permutations.

Multivariate analysis and clustering of simulated and measured expression profiles Affinity propagation (15) was used to cluster simulated translation flux profiles using the R package APCluster (16). The default `q = 0.5` resulted in 18 clusters across 333 nutrients.

Enrichment and depletion analysis Enrichment and depletion analysis was performed using hypergeometric *p*-values, with the Benjamini-Hochberg correction for multiple testing (12).

Identification of expressed and non-expressed genes from microarrays From the normalized microarray compendium from (17), we removed two conditions that did not include a replicate, yielding a total of 69 conditions. This compendium was used to identify expressed and non-expressed genes across the conditions. A hypergeometric test of replicate samples of each probe against 21 Affymetrix control probes was conducted, as described in (17). Thus, 1987 genes were found to always be expressed across the 69 conditions (hypergeometric test, *p*-value < 0.05 compared against 21 Affymetrix control probes). The core proteome

included 305 (86% of core proteome) always-expressed genes, which was a significant enrichment (hypergeometric test, *p*-value = 8.0×10^{-28}). The core proteome also included 5 genes that were never expressed (Dataset S6). Two (ompN and IptG) coded transport activities, one (cynT) was an isozyme for carbonic anhydrase activity, and two coded ribosomal RNA methylation or processing (rlmC and rluC). These few inconsistencies are explained by the known tendency of ME models to choose the most efficient enzymes due to a total proteome mass constraint (18). Improved determination of effective rate constants for enzymes is expected to improve these inconsistencies in the future.

Comparison of *B. subtilis* and the *E. coli* core proteome *B. subtilis* growth simulations were run using the iYO844 genome-scale model (19)for 95 carbon (viable in 81), 52 nitrogen, 26 phosphorous, and 12 sulfur sources. Of the *B. subtilis* genes in the iYO844 genome-scale model, we identified 225 functional homologs with *E. coli* using the SEED (*E*-value < 1e-5). Homologs that were in the core proteome, as defined using the *E. coli* ME model, showed significantly higher predicted frequency of utilization across 171 simulations (Wilcoxon rank-sum test, *p*-value = 5.2×10^{-7}). The median percentage of *B. subtilis* simulation conditions where core genes were predicted to be expressed was 84%. Finally, we identified 20 homologous core genes that were predicted to not be expressed in *B. subtilis*. Of these, 8 genes are essential in *B. subtilis (20)*, while 11 are essential in either *B. subtilis* or *E. coli (21)*. The remaining 9 consisted of cytochrome oxidase, leucine, histidine, and glycogen biosynthesis genes.
We also compared the *B. subtilis* model with Gil and Mushegian gene sets. First, we found that the limited scope of the *B. subtilis* metabolic model resulted in a small overlap with these gene sets. Specifically, Gil *et al.* identified 178 *B. subtilis* homologous genes in their paleome, of which only 16 were present in the iYO844. In part, this is largely due to the fact that many of the core genes (200/356), as well as paleome genes from Gil *et al.* and Mushegian *et al.* have functions related to the transcription and translation machinery, and thus are found in the E (expression) portion of the ME model.

ROC curves for genome overlap with *B. aphidicola* and *M. genitalium* ROC curves were constructed in R using the pROC package (22). The 564 *E. coli* orthologs of *B. aphidicola* APS were obtained from BuchneraBase (23).

ME simulation across 333 growth conditions Simulations were carried out using iOL1554-ME, the genome-scale model of *E. coli* K-12 MG1655 metabolism and expression (24). A total of 333 ME-model simulations were performed for a base media of glucose/M9-minimal media. In each simulation, the main carbon, nitrogen, phosphate or sulfate source in the medium was changed, with the other three nutrient sources held constant. In total, 180 different carbon sources, 49 phosphorus sources, 93 nitrogen sources, and 11 sulfur sources were tested.

Sensitivity analysis of core proteome and rich media ME simulations We tested the robustness of the core proteome, and rich media-based gene set, to ME model parameter uncertainties (Dataset S2). To this end, ME simulations were run with 300 randomly sampled effective rate constants (SBO:0000611). For the core proteome, simulations were run on glucose minimal medium; and rich medium for the rich medium gene set. For glucose minimal media, the median percentage of samples where a core gene was expressed 100%, and the mean was 95%; non-core genes had median 0% and mean of 13% (Fig. S12A). For rich media, the percent expressed for core genes was 100% (median) and 72% (mean); non-core genes were expressed 0.67% (median) and 10% (mean) (Fig. S12B).

Furthermore, out of 300 random samples in glucose minimal medium, cytochrome bo oxidase, bd-I oxidase, and bd-II oxidase were used 45%, 23%, and 56% of the time (Dataset S7). Therefore, although with fixed parameters, cytochrome bo oxidase was selected 100% of the time across 333 conditions, all three cytochrome oxidases might be considered as a set, when parameter uncertainty is considered.

Comparison of *E. coli*, *M. genitalium, and B. subtilis* functional annotations The 564 *E. coli* orthologs of *B. aphidicola* APS were obtained from BuchneraBase (23). Functional overlaps between *E. coli* and *M. genitalium* were compared using RAST annotated FIGfams (25–27). RAST (25, 26) annotations of *E. coli* K-12 MG1655 [RASTID: 511145.6] and *M. genitalium* [RASTID: 243273.1] were obtained using the SEED servers and RAST API (27). Functional overlaps were compared using RAST annotated FIGfams. FIGfams are assigned based on grouping of proteins that can reliably be asserted to implement identical function. Therefore FIGfams serve as a consistent way to compare overlapping functions between different organisms, even those as distantly related as *E. coli* and *M. genitalium*. A similar procedure was followed to obtain overlapping functions between *E. coli* and *B. subtilis* str. 168 [RASTID: 224308.1].

Selection and preparation of microarray compendium The EcoMAC microarray compendium (28) was used for clustering and enrichment analysis. We included only a relevant subset of all conditions. Thus, we excluded regulatory rewiring samples, as they would not represent the naturally-evolved expression patterns; we removed microgravity and magnetic treatment conditions; we kept only strains labeled as K12, MG1655, BW25113, and W3110; we removed time-dependent samples (i.e., kept arrays with Time labeled *blank*, WT, exponential, mid log phase, and mid-log phase. Finally, we had 444 relevant conditions.

Biclustering data preparation and analysis Biclustering was performed using cMonkey (29). The data was provided as log2 fold-change of each expression profile, relative to a set of reference profiles. The reference profiles were chosen to be three MG1655 samples, all in aerobic, glucose M9 medium. cMonkey uses a Markov chain modeling process to determine the probability of each gene and column being in a specific bicluster based on a predetermined scoring algorithm. By default, the scoring algorithm weights for the presence of cis-regulatory motifs and network association which might bias the biclustering results. To account for this, four separate runs were performed with two different scoring algorithm parameters (i.e., default parameters, and zero motif and network scoring), each in replicate, to eliminate biclusters sensitive to algorithm parameters. Each run was then tested for enrichment for KEGG pathways, as well as for the core proteome. Bicluster sizes ranged between 8 to 39 genes (median=19) by 210 to 235 conditions (median=219) (SI Dataset S). Upon filtering out biclusters not enriched for either gene set, we identified six KEGG pathways that were significantly enriched in all runs of cMonkey. These were Ribosome; Valine, leucine and isoleucine biosynthesis; C5-Branched dibasic acid metabolism; Thiamine metabolism; Histidine metabolism; and Phenylalanine, tyrosine and tryptophan metabolism.

Determining gene essentiality from Tn-seq Genes were considered essential if the FDR-adjusted *p*-value was < 0.05 and the $\log_2$ ratio of (normalized) measured-over-expected number of reads per gene was below the optimal cutoff as predicted by ESSENTIALS (30).

Tn-seq

Bacterial strains

Tn-seq experiments were performed using the ΔhsdR strain of E. coli K-12 BW25113 obtained from the KEIO collection (31). Note that BW25113 has a well-defined pedigree and is closely related to MG1655 (21). Random

transposon mutagenesis yields in strain BW25113 were higher than that of wild-type E. coli K-12 MG1655, partly due to the inactivation of hsdR. Because hsdR is a non-essential, non-metabolic, and non-regulatory gene, deleting the gene is expected to have minimal effect on metabolism and expression machinery. Note also that the parental BW25113 strain already has a frameshift mutation, inactivating hsdR (32); therefore, the full deletion of the gene will likewise have little effect.

Transposon library construction

E. coli was grown overnight in LB media (3 mL). For each transposome reaction, 20 LB Kan plates were prepared. For each transposome reaction, three 50 mL aliquots of sterile 10% glycerol were prepared in 50 mL Falcon tubes and stored at 4 °C. Cells were harvested at an OD600 of 0.6, and electroporated (Eppendorf Eporator at 2500 V; Bio-Rad electroporation cuvettes, 0.1 cm gap). Immediately after electroporation, the cells were incubated in 800 mL SOC medium at 37 °C. Cells were then plated onto LB Kan plates and incubated at 37 °C overnight. Colonies ranged between 60,000 - 90,000 colonies per library. The pooled libraries were scraped of the LB plates and washed 3 times with 50 mL PBS (phosphate buffered saline) to reduce nutrient carryover from the LB plates. They were then further incubated with shaking at 37 °C for 12 hours in M9 glucose media with Kan. The pooled libraries were then passaged three times, centrifuged and frozen.

Subsequently, TruSeq (Illumina) DNA libraries were prepared by first fragmenting DNA using 350 bp insert size Covaris settings (5% DF, 175 W PIP, 200 cycles/burst, 50 seconds), followed by clean-up. End repair and size selection was then performed using settings for 350 bp insert size. The 3' ends were then adenlyated, followed by ligating of the adaptors to the transposons.

Tn-specific PCR (for primers, see **Sequences of primers and library intermediates**) was performed to enrich the library for sequences containing the transposon-chromosome junction. The following reaction was set up in a low-profile optical lid PCR tube: a) TruSeq gDNA library (25 ng), b) Phusion HS buffer, 5x (10 µL), c) 10 mM dNTPs (2.5 µL), d) 10x SYBR Green (1 µL), e) Tn primer (10 µM) (2.5 uL), f) Tn indexed primer (10 µM) (2.5 uL), g) Phusion (1 uL), h) H_2O filled to a final volume of 50 uL. These tubes were placed in the qPCR machine and run at 95 ^oC for 2 minutes, and 25 cycles 95 ^oC for 15 seconds, 60 ^oC for 30 seconds, and 72 ^oC for 30 seconds. Note that reactions were stopped ~3-5 cycles after PCR reached log amplification (~15 to 20 cycles), and tubes were transferred to another PCR machine for 10 minutes of final extension at 72 ^oC. A standard clean-up protocol was followed using sample purification beads (TruSeq kit). Specifically, the bead to DNA ratio was 50 µL SPB to 50 µL PCR reaction, and the final elution from the beads was 30 µL of water. Finally, ~30 uL of the Tn-seq library was recovered. The final library concentration was measured using Qubit, while fragment size distribution was characterized on an Agilent Bioanalyzer with an HS DNA chip. Average sizes for each library ranged between 400 and 600 bp. Lastly, libraries were sequenced on a MiSeq using a custom sequencing primer for Read 1: GCATGCAAGCTTCAGGGTTGAGATGTGTATAAG.

Sequences of primers and library intermediates

<KAN2> transposon:

CTGTCTCTTATACACATCTCAACCATCATCGATGAATTGTGTCTCAAAATCTCTGATGTTACATTGCACAAGATAAAAATAT
ATCATCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGT
CTTGCTCGAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAAT
CAGGTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCC
AATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCATTTTATCCGT
ACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCCGGAAAAACAGCATTCCAGGTATTAGAAGAATATCCTGATTCA
GGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGA
TCGCGTATTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAA
TGGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGAT
TTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTCGGAATCGCAGACCGAT
ACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATT
GATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATCAGAATTGGTTAATTGGTTGTAACA
CTGGCAGAGCATTACGCTGACTTGACGGGACGGCGGCTTTGTTGAATAAATCGAACTTTTGCTGAGTTGAAGGATCAGATC
ACGCATCTTCCCGACAACGCAGACCGTTCCGTGGCAAAGCAAAAGTTCAAAATCACCAACTGGTCCACCTACAACAAAGC
TCTCATCAACCGTGGCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTCAGGGTTGAGATGTGTATAAGAGACA
G

Tn-specific Priming site

Sheared DNA (avg. = 350 bp)

5' - [<KAN2>DNA]AGGCATGCAAGCTTCAGGGTTGAGATGTGTATAAGAGACAG[chromosomalDNA] - 3'

End repair and adenylation

5' - [<KAN2>DNA]AGGCATGCAAGCTTCAGGGTTGAGATGTGTATAAGAGACAG[chromosomalDNA]A - 3'

Adapter ligation

5' -
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT---T[<KAN2>DNA]AGGCATGCAAGCTTCA
GGGTTGAGATGTGTATAAGAGACAG[chromosomalDNA]A---GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTC
GTATGCCGTCTTCTGCTTG - 3'

Universal Adapter
5' - AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT - 3'

Adapter Index 4
5' - GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG - 3'

P5: 5'- AATGATACGGCGACCACCGA - 3' *(+) on universal adapter
P7: 5' - CAAGCAGAAGACGGCATACGA - 3' *(-) on indexed adapter

Tn-specific PCR

Tn-specific primer
AATGATACGGCGACCACCGAGATCTACAAGGCATGCAAGCTTCAGGGTTGAGATGTGTATAAG

Indexed Primer (AD004)
CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD005)
CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD007)
CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD012)
CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD018)
CAAGCAGAAGACGGCATACGAGATGCGGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD002)
CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD003)
CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD006)
CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Indexed Primer (AD008)
CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Final library (ready for sequencing)
example: AD004
5' -
AATGATACGGCGACCACCGAGATCTACAAGGCATGCAAGCTTCAGGGTTGAGATGTGTATAAGAGACAG[chromosomalDNA]A--
-(A)GATCGGAAGAGCACACGTCTGAACTCCAGTCAC**TGACCA**ATCTCGTATGCCGTCTTCTGCTTG - 3'

Sequencing primers:
Seq1_Tn: GCATGCAAGCTTCAGGGTTGAGATGTGTATAAG (on Tn-specific primer)
SeqIn: GATCGGAAGAGCACACGTCTGAACTCCAGTCAC (on index primer)
Seq2: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT (on index primer)

References

1. LaCroix RA, et al. (2015) Discovery of key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal media using adaptive laboratory evolution. *Appl Environ Microbiol* 81(1):17–30.

2. Taniguchi Y, et al. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533–538.

3. Peebo K, et al. (2014) Coordinated activation of PTA-ACS and TCA cycles strongly reduces overflow metabolism of acetate in Escherichia coli. *Appl Microbiol Biotechnol* 98(11):5131–5143.

4. Kim D, et al. (2012) Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling. *PLoS Genet* 8(8):e1002867.

5. Cho B-K, Kim D, Knight EM, Zengler K, Palsson BO (2014) Genome-scale reconstruction of the sigma factor network in Escherichia coli: topology and functional states. *BMC Biol* 12:4.

6. Poliakov A, et al. (2011) Large-scale label-free quantitative proteomics of the pea aphid-Buchnera symbiosis. *Mol Cell Proteomics* 10(6):M110.007039.

7. Diedenhofen B, Musch J (2015) cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS One* 10(3):e0121945.

8. Thiele I, Jamshidi N, Fleming RMT, Palsson BØ (2009) Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* 5(3):e1000312.

9. Anders S, Pyl PT, Huber W (2014) HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*.

10. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550.

11. Whitlock MC (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol*.

12. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*.

13. R Core Team (2014) *R: A language and environment for statistical computing* (Vienna, Austria).

14. Zou GY (2007) Toward using confidence intervals to compare correlations. *Psychol Methods* 12(4):399–413.

15. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. *Science* 315(5814):972–976.

16. Bodenhofer U, Kothmeier A, Hochreiter S (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics* 27(17):2463–2464.

17. Lewis NE, Cho B-K, 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(11):3437–3444.

18. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol Syst Biol* 9(1):693.

19. Oh Y-K, Palsson BO, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* 282(39):28791–28799.

20. Commichau FM, Pietack N, Stülke J (2013) Essential genes in Bacillus subtilis: a re-evaluation after ten years. *Mol Biosyst* 9(6):1068–1075.

21. Baba T, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2:2006.0008.

22. Robin X, et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77.

23. Prickett MD, Page M, Douglas AE, Thomas GH (2006) BuchneraBASE: a post-genomic resource for Buchnera sp. APS. *Bioinformatics* 22(5):641–642.

24. Lerman JA, Chang RL, Hyduke DR (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol Syst Biol*.

25. Aziz RK, et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.

26. Overbeek R, et al. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42:D206–14.

27. Aziz RK, et al. (2012) SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS One* 7(10):e48053.

28. Carrera J, et al. (2014) An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. *Mol Syst Biol* 10:735.

29. Reiss DJ, Baliga NS, Bonneau R (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* 7:280.

30. Zomer A, Burghout P, Bootsma HJ, Hermans PWM, van Hijum SAFT (2012) ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS One* 7(8):e43012.

31. Baba T, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2(1):2006.0008.

32. Grenier F, Matteau D, Baby V, Rodrigue S (2014) Complete Genome Sequence of Escherichia coli BW25113. *Genome Announc* 2:e01038–14.