Supporting Information

Network-level architecture and the evolutionary potential of underground metabolism

Richard A. Notebaart, Balázs Szappanos, Bálint Kintses, Ferenc Pál, Ádám Györkei, Balázs Bogos, Viktória Lázár, Réka Spohn, Bálint Csörgő, Allon Wagner, Eytan Ruppin, Csaba Pál, Balázs Papp

Supporting Text

Estimating the degree of overexpression from the ASKA plasmids

The typical protein expression level difference between the native genomic locus and the high-copy overexpression plasmid was estimated by two approaches. In the first approach, we used literature data to calculate the fold change overexpression of a typical enzyme. A quantitative systems-wide analysis of protein copy numbers in individual *E. coli* cells provides information for 291 enzymes (1). The median expression level of these enzymes in the native genomic context is 22.6 copy/cell. While a similar comprehensive dataset for the expression levels from the high copy plasmid pCA24N is not available, the protein copy number for one overexpressed enzyme has been estimated as $>1.8x10^5$ copies/cell (2). This gives a typical expression level difference of ~4 orders of magnitude.

In the second approach, we experimentally measured the activity of one of the enzymes (LacZ) which was identified as a positive hit both in our *in silico* and experimental surveys for phenotypic novelties (i.e. LacZ increases fitness on phenylgalactoside when strongly overexpressed). Following standard protocols (3), we carried out enzyme activity measurements using the lysate of *E. coli* cells that were grown in the same conditions as the samples for our genome-wide overexpression screen (see SI Materials and Methods for details). Comparing the β-galactosidase activity from cells harbouring the plasmid pCA24N-LacZ-(gfp-) and the empty vector PCA24N (without an ORF) using the substrate ortho-nitrophenyl-β-galactoside (3), we detected an expression level difference of ~500. We note that this figure is likely to be an underestimate of the typical expression difference as the 50 μM isopropyl-β-D-thiogalactopyranoside (IPTG) induces not only the expression of the plasmid-encoded LacZ, but

also that of the native gene copy as well. Taken together, we estimate that the typical fold change increase in protein expression level in our experimental setup is in the range of 3 – 4 orders of magnitude.

SI Materials and Methods

Reconstruction of the *Escherichia coli* underground metabolism

The *Escherichia coli* underground metabolism reconstruction (hereby termed iRN1260u) is an extension to the native metabolic network of *Escherichia coli* K-12 MG1655 (4), termed iAF1260. For each EC number in the iAF1260 reconstruction we collected and evaluated weak side activities from experimental studies included in the BRENDA database. Reactions were considered as side reactions (underground reactions) when the corresponding substrates were listed in the BRENDA 'substrate' section, but not in the 'natural substrates' section. To evaluate the correctness of the classification, we examined the associated literature to ensure that the reaction was indeed experimentally studied for that particular enzyme. Then we examined the chemical properties (e.g. atom composition) of the substrates and products of the reactions as stored in the KEGG, PubChem, ChEBI, and ChemIDPlus databases. If the information extracted from the individual databases was consistent, then the reaction was further examined as a whole for correct stoichiometry, i.e. atom and electron compositions of the substrates equal that of the products. Finally, the reactions were added to the existing iAF1260 reconstruction and they were linked to the corresponding enzymes. The metabolite abbreviations were taken from the iAF1260 model when possible. In addition to BRENDA, underground reactions were added directly from the literature (for details see Dataset S1). In particular, numerous underground activities from different haloacid dehalogenase-like hydrolases were obtained from a systematic medium-scale study by Kutznetsova et al. (5) and were included in the reconstruction.

Beyond the reconstruction of underground reactions, our framework also allowed the extension of the *E. coli* iAF1260 network by 122 new native reactions. The data comes from

literature information, the BRENDA (section 'natural substrates'), and the BiGG (6) databases. To broaden the space of possible growth environments, this list of 122 new native reactions also includes 65 new exchange (transport) reactions. These transport processes were included in the reconstruction to maximize the coverage of nutrient metabolites that can possibly be taken up by the cell. Among these are the transporters involved in true-positive predictions of the metabolic model, i.e., predicted and experimentally confirmed novel growth phenotypes provided by underground metabolism (see results in the main text). The complete dataset can be found in the Supplementary information as an Excel file (Dataset S1) and as a computational SBML model (Dataset S6, also downloadable from http://group.szbk.u-szeged.hu/sysbiol/papp-balazs-lab-resources.html). Throughout the paper, the *native* network (*E. coli* iAF1260 + 122 new native reactions) was tested against the *extended* network (*native* network + one or more underground reactions, see below).

Evaluation of the underground metabolic reconstruction

The general approach of adding underground reactions to the native metabolic repertoire as described above was evaluated by examining published enzyme kinetic parameters and large-scale metabolomic data sets. We performed two analyses to demonstrate that the assembled underground reactions occur at very low rates in *E. coli*.

First, we tested whether underground reactions are catalyzed by lower kinetic efficiency compared to the native reactions of the same enzymes. To this end, we collected reactions from BRENDA with measured k_{cat} and K_m values. We were able to retrieve high quality data for 13 different enzymes for which the *in vitro* conditions for native and underground substrates were

the same (see Table S1). Subsequently, the catalytic efficiencies (k_{cat} / K_m) of the native and underground reactions of the same enzyme were compared using a Wilcoxon matched-pairs signed rank test.

Second, we tested whether *underground metabolites* (i.e. those consumed and/or produced by underground reactions only) are less likely to be present in the metabolome of *E. coli* compared to *native metabolites* (i.e. those consumed and/or produced by at least one native reaction). We used two complementary information sources for this test: metabolite presence from i) the *E. coli* Metabolome Database (ECMDB) (7) and ii) from two large-scale metabolomics studies (8, 9). 730 compounds from our reconstruction could be matched with metabolites from ECMDB using KEGG, CAS, or ChEBI identifiers. In a similar vein, 488 metabolites from our reconstruction could be matched with compounds in the union of the two large-scale metabolomics studies (8, 9).

Network distance and shared subsystems between native and underground reactions

The dataset

To test whether pairs of native and underground reactions belonging to the same enzyme are non-randomly distributed in the network, we focused on a subset of our compiled underground reaction set. We used data from a systematic measurement of phosphatase activities in the haloacid dehalogenase-like phosphatase family of *E. coli* (5). It includes enzymes acting on a wide range of phosphorylated metabolites, including carbohydrates, nucleotides, organic acids, coenzymes, and small phosphodonors. This study investigated the activity of 23 enzymes against a panel of 80 substrates in an all-against-all fashion, thereby providing comprehensive data on pairs of native – underground reactions that are catalyzed by the same enzyme. We considered a

reaction to be underground if the activity of the enzyme towards it was an order of magnitude lower than the activity towards the primary substrate (i.e. the one with the highest activity). We chose this threshold because it captured 90% of the native activities published in the iAF1260 reconstruction for the corresponding enzymes. Processing this dataset yielded 350 native-underground reaction pairs for our analysis (Dataset S5).

Network distance and subsystem classification

The network distance between each native – underground reaction pair was computed using the shortest distance method implemented in the igraph software package (10). For this purpose, we built a graph representation of the metabolic network in which two reactions were defined as adjacent if the product of one reaction is the substrate of the other. Cofactors and small molecules with high abundance in the network (H_2O, NH_3, etc.) were excluded from the graph.

Subsystem classification of reactions was taken from the iAF1260 reconstruction. Functional classification was not available for 32% of reactions in our dataset. In such cases, the corresponding subsystem classes were predicted based on the classification of the closest native reactions in the network. Specifically, for each reaction with unknown subsystem classification, we first determined its neighboring reactions based on the graph representation of the network. Next, we assigned the classes of the neighbors to the unclassified reaction by taking into account the number of reactions (n) in which their intermediate metabolite is participating. In particular, each assigned class was weighted by $1/n$ of the metabolite linking the unclassified and the classified reaction (thus, reactions connected through rare metabolites had a stronger influence on the class assignment). Finally, the reaction with the unknown subsystem classification was

assigned the class with the highest score. Leave-one-out validation on native reactions showed that this simple prediction algorithm has 72% accuracy for predicting subsystem memberships.

Statistical procedure

The statistical significances of the network distance and subsystem co-membership observed for the 350 native – underground reaction pairs were assessed by employing a randomization procedure. To this end, we randomized the enzyme-reaction associations in the dataset in a series of iterative steps as follows. Each enzyme-reaction association in our dataset is labelled as native, underground or no association. In each step we randomly selected two enzyme-reaction associations E_1-R_1 and E_2-R_2 in such a way that the label of E_1-R_1 is identical to that of E_2-R_2 (and is either native or underground) and the label of E_1-R_2 is identical to that of E_2-R_1 and differs from E_1-R_1. Next, we swapped the labels of E_1-R_1 and E_1-R_2 and those of E_2-R_1 and E_2-R_2.This randomization process was repeated for another pair of reaction-enzyme associations until all were targeted at least once. This procedure conserves the degree distribution of both enzymes and reactions while randomizing the labels of enzyme-reaction associations. We note that, on average, the randomization process resulted in four changes per enzyme-reaction association. The set of randomized native-underground reaction pairs was generated from the randomized enzyme-reaction associations. We generated 10,000 such random variant native-underground sets for statistical analyses.

To investigate whether the network-level distribution of underground – native reaction pairs is influenced by the chemical similarity of their substrates, we also performed a modified statistical analysis in which the chemical similarity of the substrates was constrained. Chemical similarity of substrate pairs was measured by the Tanimoto similarity of the corresponding

chemical fingerprints (11). During randomization, swapping was allowed only if the chemical similarities of the substrates of reaction A and B were above a certain similarity threshold. Results of this modified randomization procedure are shown in Figure S2.

Identifying reactions that are capable of carrying a non-zero flux

The capability of each reaction of carrying a non-zero flux in steady state was determined as previously (12). Briefly, using the constraint-based modelling framework of flux balance analysis (13), we minimized and maximized in turn each individual reaction flux under the governing biochemical constraints and identified those reactions for which both the minimal and maximal achievable flux was zero. These reactions were considered to be incapable of carrying a flux in steady state (i.e. blocked). To estimate the set of reactions that can carry a flux in the *native* and *extended* network, respectively, we considered a condition where all external nutrients are available for uptake and secretion; this scenario identifies reactions that can possibly be active under at least some conditions. We identified 1257 and 1393 non-transport reactions that are capable of carrying a non-zero flux in the *native* and in the *extended* network, respectively.

Elementary Flux Mode analysis

Elementary Flux Mode (EFM) analysis was employed for two purposes: i) to examine whether underground reactions participate in biomass-forming pathways and ii) to investigate the properties of these pathways. EFM (14) is a mathematical representation of a biochemical

pathway and is defined as a minimal set of reactions that can operate at steady state while taking into account the directionality of irreversible reactions. Since enumerating all EFMs in a genome-scale metabolic network is computationally infeasible (15), we employed and modified a previously published method (16) to obtain a sample of EFMs.

According to the sampling algorithm of Kaleta et al. (16), a single random EFM can be obtained by minimizing the sum of fluxes if i) the network is in steady state, ii) all reversible reactions are represented as two irreversible reactions, both carrying only positive fluxes, iii) an objective reaction is constrained to carry a non-zero flux, and iv) a random set of reactions is removed from the network. We modified this method to gain a random sample of EFMs containing two objective reactions (e.g. a biomass reaction and a selected other reaction of interest). Accordingly, we constrained two objective reactions to carry non-zero fluxes and applied the minimization method of Kaleta et al. (16). However, the minimization of the sum of fluxes in a network where two reactions are constrained to carry non-zero fluxes might result in a steady-state flux distribution that is not an EFM (e.g. two separate reaction cycles each containing one of the objective reactions). To filter out these biologically irrelevant cases, we evaluated each candidate EFM by performing a second minimization in which only one objective was constrained to a non-zero level while all reactions that were not active in the candidate EFM were removed. If the flux distribution after the second minimization step matched the candidate EFM, it was accepted as a verified EFM. We repeated the above procedure for each reaction of interest until we gained a sample of 100,000 verified EFMs. Subsequently, all EFMs were normalized by the glucose uptake flux, to ensure that the yield of an EFM was equivalent to the flux of biomass reaction. In addition to calculating biomass yield, we also counted the number of reactions in each EFM and used as a measure of pathway length.

Properties of EFMs that contain either the native or the underground reaction of the same enzyme were compared using a one-sample t-test on the rank-sum test statistics of the paired samples. This test is especially robust to differences in the distributions of the paired samples. Accordingly, for each enzyme with an assigned underground reaction, we first divided EFMs into two groups based on whether they contained the native reaction(s) or the underground reaction(s) of the enzyme. Next, we ranked all EFMs from these two sets based on their property of interest (e.g. yield) and calculated the rank-sum test statistic to compare these two populations. Such a test statistic was in turn determined for each enzyme. Finally, for both pathway yield and length, we applied a one-sample t-test on the rank-sum test statistics to examine whether EFMs containing the underground reaction show an overall difference from those containing the native reaction.

Metabolite toxicity analysis

To characterize the toxicity of metabolites, we predicted IC_{50} values (half maximum inhibitory concentration) of both *native* and *underground* metabolites using a QSAR-based algorithm called EcoliTox (17). The algorithm was trained on experimentally determined IC_{50} values of a diverse set of 94 compounds in *E. coli* and predicts IC_{50} based on molecule structure with high accuracy (R^2=0.71 between predicted and measured values). We note that the set of 94 training compounds was selected to provide a representative set of chemicals with maximal chemical diversity and is not specific to the *E. coli* metabolome. Thus, the method is expected to accurately predict the toxicity of both native *E. coli* metabolites and that of compounds not present in *E. coli*. Molecule structures in mol file format were automatically extracted from

KEGG and ECMDB databases using KEGG identifiers of the metabolites and supplemented with manually retrieved structural data when KEGG ID was unavailable.

To test whether *underground metabolites* show especially low IC_{50} values, we performed pairwise tests as follows. First, we selected underground reactions in which either the substrate or the product (or both) is a novel metabolite in the network. Then we compared, in a pairwise manner, the IC_{50} values of the *underground* and the corresponding *native metabolites* associated with the same enzymes using a Wilcoxon rank-sum test. In the case of multiple substrates and/or products per enzyme, we took their mean value before conducting the pairwise comparison. Repeating the analysis by separately considering substrates and products for each enzyme gave very similar results (P=0.77).

Expression Dependent Gene Effects (EDGE) analysis

The EDGE algorithm (18) quantifies the consequences of inducing the expression of metabolic genes on the objective function (e.g., maximum growth flux) within a genome-scale metabolic network. The EDGE score of a reaction determines whether the enzyme is (i) *beneficial* - contributing towards the realization of the objective (positive score), (ii) *detrimental* to the objective (negative score), or (iii) *neutral* with respect to the objective (zero score). The score is the result of comparing the maximal value of the objective function between two functional states of the reaction: (a) when the reaction carries a minimal finite flux and (b) when the reaction is constrained to carry no flux. In the case of reversible reactions in state (a), a minimal flux through either direction is required, and the minimum value of the objective function is taken into account:

$$E_r = min\left(GR_{up}^{forward}, GR_{up}^{backward}\right) - GR_{ko}$$

where E_r is the EDGE score of reaction *r*, *GR* is the objective function and *up* stands for state (a) while *ko* stands for state (b).

We calculated the EDGE scores for all the native and underground reactions and defined their toxic (detrimental, i.e., negative score) or non/toxic (beneficial or neutral, i.e., positive or zero score, respectively) character. Subsequently, we compared the frequency of toxic reactions among the native and the underground reactions associated with the same enzyme using a paired Wilcoxon test. Essential reactions were excluded from both tests because, by definition, only native reactions can be essential under standard nutrient conditions and, according to the definition of EDGE score, essential reactions cannot be toxic.

In silico analysis of the adaptive potential of underground metabolism in novel environments

Identifying environments in which underground reactions confer growth advantage

To assess whether underground reactions confer growth in non-standard conditions, we defined a comprehensive set of minimal media to test for *in silico* growth. The list includes 2754 minimal media encompassing the complete range of nutrient sources that can be imported into the network (Dataset S2). First, we determined which nutrients can be utilized as carbon (C), nitrogen (N), phosphorus (P), and sulfur (S) sources in a minimal medium in the presence and absence of oxygen by the native metabolic model and by the native model extended with all underground activities (extended model). We employed flux balance analysis (FBA) to calculate growth capabilities across conditions using the Sybil package (19) in the R programming

environment (20). FBA is a computational technique for identifying the maximal biomass yield (a proxy for growth) of large-scale metabolic networks in steady-state (13). The *native* and the *extended* models were able to grow (i.e. produce biomass) in 645 and 664 conditions, respectively. For further analyses we used those conditions in which the *extended* model can grow. Conditions in which the underground reactions conferred a benefit were defined as those where adding the underground reaction set increased biomass yield by at least 5% over the native network. We applied this threshold to ignore marginally small and potentially irrelevant gains in biomass yield. The mathematical framework of flux balance analysis relies on maximizing the biomass yield within the feasible solution space defined by the structure of the network and the constraints imposed. Because adding hundreds of reactions to the native network increases the size of the feasible solution space, the *extended network* is expected to have a slightly increased maximum yield under most conditions.

Identifying underground reactions that confer a growth advantage in novel nutrient environments

To identify underground reactions which confer a growth advantage either individually or in combination with other underground reactions, we focused on conditions where the extended network had at least a 5% growth advantage. First, we employed an exhaustive *in silico* multiple reaction knockout test (21) to identify all underground reactions contributing to growth advantage in a given condition. Next, we reinserted individual reactions or reaction sets showing considerable phenotypes into the native network and tested whether they recovered the growth advantage observed for the complete set of underground reactions. We considered a phenotype

recovered if removing the reaction(s) from the extended network resulted in the loss of the growth advantage, while the subsequent reinsertion of the reaction(s) into the native network resulted in the regain of the growth advantage with as few reactions as possible. Loss and regain of a phenotype was defined as the relative growth advantage being less than 20% and more than 80% of that seen for the extended network. The rationale behind these thresholds is that only a mere 1.05% of all reaction knockout phenotypes fall between these two values. All observed growth advantages can be recovered by adding at most three underground reactions to the native network.

Genome-wide screening for genes that can enhance growth in novel environments when overexpressed

We carried out a genome-wide screen to identify underground reactions that enhance growth when their activity is amplified. The impact of *E. coli* ORF overexpression on growth efficiency was tested in a wide array of carbon sources using a protocol adopted from Soo et al. (22). A graphical overview of the workflow is presented in Figure S3.

Basic procedures of the competition experiments

In brief, the method relies on the competition of a population of cells transformed by a pooled library of overexpression plasmids in the presence of an ordered array of carbon sources, followed by the identification of ORFs that were enriched during the competition experiment. To this end, the complete set of *E. coli* K-12 Open Reading Frame Archive library (ASKA) was grown in the original host strain *E. coli* K-12 AG1 (23) in 96-well plates (growth conditions: 37

Cº, 280 rpm, LB medium). An equal aliquot of each member of the ASKA library (each well of the 96 well plates) was pooled together and the plasmid DNA (pCA24N-ORF-GFP(-)) was isolated. The resulting plasmid preparation of the ASKA library, as well as the empty vector pCA24N (without an ORF), were transformed into *E. coli* MG1655 by electroporation. The transformed samples were grown in mineral salts minimal medium (MS-minimal) that contains 0.8% glycerol as a sole carbon source to avoid glucose-mediated catabolite repression. When cell culture density reached OD_{600}=1, protein expression was induced by isopropyl-β-D-thiogalactopyranoside (IPTG) at a concentration (50 μM) where the growth effect of overexpression toxicity is reduced (22). After 2h of induction, cells were harvested by centrifugation and re-suspended in MS-minimal without carbon source and were starved for 1h at room temperature. At this point, a fraction of the cells from the ASKA pool and the negative control (empty plasmid) were used to isolate plasmid DNA. These preparations were subjected to next-generation sequencing with the SOLiD System (Life Technologies) to determine the diversity of the pooled library. Samples for sequencing were prepared as described previously (24). To evaluate the presence of each ORF, the number of sequence reads covering the corresponding ORF was compared in the two samples in the same manner as in RNA-Seq whole-transcriptome analyses. ORFs in the ASKA pool covered by more reads than 95% of the ORFs in the negative control (genomic background) were considered as present. 86.7% of the known enzyme-encoding ORFs were detected in the pooled ASKA library at the beginning of the growth experiments.

Nutrient conditions studied

The set of investigated nutrient environments included 190 different carbon sources present on PM1 and PM2A phenotype microarray plates (25) (Biolog Inc.). This list was extended by a set of 4 carbon sources on which the *in silico* analysis predicted innovation via underground reactions, but were not present on either PM1 or PM2A plates (Ethylene glycol, L-Glyceraldehyde, D-Lyxose and Phenylgalactoside). To prepare the nutrient array, carbon sources were suspended in 100 µl MS-minimal media supplemented with 50 µM IPTG and 20 µg/ml chloramphenicol, then transferred to standard 96-well plates. The final concentrations of the carbon sources not present on phenotype microarray plates were set to 20 mM. Each well of the 96-well plates was inoculated with approximately $2*10^6$ cells from the above-described cell preparations either expressing the pooled ASKA library or the empty plasmid.

Monitoring growth and detecting growth differences

Growth was monitored at OD_{600} for 7 days (30 C°, 1000 rpm in automated plate readers (BioTek Inc, USA). The pooled ASKA library and the negative control (empty vector) populations were always monitored simultaneously on two identically arranged 96-well plates. The temperature was decreased to 30 C° to minimize the heat-induced unfolding of the overexpressed proteins and to minimize evaporation of the samples. For the latter reason, plates were also sealed with gas permeable sealing foil (Breath Easy, Sigma). Two biologically independent replicates of these experiments were carried out. Following the 7^{th} day, 2 µl from each well was used to inoculate fresh media of the corresponding carbon sources and growth was monitored for an extra 5 days. After the 5^{th} day, a second transfer was carried out to let the cells grow on fresh

medium for another 5 days. This gradual library enrichment procedure was expected to enable the detection of even minor growth differences between the ASKA pool and the negative control.

Growth curves derived from OD_{600} measurements were analyzed both during and at the end of the incubation periods. To increase the number of potential hits in the initial screen, we used two criteria to identify conditions where an ORF possibly increases growth when overexpressed: (i) Visually discernible difference between the growth curve of the ASKA pool and that of the negative control or (ii) The growth curve of the ASKA pool has a reproducibly larger integral (i.e. area under the growth curve) than that of the negative control in any of the three rounds of library enrichments. Carbon sources fulfilling any of these criteria were considered for an ASKA ORF enrichment analysis (see protocol below). To ensure that no relevant carbon source remained unexplored, they were ranked based on the differences between the growth curve integrals of the ASKA pool and that of the negative control after each round of library enrichment (Dataset S3). Starting from the top of the lists (i.e. highest integral differences), we performed the ORF enrichment test on each carbon source until no further carbon sources with enriched ORFs were found (Dataset S3).

Identifying enriched ORFs

The enrichment of ORFs during the competition experiments was tested with the following protocol (adopted from Soo et al. (22) with minor modifications): cells from wells were plated out onto LB agar plates with chloramphenicol. Using vector-specific primers, the ASKA ORFs were amplified by PCR from at least 8 of the resulting colonies. The sizes of the PCR products were analyzed using agarose gel electrophoresis. PCR product pairs with the same size were purified (Zymo Clean and Concentrator kit) and sent to sequencing. ASKA ORFs represented at

least twice out of the 8 examined colonies were considered as candidates for ORFs that provide enhanced growth when overexpressed. Out of the 194 carbon sources 63 were tested for enriched ORFs and 41 of them gave candidate ORFs for further examination. In 10 of these carbon sources, the very same ORF (*eutE*) showed a marginal enrichment. As this set of 10 carbon sources proved to be very diverse, the growth effect of *eutE* was considered as aspecific and excluded from further analysis (for details, see Dataset S3).

Verification of genes that enhance growth in novel environments when overexpressed

To verify the beneficial effect of the identified ORFs, additional growth experiments were carried out with specific overexpressing ASKA strain – carbon source pairs only. In brief, instead of competing a pooled collection of ASKA strains, here we tested the impact of individual ORFs on growth in specific environments. 27 ORFs identified in the screening procedure were re-tested on one or more of the 31 corresponding carbon sources (Dataset S3). Each ORF was isolated from the original ASKA collection and, following PCR verification, was retransformed into *E. coli* MG1655. Growth conditions and ORF induction were as described above. Growth measurements were carried out in three biological replicates for 4 days. Following the calculation of the growth curve integrals, a statistical procedure (one-sided paired t-test) was applied to determine the statistical significance of the growth differences between the ASKA strains and the corresponding negative controls. Growth curves of those ORF – nutrient pairs which showed a significant difference are plotted in Figure S4 and Figure 4B. This dataset was complemented by those ORF – carbon source pairs which were predicted *in silico* as novel environments where an underground activity is essential for growth (light green squares in Figure 3B), but were not detected in the high-throughput screen. Overall, we could test 7 such

cases experimentally for growth (see cases denoted by asterisk in Table S2B; the carbon source 5-Amino-4-oxopentanoate could not be obtained commercially and the ORF responsible for the degradation of D-Arabitol could not be identified, thus these cases were excluded from verification). Out of the 7 tested ORF – carbon source pairs only 1 displayed growth in the verification assay (*fumB* in D-Tartrate), indicating that the genome-wide screen has a low false negative rate.

Figure S1. Novel metabolites introduced by underground reactions are underrepresented among empirically observed metabolites in metabolome datasets.

Number of native and underground metabolites that are present / absent in large-scale metabolomics studies (8, 9) (see SI Materials and Methods for details). The difference is highly significant ($P<10^{-15}$, Fisher's exact test). Underground metabolites are defined as those that are consumed and/or produced by underground reactions only.

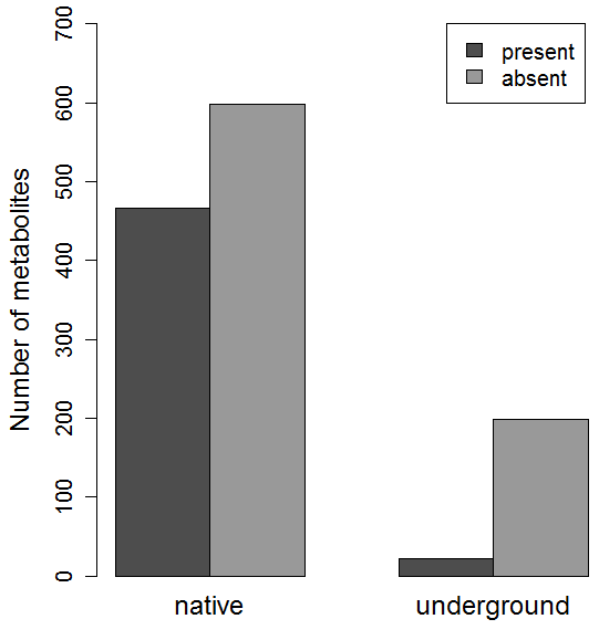
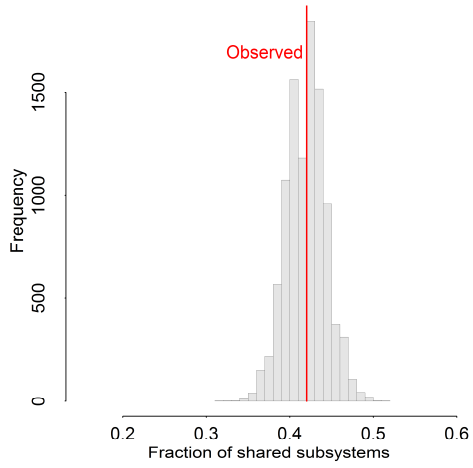


Figure S2. Functional relatedness and short network distance between native and underground reactions is a by-product of chemical similarity.

As shown in the main text (Figure 2 B-C), native and underground activities of the same enzyme are close in the network and often belong to the same metabolic subsystem. Here we test whether this non-random pattern remains statistically significant after controlling for the chemical similarity between reactants of native and underground reactions. To this end, we performed the same randomization test as described in SI Materials and Methods, but only swapped the labels of reaction pairs (i.e. native or underground) if their substrates showed a chemical fingerprint similarity of 0.4 or higher (as measured by the Tanimoto similarity coefficient (11)). Neither the high fraction of shared subsystems **(A)**, nor the short network distances **(B)** between native – underground reaction pairs remain significant when only reactions with chemically similar substrates are randomized between enzymes (P=0.46 and P=0.2, respectively; red lines indicate observed values for underground - native reaction pairs annotated to the same enzymes in our reconstruction.). This indicates that the functional relatedness between native and underground activities of the same enzymes is a by-product of chemical constraints.

A)



B)

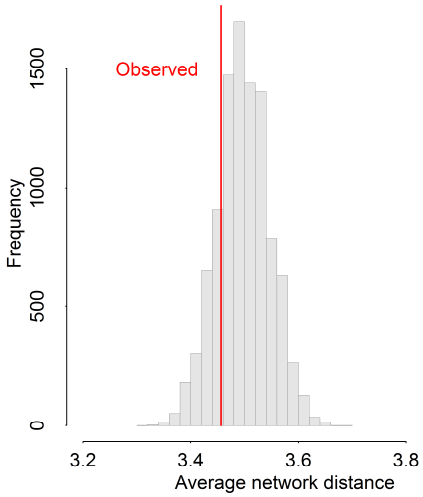

Observed
Frequency
0
500
1000
1500
3.2
3.4
3.6
3.8
Average network distance

Figure S3. Workflow of the genome-wide overexpression screen.

The complete set of *E. coli* K-12 Open Reading Frame Archive library (ASKA) was pooled together and the resulting plasmid preparation was used to transform *E. coli* MG1655 cells. These cells, as well as the negative control, were grown in liquid culture and expression was induced with 50µM of IPTG. Cells from both cultures were used to inoculate an array of 194 carbon sources. Growth of the pooled library was compared to that of the negative control at OD_{600} to identify carbon sources where the ASKA pool grows more efficiently than the negative control. To gradually enrich overexpression plasmids conferring a growth benefit, cells from each well were transferred into a fresh array of carbon sources and growth was monitored for two additional rounds. Cells from positive wells were plated onto LB agar plates. Eight of the resulting colonies were subjected to PCR analysis and sequencing to identify ORFs the plasmids of which had been enriched during growth. The procedure was adopted from Soo et al. (22) with modifications. See SI Materials and Methods for more details.

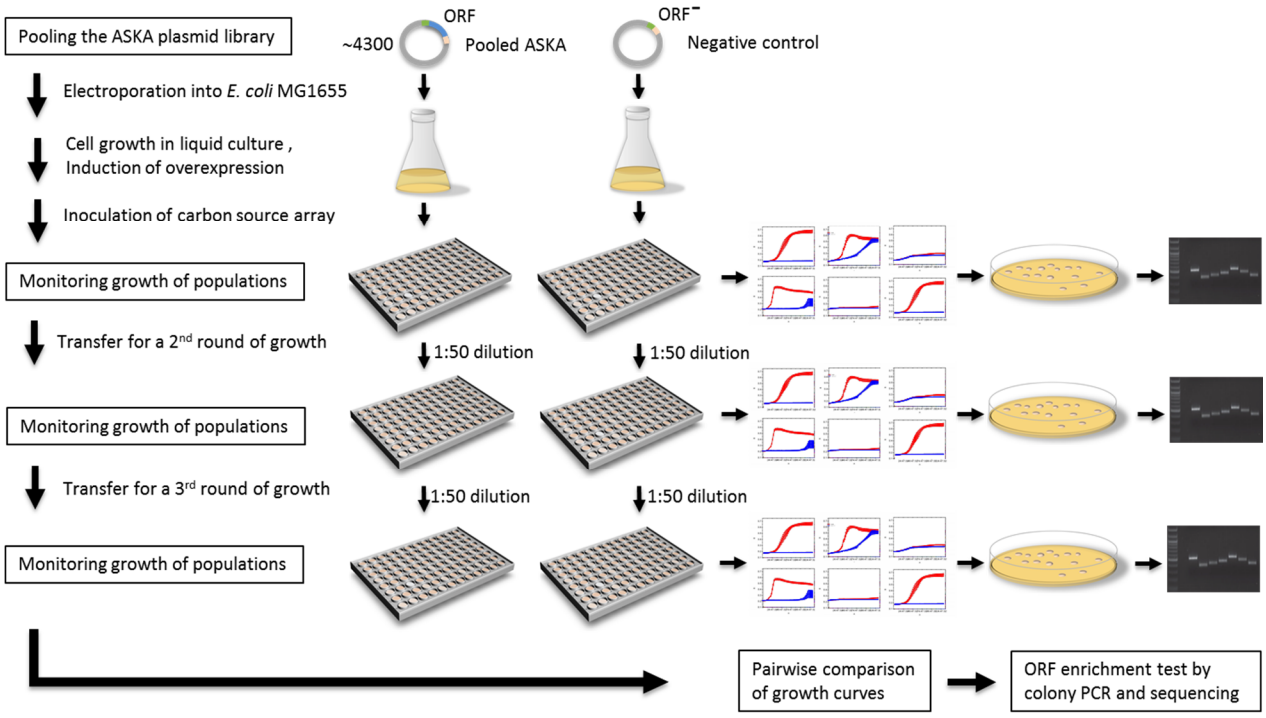
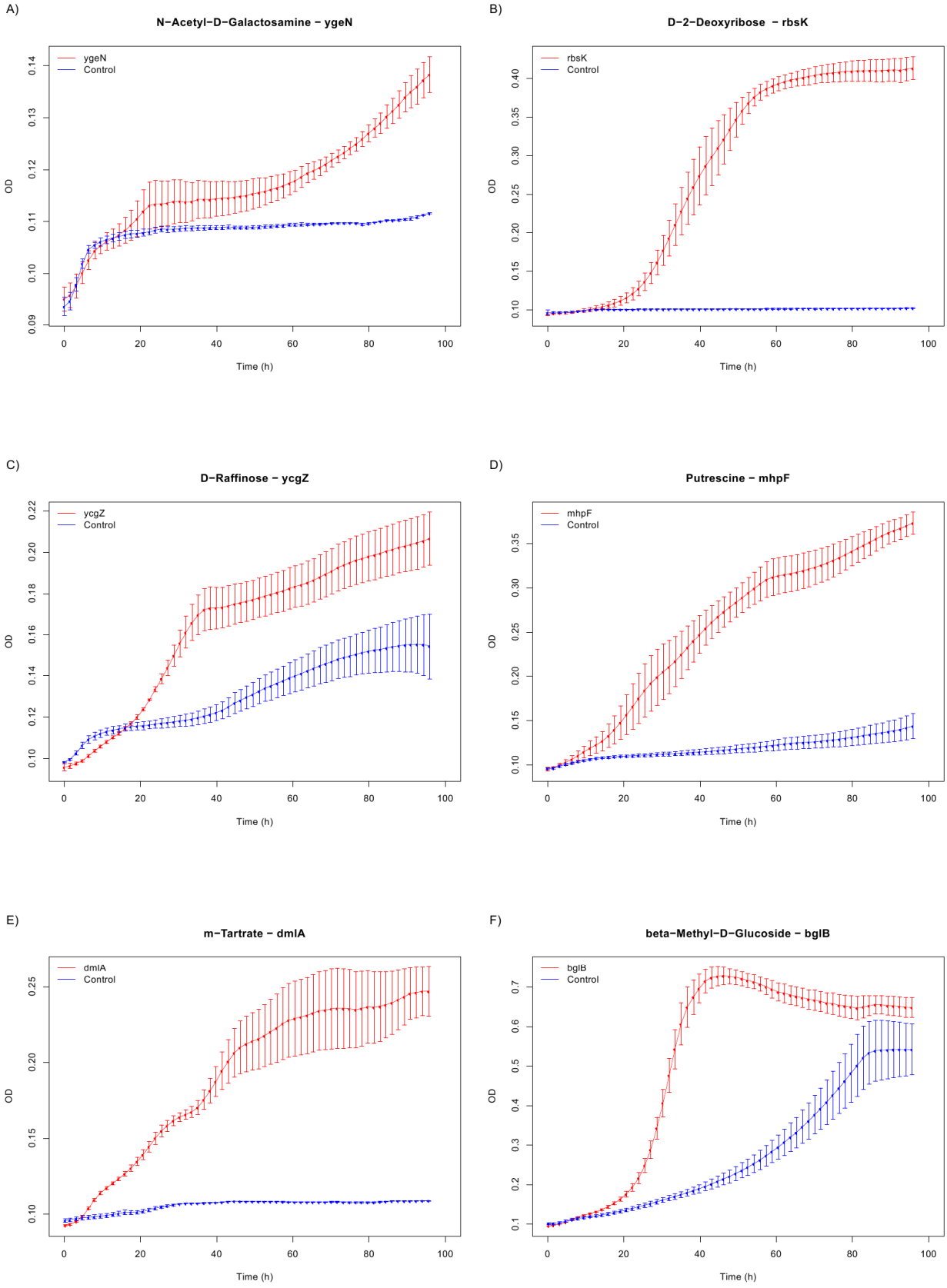
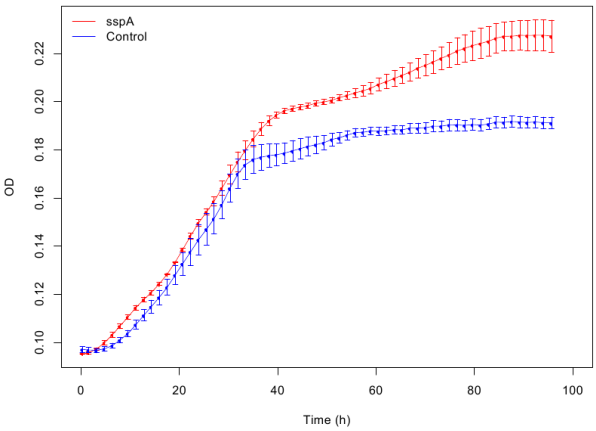


Figure S4. Growth curves of *E. coli* cells overexpressing ORFs that enable or improve growth in novel environments.

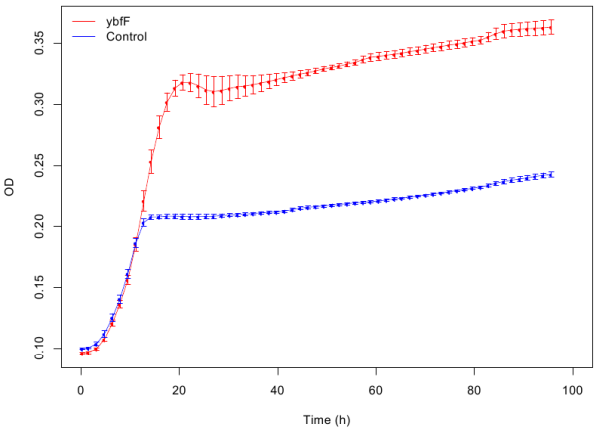
Growth experiments confirmed the beneficial growth effect of ORF overexpression in 18 cases (panels A-R). Red curves show the growth of the overexpressing cells, while the blue one is the negative control (cells harboring the empty plasmid). Each curve represents the average of three biologically independent replicates and their standard error. Detailed description of the growth conditions and the data analysis are presented in SI Materials and Methods. We note that panels B, O, P, Q and R are also presented in the main text (Figure 4B).


A)
N−Acetyl−D−Galactosamine − ygeN
ygeN
Control
OD
Time (h)
B)
D−2−Deoxyribose − rbsK
rbsK
Control
OD
Time (h)
C)
D−Raffinose − ycgZ
ycgZ
Control
OD
Time (h)
D)
Putrescine − mhpF
mhpF
Control
OD
Time (h)
E)
m−Tartrate − dmlA
dmlA
Control
OD
Time (h)
F)
beta−Methyl−D−Glucoside − bglB
bglB
Control
OD
Time (h)

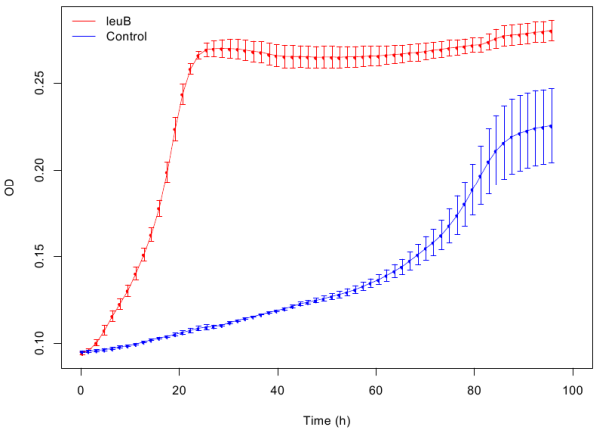
G)


Glyoxylic Acid − sspA
sspA
Control
OD
Time (h)

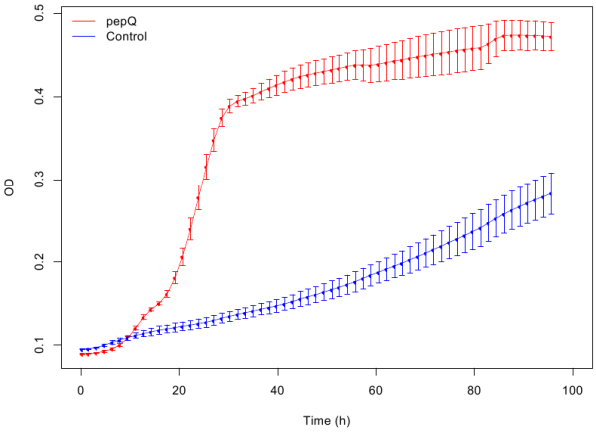
H)


Monomethyl Succinate − ybfF
ybfF
Control
OD
Time (h)

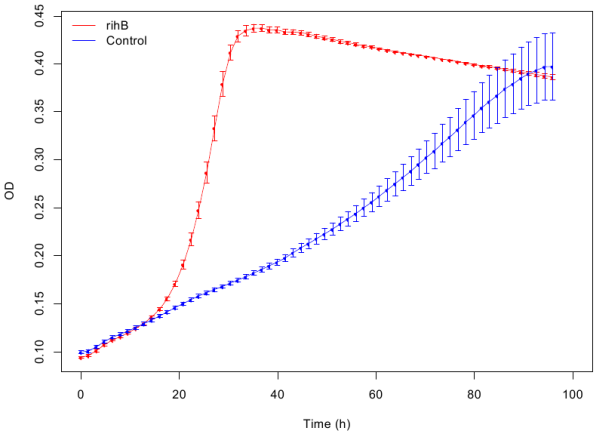
I)


D−Malate − leuB
leuB
Control
OD
Time (h)

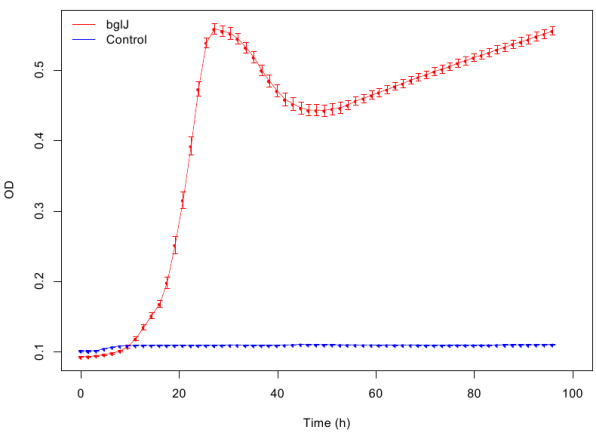
J)


Gly−Pro − pepQ
pepQ
Control
OD
Time (h)

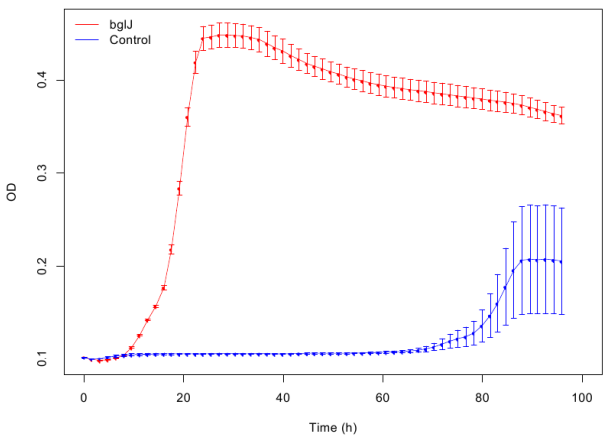
K)


Uridine − rihB
rihB
Control
OD
Time (h)

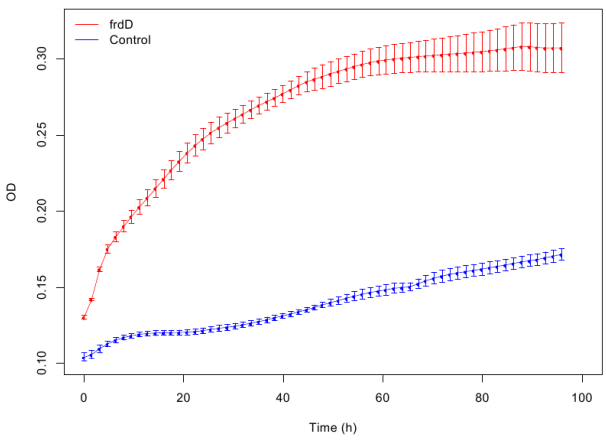
L)


Arbutin − bglJ
bglJ
Control
OD
Time (h)

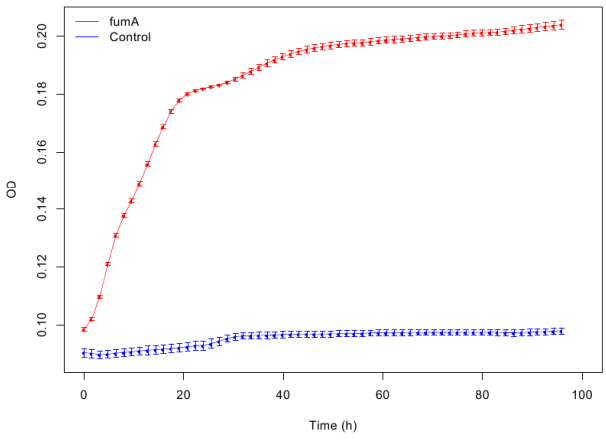
M)


Salicin − bglJ
bglJ
Control
OD
Time (h)

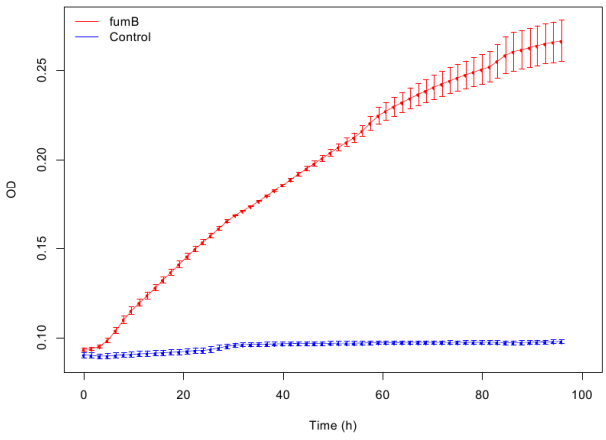
N)


L−Alaninamide − frdD
frdD
Control
OD
Time (h)

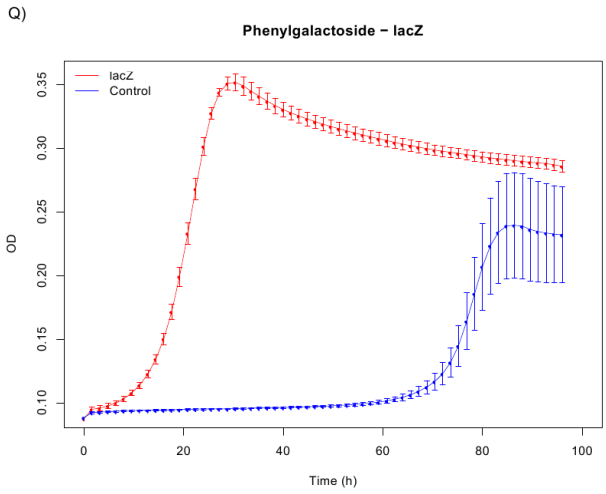
O)


D−Tartrate − fumA
fumA
Control
OD
Time (h)

P)


D−Tartrate − fumB
fumB
Control
OD
Time (h)

Q)


Phenylgalactoside − lacZ
lacZ
Control
OD
Time (h)

R)

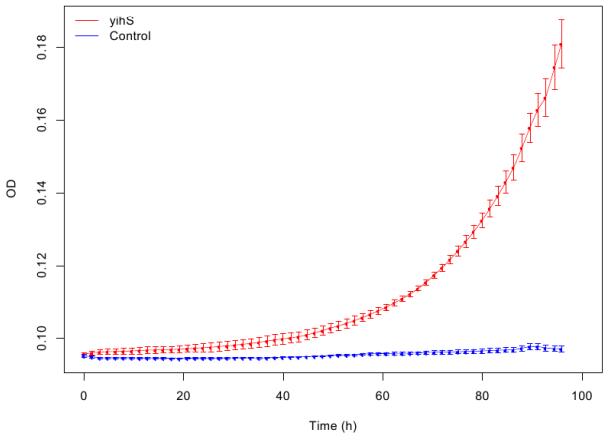

D−Lyxose − yihS
yihS
Control
OD
Time (h)

Table S1. List of published catalytic efficiencies (k_{cat}/K_m) for native and underground reactions of the same enzymes.

In cases of multiple substrates the median k_{cat}/K_m values are shown. Data were retrieved from the BRENDA database and associated literature. A Wilcoxon matched-pairs signed rank test demonstrates that underground reactions are catalyzed with significantly lower efficiency (k_{cat}/K_m) than native reactions of the same enzymes ($P<0.001$).

EC Number	Native Substrate(s) k_{cat}/K_m ($s^{-1}mM^{-1}$)	Underground Substrate(s) k_{cat}/K_m ($s^{-1}mM^{-1}$)
1.1.5.2	2344.16	10.47
1.8.1.2	3680.98	37.98
1.13.11.16	1115.38	21.39
2.6.1.16	72	3.4
2.7.1.15	309.68	0.03
2.7.1.16	430.85	17.91
2.7.1.17	11618.64	1.87
2.7.1.39	5.2	0.1
2.7.7.58	2037.04	9.53
3.1.3.10	262.92	0.23
4.2.1.2	4350	7.5
5.1.3.2	2941.18	0.0014
6.1.1.15	280	0.01

Table S2. Evaluation of *in silico* predicted growth advantages conferred by underground reactions

A) Agreement between growth predictions of the extended metabolic network model in specific environments and the *in vivo* overexpression screen are summarized in a contingency table. We only considered enzyme-encoding ORFs which have underground reactions associated in the reconstruction and are also present in the ASKA library (112 ORFs). Carbon sources not present in the extended network were excluded from the analysis, leaving 105 conditions. We note that some of the carbon sources presented in Figure 4B were therefore excluded from this comparison. Thus, a total of 112 * 105 ORF – carbon source pairs were evaluated here. Statistical significance of the overlap between *in silico* and experimental results was assessed by Fisher's exact test ($P<10^{-13}$). All *in silico* predicted and experimentally evaluated ORF – carbon source pairs are listed in Table S2B (see below). The 5 true positive cases in the upper left cell of the table correspond to the 4 successfully predicted carbon sources and 5 ORFs listed in Figure 4A. The one false negative case in the lower left cell is the *mhpF* – putrescine pair in Figure 4A.

		Experimentally determined growth advantage	
		Yes	**No**
***In silico* predicted growth advantage**	**Yes**	5	11
	No	1	11743

B) List of ORF – carbon source pairs that were *in silico* predicted to confer growth advantage via underground reactions and could be evaluated experimentally. (++) indicates that the underground activity of the enzyme is essential for growth *in silico*, while (+) indicates that the addition of the underground activity to the native network increases growth *in silico*. (●) and (○) indicate experimentally determined growth advantage and lack of growth advantage, respectively. Asterisk denotes cases which were predicted *in silico* as novel environments where an underground activity is essential for growth (light green squares in Figure 3B), but were not detected in the genome-wide screen. These ORF – carbon source pairs were nevertheless subjected to the verification assay as described in SI Materials and Methods.

Carbon source	**Gene name**	***In silico* growth advantage**	**Experimentally confirmed**
Phenylgalactoside	*lacZ*	++	●
D-Lyxose	*yihS*	++	●
D-Tartrate	*fumA*	++	●
D-Tartrate*	*fumB*	++	●
D-2-Deoxyribose	*rbsK*	++	●
D-2-Deoxyribose*	*deoC*	++	○
D-Tartrate*	*fumC*	++	○
L-Glyceraldehyde*	*fucO*	++	○
D-Arabinose*	*rbsK*	++	○
L-Sorbose*	*rhaD*	++	○
Ethylene glycol*	*fucO*	++	○
L-Glyceraldehyde	*glpK*	++	○
L-Arginine	*nadE*	+	○
Glycolate	*dxs*	+	○
L-Proline	*adhE*	+	○
L-Proline	*mhpF*	+	○

Table S3. Evaluation of growth advantages predicted by an alternative reconstruction based on a more recent version of the *E. coli* native network.

Here we ask whether incorporating underground reactions into a more recent version of the *E. coli* native metabolic network (iJO1366, ref. (26)) has an effect on predicting metabolic novelties across carbon sources. In brief, we integrated our list of underground reactions into iJO1366 as a base network. Next, we generated *in silico* predictions across those carbon sources that were tested in our genome-wide overexpression screen and identified conditions where the iJO1366 network extended with underground reactions showed at least 5% increased biomass production compared to the native iJO1366 network. Comparison of these *in silico* predicted gene – carbon source pairs with experimentally determined ones revealed a highly significant overlap ($P<10^{-7}$, Fisher's exact test). The contingency table of this comparison is shown below (this is analogous to Table S2A). We note that the underground reconstruction based on iJO1366 has somewhat lower prediction accuracy (e.g. 3 true positives instead of 5). Discrepancies between the two models can be largely attributed to the fact that the native iJO1366 network is already capable of utilizing D-tartrate and this growth phenotype relies on the very same reaction that we identified as underground (D-tartrate dehydratase associated with FumB). Published enzyme kinetic data clearly supports the underground nature of this reaction: the catalytic efficiencies of both FumA and FumB are 3 orders of magnitude lower for D-tartrate than for their native substrates (27). Thus, in the case of D-tartrate utilization, innovation is correctly predicted by our original iRN1260u reconstruction (through both *fumA* and *fumB*), but falsely predicted by the extended network built from iJO1366. We also found an opposite example: absence of growth on L-arginine through overexpressing *nadE* is correctly predicted by the extended network based on iJO1366, but falsely by our original model.

		Experimentally determined growth advantage	
		Yes	**No**
***In silico* predicted growth advantage**	**Yes**	3	10
	No	3	11429

Other Supporting Information Files

Dataset S1 (XLSX) Underground metabolic reconstruction of *E. coli* (iRN1260u).

Dataset S2 (XLSX) List of nutrient environments defined for the *in silico* growth prediction analysis.

Dataset S3 (XLSX) Summary of the results of the *in vivo* genome-wide screen.

Dataset S4 (XLSX) Information on known or putative underground activities which allow / improve growth on specific carbon sources when amplified.

Dataset S5 (XLSX) Set of native – underground reaction pairs derived from a systematic study of the haloacid dehalogenase-like phosphatase family of *E. coli* (Kuznetsova et al., 2006).

Dataset S6 (XML) The underground metabolic reconstruction (iRN1260u) of *E. coli* as provided in a Systems Biology Markup Language (SBML) file.

References

1. Taniguchi Y, *et al.* (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329(5991):533-538.
2. Patrick WM & Matsumura I (2008) A study in molecular contingency: glutamine phosphoribosylpyrophosphate amidotransferase is a promiscuous and evolvable phosphoribosylanthranilate isomerase. *J Mol Biol* 377(2):323-336.
3. Griffith KL & Wolf RE, Jr. (2002) Measuring beta-galactosidase activity in bacteria: cell growth, permeabilization, and enzyme assays in 96-well arrays. *Biochem Biophys Res Commun* 290(1):397-402.
4. Feist AM, *et al.* (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.
5. Kuznetsova E, *et al.* (2006) Genome-wide analysis of substrate specificities of the Escherichia coli haloacid dehalogenase-like phosphatase family. *J Biol Chem* 281(47):36149-36161.
6. Schellenberger J, Park JO, Conrad TM, & Palsson BO (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11:213.
7. Guo AC, *et al.* (2013) ECMDB: the E. coli Metabolome Database. *Nucleic Acids Res* 41(Database issue):D625-630.
8. Ishii N, *et al.* (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science* 316(5824):593-597.

9. van der Werf MJ, Overkamp KM, Muilwijk B, Coulier L, & Hankemeier T (2007) Microbial metabolomics: toward a platform with full metabolome coverage. *Anal Biochem* 370(1):17-25.

10. Csardi G & Nepusz T (2006) The igraph software package for complex network research. *InterJournal* Complex Systems.

11. Leach AR & Gillet VJ (2003) *An introduction to chemoinformatics* (Springer).

12. Burgard AP, Nikolaev EV, Schilling CH, & Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* 14(2):301-312.

13. Price ND, Reed JL, & Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2(11):886-897.

14. Schuster S, Dandekar T, & Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* 17(2):53-60.

15. Klamt S & Stelling J (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep* 29(1-2):233-236.

16. Kaleta C, de Figueiredo LF, Behre J, & Schuster S eds (2009) *EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks* (Gesellschaft für Informatik, Bonn), Vol 157, pp 179-189.

17. Planson AG, Carbonell P, Paillard E, Pollet N, & Faulon JL (2012) Compound toxicity screening and structure-activity relationship modeling in Escherichia coli. *Biotechnol Bioeng* 109(3):846-850.

18. Wagner A, *et al.* (2013) Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious. *Proc Natl Acad Sci U S A.*

19. Gelius-Dietrich G, Amer Desouki A, Fritzemeier CJ, & Lercher MJ (2013) sybil inverted question mark Efficient constraint-based modelling in R. *BMC Syst Biol* 7(1):125.

20. Team RC (2007) R: A Language and Environment for Statistical ComputingVienna, Austria).

21. Deutscher D, Meilijson I, Kupiec M, & Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* 38(9):993-998.

22. Soo VW, Hanson-Manful P, & Patrick WM (2011) Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. *Proc Natl Acad Sci U S A* 108(4):1484-1489.

23. Kitagawa M, *et al.* (2005) Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res* 12(5):291-299.

24. Lazar V, *et al.* (2013) Bacterial evolution of antibiotic hypersensitivity. *Mol Syst Biol* 9:700.

25. Bochner BR, Gadzinski P, & Panomitros E (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* 11(7):1246-1255.

26. Orth JD, *et al.* (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol* 7:535.

27. van Vugt-Lussenburg BM, van der Weel L, Hagen WR, & Hagedoorn PL (2013) Biochemical similarities and differences between the catalytic [4Fe-4S] cluster containing fumarases FumA and FumB from Escherichia coli. *PLoS One* 8(2):e55549.