Supplementary Information

Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious

Allon Wagner[1], Raphy Zarecki[1], Leah Reshef[3], Camelia Gochev[3], Rotem Sorek[4], Uri Gophna[3,5], Eytan Ruppin[1,2]

[1] The Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel
[2] The Sackler School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel
[3] Department of Molecular Microbiology and Biotechnology, Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel
[4] Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel
[5] National Evolutionary Synthesis Center, Durham, North Carolina 27705, USA

Correspondence to: Eytan Ruppin, ruppin@post.tau.ac.il.

Table of Contents

Supplementary Notes

1. Growth experiments

Since Kitagawa *et al*. (1) have already reported which genes inhibit growth when over-expressed in a rich LB medium, we chose to study growth inhibition due to over-expression in a minimal medium. In addition, the chemical composition of a synthetic medium is well-defined (in contrast, for instance, to LB medium whose composition is not fully characterized), and can be therefore accurately simulated *in-silico*, which leads to better model predictions. We calculated EDGE scores for all metabolic genes in the iAF1260 model (2), and ranked them by their absolute values, which were taken as a correlate of the predictions' confidence. 26 high-ranking genes, 12 with negative EDGE scores and 14 with positive ones, were selected for the subsequent growth experiments (Supplementary Table S1, and see details below). As detailed in the main text, we chose to use the MG1655 wild-type strain. We removed the GFP tag sequence in order to minimize confounding effects.

We grew clones harboring the IPTG-inducible plasmids of choice in a minimal glucose-supplemented M9 medium, supplemented with either 0, 50, 150, 450, or 1000µM IPTG. Supplementation with 150µM IPTG or higher concentrations led to severe growth inhibition in almost all cases, and was therefore considered less informative. Each strain/medium experiment was carried in duplicate wells, and OD_{595} of both duplicates were averaged to produce a growth curve. The slope of the linear phase of the logarithmic growth curve was extracted, and for each strain harboring a plasmid of choice the ratio of the slope of the 50µM IPTG-supplemented curve to that of the no IPTG supplementation was computed. That ratio was considered the growth inhibition observed at that experiment. All the ratios corresponding to different repetitions of the same strain were averaged to produce the growth inhibition associated with a particular gene's over-expression.

The 26 genes that were selected for the growth experiment were selected as follows: The iAF1260 model had 656 genes which were not associated with a blocked reaction, and were assigned a non-zero EDGE score for aerobic growth on glucose-supplemented M9 minimal medium. We sought to test ~30 genes, and thus obtain a ~5% coverage of that set. The absolute value of the EDGE score was used as a proxy for the prediction's confidence. We considered positive (non-toxic) and negative (toxic) scores separately since the scales of their magnitude were generally not symmetric. Genes were selected from the most confident predictions, while keeping equal proportions of genes that had positive and negative EDGE scores. We did not include isozymes (two genes that are indistinguishable in their function, at least as far as the GSMM is concerned), and avoided selecting multiple genes that participated in one metabolic pathway in order to obtain a better coverage of the metabolic network. The genes *trpB*, *trpC*, *trpD*, *trpE* were an exception to this rule, and were all included because the ASKA study (1) had reported that their over-expression leads to medium or severe growth inhibition, while EDGE assigned them a high positive score. Several genes were subsequently dropped due to technical issues, such as extremely divergent results between technical or biological duplicates, resulting in the set of 26 genes described in this study.

2. Evaluating EDGE against experimental over-expression libraries

EDGE was evaluated by comparing its scores to two genome-scale over-expression libraries: the ASKA library for *Escherichia coli* (1) and the Yeast GST-Tagged collection for *Saccharomyces cerevisiae* (3).

ASKA (1) divided *E. coli* into 3 categories based on the growth inhibition they cause when over-expressed: severe, moderate or no growth inhibition. Inhibition was measured by growing patches of cells on LB agar medium with or without 1 mM IPTG. We considered a gene to be toxic according to the ASKA study if it belongs to the class of genes causing severe growth inhibition.

Sopko *et al*. (3) constructed an array of yeast strains, with each strain carrying a different ORF expressed from the inducible GAL1/10 promoter. They transferred the array to both galactose and glucose-containing media, and systematically searched for strains showing a galactose-specific slow-growth phenotype. Each gene was assigned a score between 1 (lethal) and 5 (WT growth) to denote the growth inhibition it had caused. We follow Sopko *et al*.'s classification and consider a gene as toxic when its score is less than 5. We verified that we obtain similar results when considering genes with a score higher than or equal to (a) 4, or (b) 3.5 as non-toxic.

3. Evaluating EDGE against PandaTox

The PandaTox dataset lists genes that are unclonable in *E. coli*, or have significantly reduced clone coverage when replicated in *E. coli,* due to the toxic effects they cause (4, 5). For the purpose of the current study, we considered only unclonable genes as “toxic genes” (similar results were obtained when considering the reduced-coverage genes as toxic as well).

We simulated the process of cloning a gene into *E. coli* using automatically-generated models, built by the Model SEED infrastructure (6). The Model SEED had metabolic models available for 349 microorganisms out of the 393 which PandaTox lists, and we excluded species whose SEED metabolic model contained less than 5 toxic genes according to PandaTox. The remaining 138 bacterial models served in our simulation as the donor organisms. The SEED model for *E. coli* K-12 (NCBI taxon: 83333) served as the recipient organism. Each gene within each donor was cloned *in-silico* into the recipient organism. This was done as follows:

For each reaction associated with the donor gene within the donor organism:

- If the same reaction occurs also within the recipient, a gene-to-protein association was created between the cloned gene and the reaction.
- If the reaction does not occur within the recipient, it was added to its metabolic network, along with any metabolites associated with the reaction that were not already present in the recipient's network. We also added transport reactions and exchange reactions for those metabolites, and supplemented the *in-silico* growth medium with them. This was done in order to ensure that the cloned reactions will not be blocked on account of the newly added metabolites.

The *in-silico* growth medium of the newly-generated model simulated a rich LB medium, (as defined by Henry *et al.* (7)) in order to mimic experimental conditions. It was supplemented with newly-added metabolites as described above. We then computed the EDGE scores for the cloned gene within the context of the recipient organism.

We used two measures as correlates of promoter similarity between the source organism and the recipient *E. coli*:

1. Most recent common ancestor node in the NCBI taxonomy tree (http://www.ncbi.nlm.nih.gov/taxonomy, accessed July 2012), which can be either (from the farthest to the closest): bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriaceae, Escherichia coli. The latter is for cases of gene transfer between different *E.coli* strains.
2. Phylogenetic distance, according to MicrobesOnline (8) , measured as the distance between the source organism and the standard *E. coli* sequencing strain (DH10B, taxon ID: 316385). 8 out of the 138 bacteria in the data set do not appear in that tree.

We note that we chose to use the SEED's *E. coli* model as the recipient organism, rather than the *E. coli* iAF1260 model that was employed in the rest of this study. The SEED metabolic models share a unified annotation system for genes, reactions, metabolites etc. This allowed us to simulate *in silico* gene transfer from the SEED models of various organisms into SEED's *E. coli* model in a straightforward way, according to the steps described above. Had we used iAF1260 for this test, a costly and error-prone translation of the annotation would have been required. Moreover, sometime there are minute differences between gene-protein associations, reaction details etc., between the two systems, which would have to be manually resolved. These reasons led to the preference of SEED's *E. coli* model for this test.

We further tested whether the small group of organisms, for which the EDGE-based classifier predictions seemed worse than a random classifier (at the bottom right of Figure 3), might have had an underlying characteristic. Such a characteristic might explain why EDGE seemed to predict toxic genes to be non-toxic and vice-versa in these organisms. Alternatively, it might be that there was no such characteristic, and that the AUC values obtained by EDGE represent the usual fluctuations of a random variable (the AUC in this case) around its expected value (0.5 in this case, because of features of the experimental data that concern species that are phylogenetically far from *E. coli*, as discussed in the main text). We concluded that the latter was true. First, we manually inspected the relevant data points and verified there was no conspicuous trait that separated them from data points that obtained an AUC which was a higher larger than 0.5. We then ran a computational test, which supported this observation, and is described next.

We computed for each of these organisms the AUC of a truly random classifier, but one which still maintained the ratios of positive and negative predictions as the EDGE-based classifier, by shuffling the EDGE scores of genes within each organism. This process was repeated N = 10000 times. This way, we obtained the distribution of the AUC random variable under the null hypothesis, and were able to judge the AUC of the EDGE-based classifier against it. If the EDGE-based predictor had not conformed to the null hypothesis, we would have concluded that these organisms represent cases in which the EDGE classifier performed significantly worse than random, and this in turn would have supported the hypothesis that there was an underlying biology behind these “worse than random” data points. However, this was not the case. Almost all of the EDGE AUCs were within one standard deviation of the mean and all of them were within 1.5 standard deviations from the mean. Supplementary Table S8 lists the mean and the standard deviation of the distribution of the random AUC values (i.e., AUCs under the null hypothesis) for each taxon (i.e., data point) of interest, as well as the AUC that the EDGE-based classifier had obtained for this taxon. The p-value of the EDGE-based predictor under the null hypothesis is also noted. We conclude that the EDGE-based classifier for these data points conforms to the null hypothesis of random prediction. This means that EDGE cannot be used to predict the outcomes of the experimental procedures in the case of organisms that are phylogenetically far from *E. coli* (due to the reasons discussed in the main text), but that there is no evidence of worse than random predictions that might point to some interesting biological feature of these data points.

4. Evaluating EDGE against transcriptomic data

The growth media of each of the following arrays was reconstructed *in-silico* (Supplementary Methods) and EDGE scores were computed accordingly. We tested the correspondence between gene expression and EDGE-predicted toxicity by testing whether genes that EDGE classified as toxic (i.e., had negative EDGE scores) were lowly-expressed compared to genes which were not predicted to be toxic (i.e., had positive EDGE scores). We report both the Wilcoxon rank sum test statistic, transformed into a standard normal variable, and the corresponding p-value (against a one-sided alternative) – see Figure 4b in the main text. Similar results were obtained in all cases when measuring the correspondence through the Spearman correlation coefficient between the gene expression and the EDGE score instead of through the Wilcoxon rank sum statistic.

In each of the following datasets, all the arrays that correspond to the same experimental conditions were averaged.

4.1. Escherichia coli

We analyzed the following gene expression data:

- *E. coli* K-12 MG1655 grown on 13 different media and in different experimental conditions (9, 10) (Supplementary Table 3A). Also available from: http://systemsbiology.ucsd.edu/InSilicoOrganisms/Ecoli/EcoliExpression2 (last accessed Feb. 2013). Only arrays that measured wild-type strains (in contrast to KO strains) were analyzed in this study.
- *E. coli* K-12 MG1655 grown in minimal MOPS medium supplemented with 6 carbon sources of varying quality (11) (Supplementary Table 3B). Publicly available from GEO (GDS 1099).
- *E. coli* K-12 MG1655 grown on LB medium (12) (Supplementary Table 3C). Publicly available from GEO (GSM511651).

4.2. Yeast

We analyzed the following gene expression data:

- *S. cerevisiae* grown on galactose-supplemented, defined minimal medium (13) (Supplementary Table 3D). Publicly available from GEO (GSE461). Only control arrays were used in this study (GSM7490, GSM7491, GSM7492).
- *S. cerevisiae* grown on rich YP medium supplemented with various carbon sources (14) (Supplementary Table 3E). Publicly available from GEO (GSE18). Arrays that measured stress conditions were not used in this study. Rather, only arrays that measured growth on various carbon sources were analyzed (GSM907, GSM990, GSM991, GSM997, GSM999, GSM1001)

4.3. Human

We analyzed the following gene expression data:

- Samples taken from 79 human tissues, 6 of which are cancerous and the rest healthy (15) (Supplementary Table 4). Publicly available from GEO (GSE1133). This dataset also contains mouse arrays, which were not used in the current study.
- Samples taken from the NCI-60 collection of cancer cell lines (16) (Supplementary Table 5). Publicly available from GEO (GSE5846).

4.4. Arabidopsis thaliana

We analyzed the following gene expression data:

- 79 arrays covering many developmental stages of the plant and diverse organs (17) (Supplementary Table 6). Publicly available from ArrayExpress

(E-TABM-17). The results presented in the main text were obtained by using the non-photosynthetic medium for all samples. Replacing that medium with the photosynthetic medium in the case of photosynthetic plant tissues gives similar results (73 of the 79 arrays have Wilcoxon rank sum $p < 0.05$, median $p < 1.07e-4$).

5. Interpretation of genetic reprogramming in cancer through EDGE

Four gene expression datasets were used in the course of the analysis, as detailed in the main text and in Supplementary Table S7. They were subjected to the same analysis procedures for gene expression data as described previously in Supplementary Notes, “Evaluating EDGE against transcriptomic data”. All datasets are publicly available from GEO. Their GEO identifiers and bibliographic references are:

1. GDS3592 – Bowen *et al*. (18)
2. GDS2342 – Diaz-Blanco *et al*. (19)
3. GDS3289 – Tomlins *et al*. (20)
4. GDS3716 – Graham *et al*. (21)

Note that in the case of GDS3716 the test group was actually not diagnosed with active cancer, but rather only with high risk to develop one.

6. Correspondence of EDGE scores and actual gene expression is indicative of cellular proliferative capabilities

As in Supplementary Notes, “Evaluating EDGE against transcriptomic data”, the magnitude of the correspondence was quantified through the Wilcoxon rank sum test statistic, transformed into a standard normal variable, for testing whether genes which EDGE predicted to be toxic (i.e., had negative EDGE scores) where down-regulated compared to genes which were predicted to be non-toxic (i.e., had positive EDGE scores). We ranked the samples in the aforementioned dataset of gene expression from 79 human tissues (15), 73 of which are healthy and 6 cancerous, according to that magnitude, which quantifies the correspondence of the actual, observed gene expression with the EDGE predictions. Almost identical results were obtained when quantifying the correspondence between EDGE score and gene expression through the Spearman correlation coefficient instead of the Wilcoxon rank sum test statistic.

Since cancerous tissues are expected to adhere better than normal tissues to the proliferative objective function, it was expected that they would have higher correspondence between EDGE scores and gene expression (quantified as described above). We tested whether this characteristic allows one to build a computational predictor that separates cancer samples from non-cancer samples by computing the AUC that such a predictor would have obtained. Remarkably, an AUC of 0.96 ($p < 1.64e-22$) was obtained.

Indeed, previous studies have already shown that machine learning methods can be successfully applied to classify cancer microarrays (22–25). However, the successful classification described here relies *on only one feature*, namely the correspondence between EDGE scores and gene expression. This demonstrates the power of EDGE to detect genes, whose expression might impede proliferation of human cells (and are thus suppressed in proliferative tissues).

7. Neutral genes

We conducted tests to make sure that the sets of genes that EDGE marks as neutral correspond to the intuitive interpretation of a neutral (i.e, zero) EDGE score. A gene is assigned a zero EDGE score when the cellular objective can be obtained both when forcing or eliminating its expression. Considering, for example, *E. coli*'s iAF1260 metabolic model (2) grown on glucose-supplemented M9 minimal medium in aerobic conditions, we see that the genes *grxA*, *grxC*, and *grxD* are among those that are assigned a neutral EDGE score. According to the metabolic reconstruction, these genes are responsible, among other functions, to the catalyzation of the reaction phosphoadenylyl sulfate reductase, which takes glutaredoxin as its substrate. The GSMM predicts that this reaction can carry a non-zero flux in an optimal (with respect to biomass production) flux distribution in the environment in question. On the other hand, optimal biomass production can also be obtained while suppressing this reaction completely, and in this case it is compensated by a similar reaction, phosphoadenylyl sulfate reductase, which takes thioredoxin as its substrate. Previously published experimental results are in line with these observations. On the one hand, these genes (*grxA*, *grxC*, and *grxD*) are known to be non-essential for growth on glucose-supplemented M9 minimal medium in the presence of oxygen, and their KO strains exhibit a growth rate which is close to the wild type (26, 27). On the other hand, these genes are generally highly-expressed when *E. coli* is grown in that environment nonetheless (9). In conclusion, the organism can achieve an optimal biomass production both when expressing these genes in the environment in question, and when suppressing their expression, and this fact is predicted by their neutral EDGE score for that environment.

Supplementary Methods

8. In-silico strains and media

8.1. Escherichia coli

This study used the iAF1260 metabolic model of Feist *et al.* (2) in all cases except for the results that concern the PandaTox dataset (see below). *In-silico* minimal media were based on the composition defined by Feist *et al.*, with the appropriate carbon source, nitrogen source and oxygen (when aerobic conditions are recreated). *In-silico* LB medium composition was adapted from Henry *et al.* (7). Following Feist *et al.*, whenever both glucose and oxygen are available to the model, we emulate the cellular regulatory response by shutting down 152 reactions listed in their paper.

In-silico predictions for the clonability of genes into an *E. coli* host were carried with the Model SEED's automatically-generated metabolic models (6) for the foreign organisms, and with the Model SEED's model for *E. coli* K-12 (NCBI taxon: 83333) as the recipient model. *In-silico* LB medium composition was again adapted from Henry *et al.*(7), with the addition of thiamine since SEED's *E. coli* K-12 model cannot produce biomass unless the growth medium is supplemented with thiamine.

8.2. Yeast

The metabolic model used for *S. cerevisiae* was iMM904, published by Mo *et al*. (28). *In-silico* composition of a synthetic minimal medium was adapted from Mo *et al.* as well. *In-silico* YP medium composition is given in Supplementary Table S9. Both media were supplemented with carbon sources according to the experimental conditions.

8.3. Human

Recon1 (29) was used as the metabolic model, with an *in-silico* medium that mimics RPMI-1640 (Supplementary Table S10). The biomass production pseudo-reaction for that model was taken from Folger *et al*. (30).

8.4. Arabidopsis thaliana

The AraGEM (31) model of the primary metabolic network in Arabidopsis was employed in this study. The medium definition is detailed in Supplementary Table S11.

9. Blocked reactions

Genome-scale metabolic models invariably contain blocked reactions. There are two types of blocked reactions (32):

> a) always (or unconditionally) blocked - reactions that cannot carry flux under steady-state conditions. These represent metabolic dead-ends and are generally associated with areas of the networks that are not thoroughly studied.
>
> b) conditionally blocked - reactions that can carry a steady-state flux under some conditions, but not in a particular growth medium, for instance because they require a substrate that cannot be produced from that growth medium).

Therefore, blocked reactions represent both knowledge gaps, and parts of the metabolic network that are not expected to play part under steady-state conditions in a particular environment. In this study, prior to each computational test, we found the set of blocked reactions (both unconditionally and conditionally blocked) by detecting reactions that were constrained to carry zero flux in the growth medium relevant to that test (33). Genes associated with these reactions were excluded from further analysis that pertained to the computational test in question.

10. The EDGE algorithm

For the sake of clarity, we repeat here the full formulation of the EDGE algorithm from the *Materials and Methods* section in the main text (where there are no new lines between the constraints of the mathematical programs due to space constraints).

Given a gene, g, let $T_g = \{v_{i_1}, \dots, v_{i_K}\}$ denote the set of reactions in the network that are associated with g. We define:

$$EDGE(g) \coloneqq \min_{j \in \{1,\dots,K\}} f^{UP}(T_g, j) - f^{KO}(T_g)$$

f^{KO} is the optimal objective subject to silencing g. The minuend $\min_{j \in \{1,\dots,K\}} f^{UP}(T_g, j)$ is the optimal objective subject to the most restrictive bottleneck. The difference can be further divided by epsilon for the purpose of normalization, but it was unnecessary in our study because all comparisons reported always involve the same epsilon. We note that this subtraction is prone to numerical “loss of significance” errors; for that reason, we round the result to 10 decimal places.

Let $S \in \mathbb{R}^{m \times n}$ be the stoichiometric matrix of a metabolic network (where m and n are the number of metabolites and reactions in the network, respectively). Let $\alpha, \beta \in \mathbb{R}^n$ denote lower and upper bounds, respectively, for reaction fluxes stemming from nutrient availability, thermodynamic constraints etc. α_i, β_i can also be set to $\pm\infty$ for some i’s to denote “no bound”. Let f denote a linear cellular objective function to maximize subject to the environmental constraints. In our study, f was always the biomass production.

Define $f^{KO}(T_g)$ to be the optimal objective value of the following linear program:

$$f^{KO}(T_g) \coloneqq \underset{v \in \mathbb{R}^n}{max} f(v)$$

$subject\ to$:

1) $S \cdot v = 0$
2) $\forall i = 1, \dots, n. \alpha_i \leq v_i \leq \beta_i$
3) $\forall v_i \in T_g : v_i = 0$

Define $f^{UP}(T_g, j)$ to be the optimal objective value of the following mixed-integer linear program:

$$f^{UP}(T_g, j) \coloneqq \underset{v \in \mathbb{R}^n, a \in \{0,1\}^K}{max} f(v)$$

$subject\ to$:

1) $S \cdot v = 0$
2) $\forall i = 1, \dots, n. \alpha_i \leq v_i \leq \beta_i$
3) $\forall v_{i_k} \in T_g \setminus \left\{v_{i_j}\right\}$:

$$a_k = 1 \longrightarrow v_{i_k} \geq \varepsilon$$

$$a_k = 0 \longrightarrow v_{i_k} \leq -\varepsilon$$

4) $a_j = 1 \longrightarrow v_{i_j} = \varepsilon$
5) $a_j = 0 \longrightarrow v_{i_j} = -\varepsilon$

where ε is an infinitesimal constant chosen to reflect the smallest non-negligible flux possible. However, ε cannot be arbitrarily small due to the finite precision of the floating-point representation. a_k are binary variables whose purpose is to ensure that the reversible reactions associated with g carry a flux in either direction. They participate in logical constraints that can be transformed into regular integer linear constraints via routine transformations (34). Commercial solvers are sometimes able to branch explicitly on these constraints. We note that we described the algorithm as adding an $\mathrm{a_k}$ variable for each reaction for the sake of simplicity. In practice, it is unnecessary to introduce an a_k variable for irreversible reactions because the respective constraints for those can be simply added as linear constraints. Further implementation considerations are discussed below.

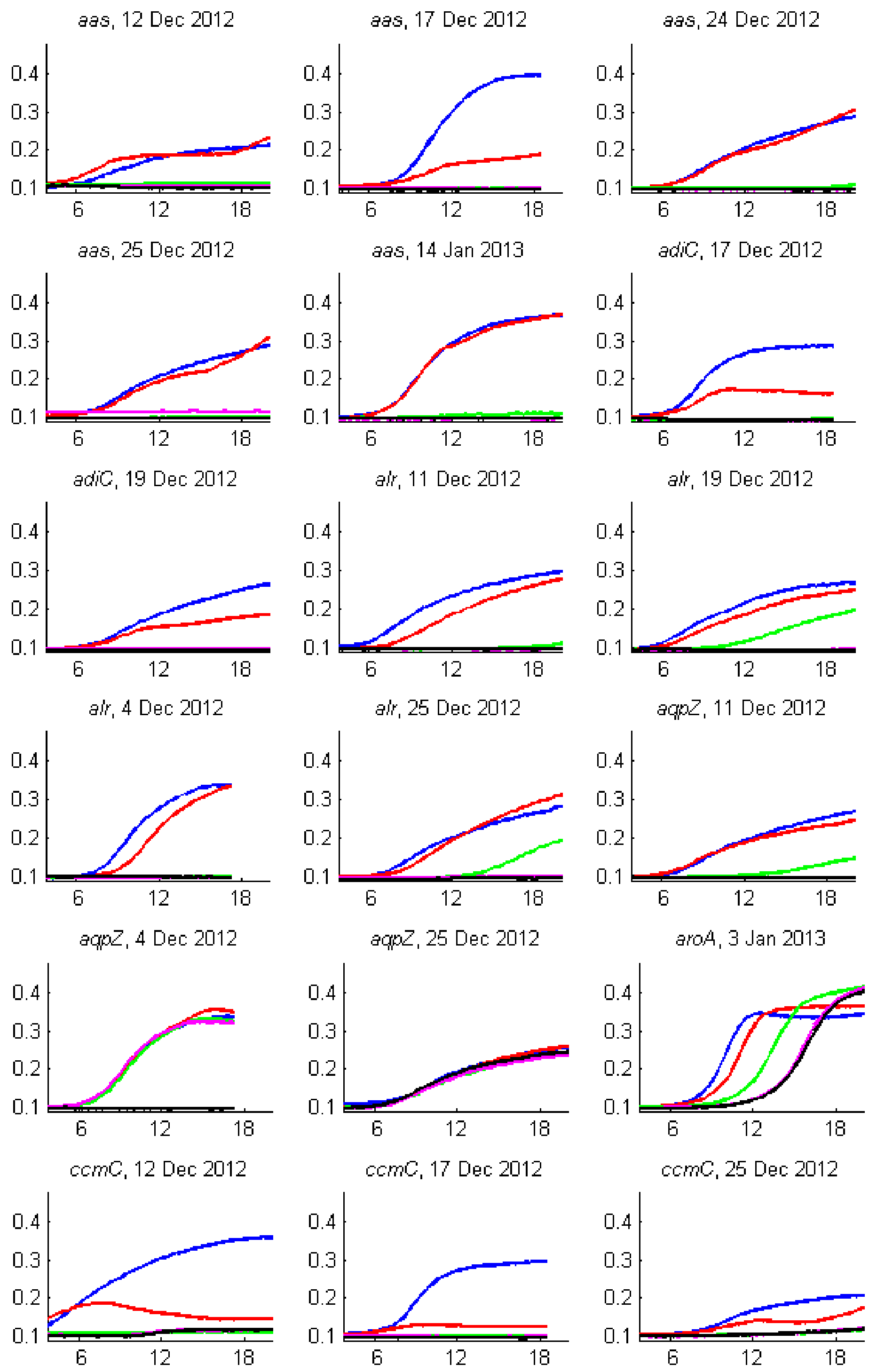
Genes were classified as toxic if they had a negative EDGE score and as non-toxic if they had a positive EDGE score. For the purpose of conducting growth experiments, we used the absolute value of the score as the prediction's confidence, with higher absolute values denoting the more confident predictions. Genes that were associated with a blocked reaction were excluded from the analysis.

11. EDGE implementation considerations

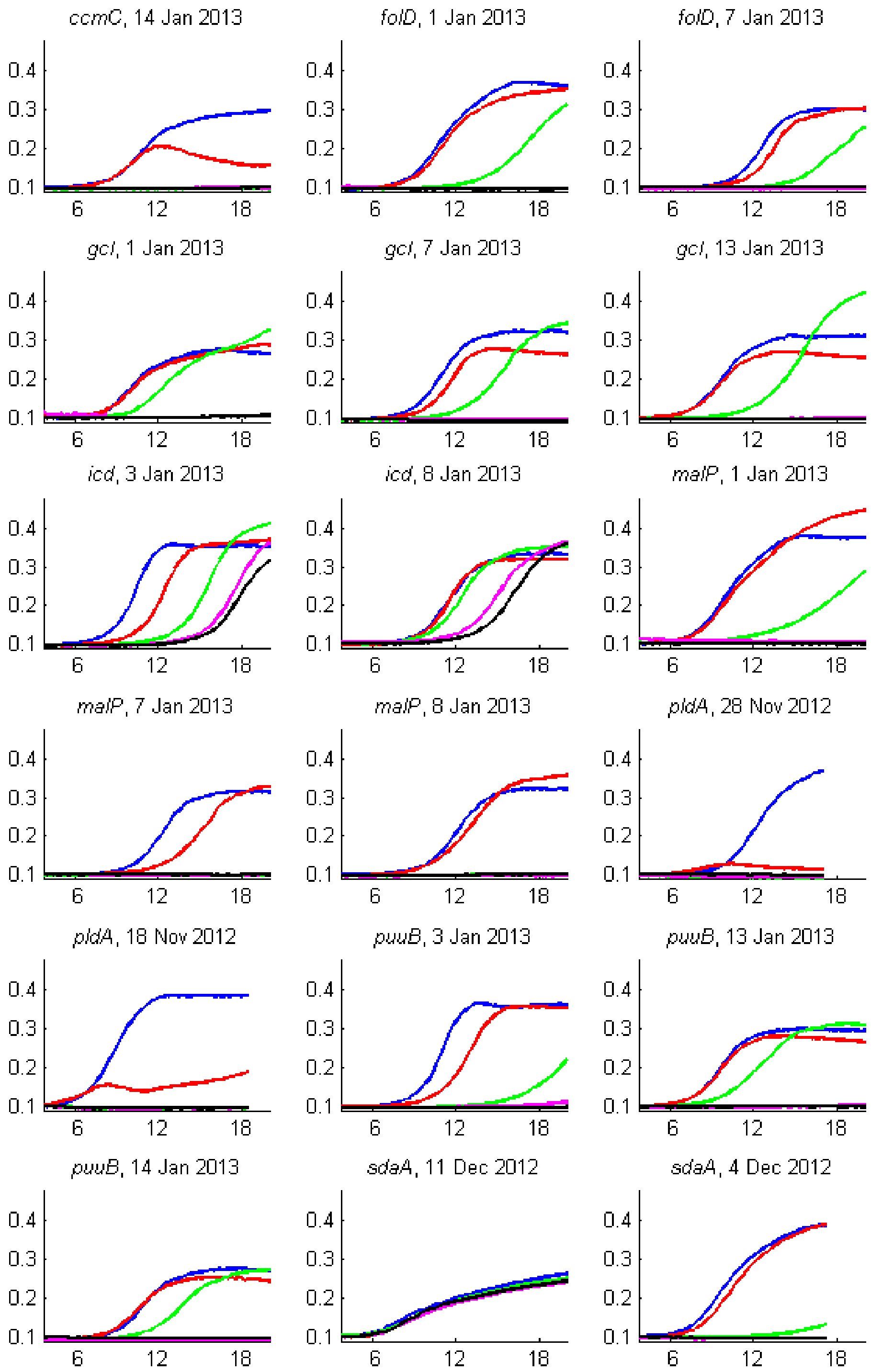
ε is an infinitesimal constant chosen to reflect the smallest non-negligible flux possible. ε cannot be arbitrarily small due to the finite precision of the floating-point representation, which cannot handle systems with constants spanning too many orders of magnitude. Thus, ε was determined according to the numerical properties of the constants of the model. ε was taken to be 10^{-5} in all manually-curated models (*E. coli* iAF1260, yeast, human, Arabidopsis), and 0.1 in the SEED automatically-constructed models.

a_k are binary variables whose purpose is to ensure that the reversible reactions associated with g carry a flux in either direction. They participate in logical constraints that can be transformed into regular integer-linear constraints via routine transformations (34). As a matter of fact, some commercial solvers, such as CPLEX (35) that we used, are sometimes able to branch explicitly on these constraints. This typically results in better numerical stability than the conventional transformations that tend to introduce trickle flows (35). We note that we described the algorithm as adding an $\mathrm{a_k}$ variable for each reaction for the sake of simplicity. In practice, it is unnecessary to introduce a a_k for irreversible reactions, since the respective constraints for those can be simply added as linear constraints, constraining the reactions to carry at least ε flux in the proper direction. In addition, when the number of reactions controlled by a gene, according to the gene-to-protein associations embedded in the model, is small enough, one can combinatorically iterate through all possible a_k combinations and solve an LP instead of solving a MILP.

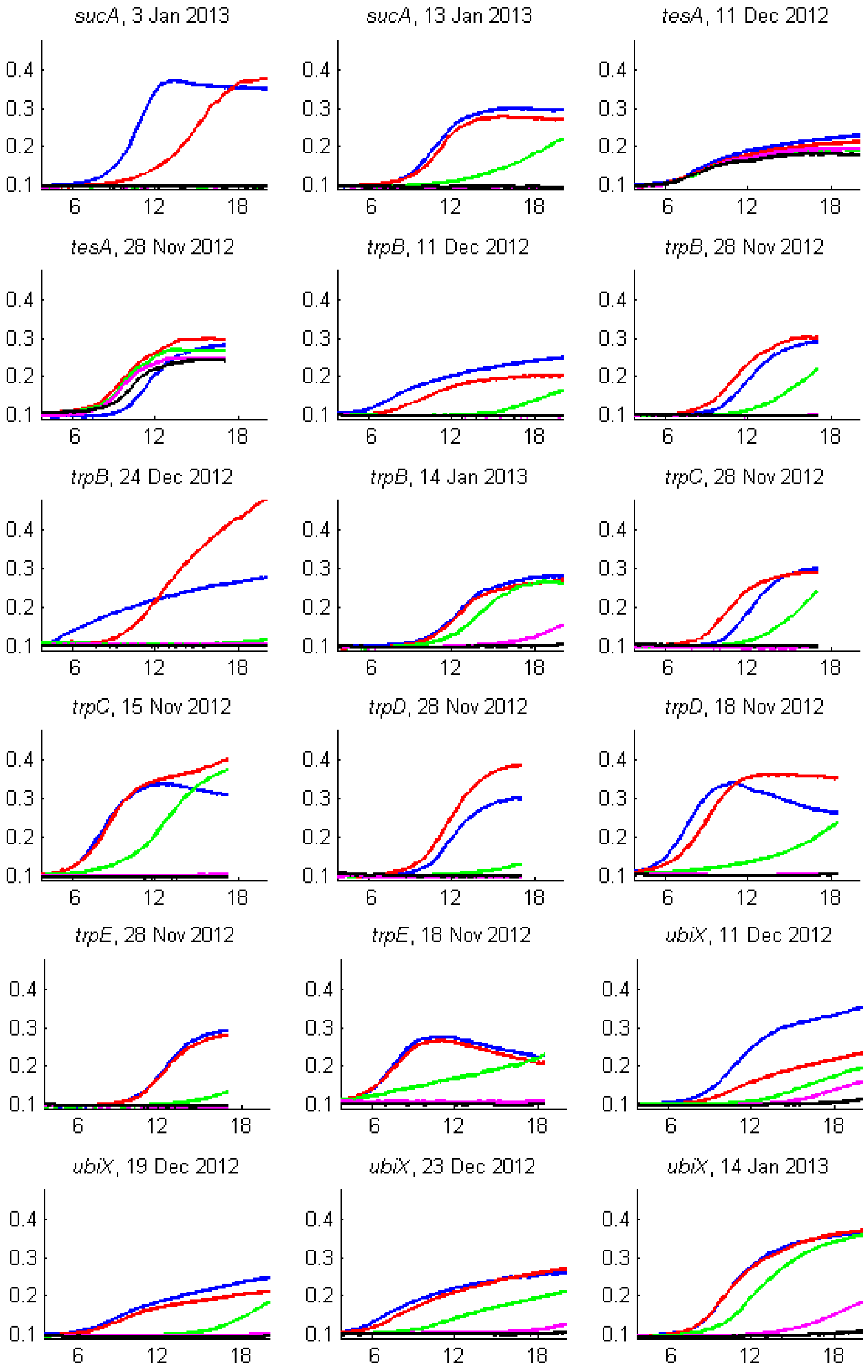
Supplementary Figures



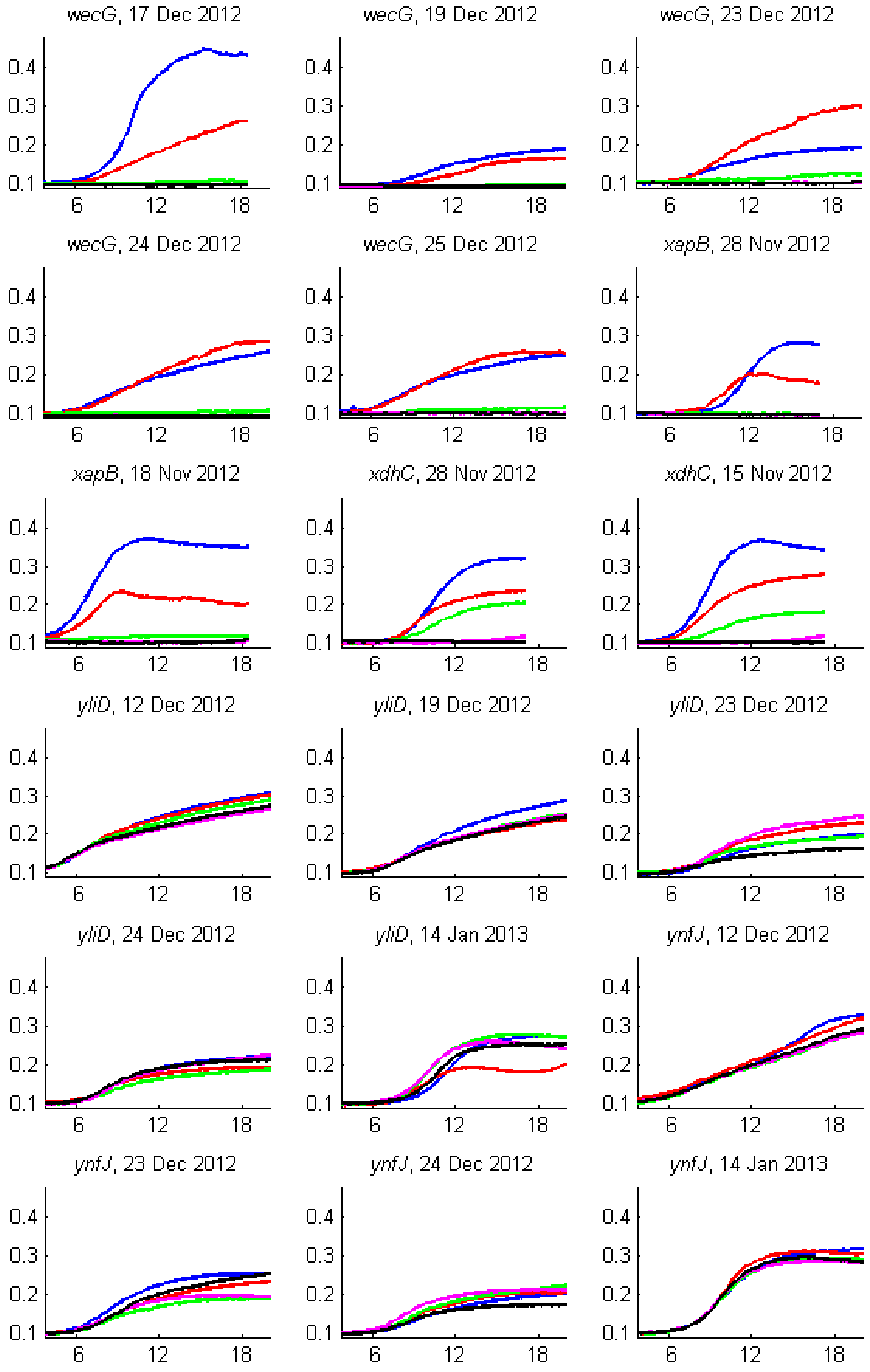
Supplementary Figure S1a. (see caption below)



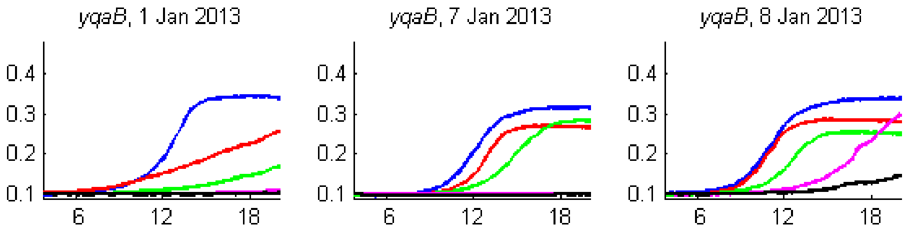
Supplementary Figure S1b. (see caption below)



Supplementary Figure S1c. (see caption below)

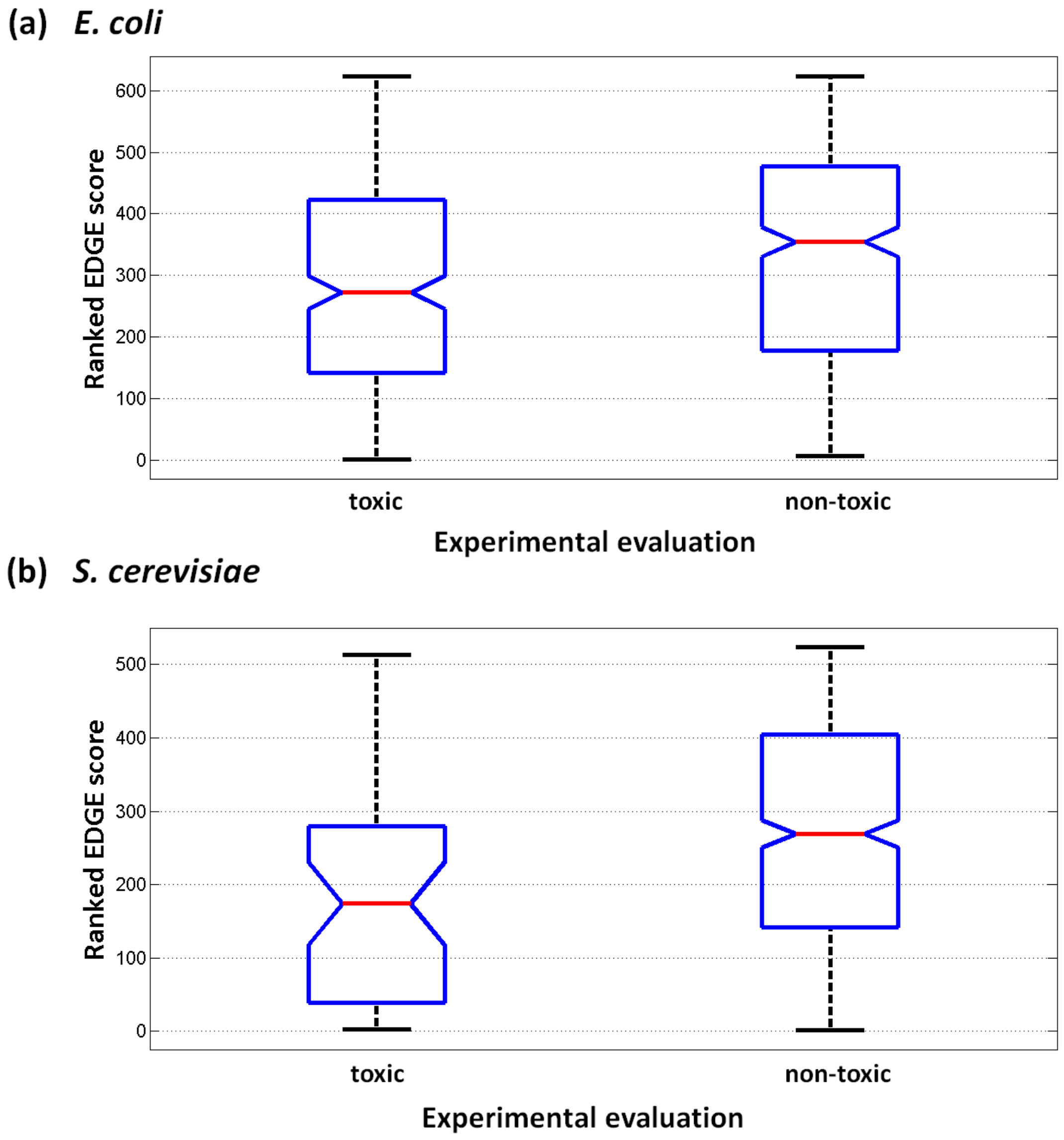


Supplementary Figure S1d. (see caption below)

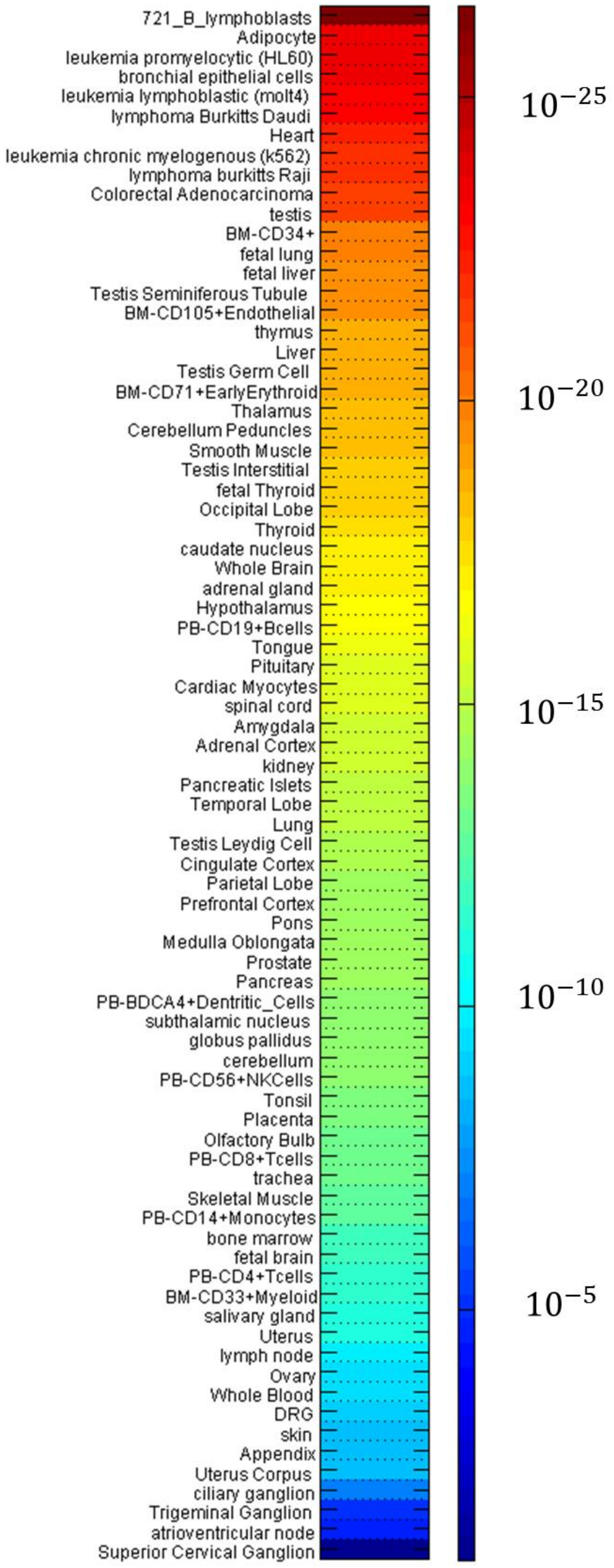


Supplementary Figure S1e. (see caption below)

Supplementary Figure S1. Growth curves for the *E. coli* growth experiments described in the main text. 26 *E. coli* metabolic genes which had a highly-confident EDGE score (i.e., high absolute values) were selected for an over-expression experiment. Plasmids (1) carrying IPTG-inducible constructs of these were transferred into a WT MG1655 *E. coli* strain, and clones were grown in a minimal M9 medium supplemented with glucose and 0, 50, 150, 450, 1000µM IPTG (blue, red, green, magenta, black curves, respectively). See the Methods section for complete details. X-axis denotes the elapsed time in hours, y-axis denotes OD_{595} values. Each subplot corresponds to one growth experiment of a certain strain (the entire growth experiment was repeated 2 to 5 times for each strain and IPTG concentration). Each curve represents an average of 2 duplicate wells.



Supplementary Figure S2. EDGE assigns significantly lower scores to genes that had previously been classified as toxic when over-expressed (i.e., their over-expression results in severe growth inhibition) in (a) *E. coli* (1), and in (b) *S. cerevisiae* (3).



Supplementary Figure S3. See figure's legend on the next page.

Supplementary Figure S3. EDGE-based analysis of the proliferative signature in the transcriptome of healthy and cancerous human tissues. EDGE relies on a hypothesized cellular objective. Since in this study we took proliferation as the cellular objective, it is not surprising to see that the magnitude by which EDGE-predicted toxic genes are down-regulated compared to EDGE-predicted non-toxic genes is highest for proliferative tissues (e.g., cancer cell lines, lymphoblasts) and lowest for non-proliferative ones (e.g., ganglia). Thus, the magnitude of the aforementioned effect is indicative of the degree to which the cellular program is geared towards proliferation. The magnitude of the effect was quantified by the test-statistic of a Wilcoxon rank sum test, as explained in the caption for Figure 4 in the main text and in the Supplementary Notes – "Correspondence of EDGE scores and actual gene expression is indicative of cellular proliferative capabilities". The corresponding p-values for these test-statistics (against a one-sided alternative) are color-coded according to the bar on the right. Gene expression data for 79 human samples was taken from Su *et al.* (15). All samples achieve p < 1.4e-5 except for the bottom one (superior cervical ganglion with p = 0.141, see Main Text). See also Supplementary Table S4.

Supplementary References

1. Kitagawa M et al. (2006) 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* 12:291–299. Available at: http://dnaresearch.oxfordjournals.org/content/12/5/291.abstract.

2. 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. Available at: http://dx.doi.org/10.1038/msb4100155.

3. Sopko R et al. (2006) Mapping Pathways and Phenotypes by Systematic Gene Overexpression. *Mol Cell* 21:319–330. Available at: http://www.sciencedirect.com/science/article/pii/S1097276505018538.

4. Sorek R et al. (2007) Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer . *Sci* 318 :1449–1452. Available at: http://www.sciencemag.org/content/318/5855/1449.abstract.

5. Kimelman A et al. (2012) A vast collection of microbial genes that are toxic to bacteria . *Genome Res* 22 :802–809. Available at: http://genome.cshlp.org/content/22/4/802.abstract.

6. Henry CS et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech* 28:977–982. Available at: http://dx.doi.org/10.1038/nbt.1672.

7. Henry CS, Zinner JF, Cohoon MP, Stevens RL (2009) iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. *Genome Biol* 10.

8. Dehal PS et al. (2009) MicrobesOnline: an integrated portal for comparative and functional genomics . *Nucleic Acids Res* . Available at: http://nar.oxfordjournals.org/content/early/2009/11/11/nar.gkp919.abstract.

9. 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* 191:3437–3444. Available at: http://jb.asm.org/content/191/11/3437.short.

10. Lewis NE et al. (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol* 6. Available at: http://dx.doi.org/10.1038/msb.2010.47.

11. Liu M et al. (2005) Global Transcriptional Programs Reveal a Carbon Source Foraging Strategy by Escherichia coli . *J Biol Chem* 280 :15921–15927. Available at: http://www.jbc.org/content/280/16/15921.abstract.

12. Habdas BJ, Smart J, Kaper JB, Sperandio V (2010) The LysR-Type Transcriptional Regulator QseD Alters Type Three Secretion in Enterohemorrhagic Escherichia coli and Motility in K-12 Escherichia coli. *J Bacteriol* 192:3699–3712. Available at: http://jb.asm.org/content/192/14/3699.abstract.

13. Bro C et al. (2003) Transcriptional, Proteomic, and Metabolic Responses to Lithium in Galactose-grown Yeast Cells. *J Biol Chem* 278 :32141–32149. Available at: http://www.jbc.org/content/278/34/32141.abstract.

14. Gasch AP et al. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol Biol Cell* 11 :4241–4257. Available at: http://www.molbiolcell.org/content/11/12/4241.abstract.

15. Su AI et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101:6062–6067. Available at: http://www.pnas.org/content/101/16/6062.abstract.

16. Lee JK et al. (2007) A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc Natl Acad Sci* 104 :13086–13091. Available at: http://www.pnas.org/content/104/32/13086.abstract.

17. Schmid M et al. (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet* 37:501–506. Available at: http://dx.doi.org/10.1038/ng1543.

18. Bowen N et al. (2009) Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. *BMC Med Genomics* 2:71. Available at: http://www.biomedcentral.com/1755-8794/2/71.

19. Diaz-Blanco E et al. (2007) Molecular signature of CD34+ hematopoietic stem and progenitor cells of patients with CML in chronic phase. *Leuk* 21:494–504. Available at: 10.1038/sj.leu.2404549.

20. Tomlins SA et al. (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 39:41–51. Available at: http://dx.doi.org/10.1038/ng1935.

21. Graham K et al. (2010) Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *Br J Cancer* 102:1284–1293. Available at: 10.1038/sj.bjc.6605576.

22. Fishel I, Kaufman A, Ruppin E (2007) Meta-analysis of gene expression data: a predictor-based approach. *Bioinforma* 23 :1599–1606. Available at: http://bioinformatics.oxfordjournals.org/content/23/13/1599.abstract.

23. Khan J et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 7:673–679. Available at: http://dx.doi.org/10.1038/89044.

24. Wang Y et al. (2005) Gene selection from microarray data for cancer classification—a machine learning approach. *Comput Biol Chem* 29:37–46. Available at: http://www.sciencedirect.com/science/article/pii/S1476927104001082.

25. Sung J et al. (2013) Multi-study Integration of Brain Cancer Transcriptomes Reveals Organ-Level Molecular Signatures. *PLoS Comput Biol* 9:e1003148. Available at: http://dx.doi.org/10.1371/journal.pcbi.1003148.

26. Orth JD et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism - 2011. *Mol Syst Biol* 7. Available at: http://dx.doi.org/10.1038/msb.2011.65.

27. Baba T et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2. Available at: http://dx.doi.org/10.1038/msb4100050.

28. Mo M, Palsson B, Herrgard M (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* 3:37. Available at: http://www.biomedcentral.com/1752-0509/3/37.

29. Duarte NC et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data . *Proc Natl Acad Sci* 104 :1777–1782. Available at: http://www.pnas.org/content/104/6/1777.abstract.

30. Folger O et al. (2011) Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* 7. Available at: http://dx.doi.org/10.1038/msb.2011.35.

31. De Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis . *Plant Physiol* 152 :579–589. Available at: http://www.plantphysiol.org/content/152/2/579.abstract.

32. Burgard AP, Nikolaev E V, Schilling CH, Maranas CD (2004) Flux Coupling Analysis of Genome-Scale Metabolic Network Reconstructions. *Genome Res* 14 :301–312. Available at: http://genome.cshlp.org/content/14/2/301.abstract.

33. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 5:264–276. Available at: http://www.sciencedirect.com/science/article/pii/S1096717603000582.

34. Bisschop J (2011) *AIMMS - Optimization Modeling* (Paragon Decision Technology).

35. International Business Machines Corporation [IBM] (2009) *IBM ILOG CPLEX V12.1: User's Manual for CPLEX* (International Business Machines).