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

Allon Wagner[a], Raphy Zarecki[a], Leah Reshef[b], Camelia Gochev[b], Rotem Sorek[c], Uri Gophna[b,d], and Eytan Ruppin[a,e,1]

[a]The Blavatnik School of Computer Science, [e]Sackler School of Medicine, and [b]Department of Molecular Microbiology and Biotechnology, Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel; [c]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel; and [d]National Evolutionary Synthesis Center, Durham, NC 27705

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved October 11, 2013 (received for review June 29, 2013)

Gene suppression and overexpression are both fundamental tools in linking genotype to phenotype in model organisms. Computational methods have proven invaluable in studying and predicting the deleterious effects of gene deletions, and yet parallel computational methods for overexpression are still lacking. Here, we present Expression-Dependent Gene Effects (EDGE), an in silico method that can predict the deleterious effects resulting from overexpression of either native or foreign metabolic genes. We first test and validate EDGE's predictive power in bacteria through a combination of small-scale growth experiments that we performed and analysis of extant large-scale datasets. Second, a broad cross-species analysis, ranging from microorganisms to multiple plant and human tissues, shows that genes that EDGE predicts to be deleterious when overexpressed are indeed typically down-regulated. This reflects a universal selection force keeping the expression of potentially deleterious genes in check. Third, EDGE-based analysis shows that cancer genetic reprogramming specifically suppresses genes whose overexpression impedes proliferation. The magnitude of this suppression is large enough to enable an almost perfect distinction between normal and cancerous tissues based solely on EDGE results. We expect EDGE to advance our understanding of human pathologies associated with up-regulation of particular transcripts and to facilitate the utilization of gene overexpression in metabolic engineering.

systems metabolic engineering | metabolic modeling | constraint-based modeling | flux balance analysis | horizontal gene transfer

Deducing phenotype from genotype is a fundamental goal of modern biology. Traditionally, experimentalists delete or suppress genes to annotate them and to detect phenotypes of interest (1), and such studies are routinely conducted in model organisms. However, it has long been recognized that gene overexpression is a complementary tool for linking genotype to phenotype and usually provides phenotypes that are different from those observed in loss-of-function studies (2).

Beyond basic science, complex human disease is often associated with abnormally up-regulated transcripts (3–5). Studies in murines and in *Drosophila* have shown that gene overexpression can induce disease on the one hand (6–8) and serve therapeutic purposes on the other (9–11). In addition, phenotypes arising due to overexpression are of prime interest in metabolic engineering, where selected native or heterologous genes are overexpressed to produce novel metabolic pathways in industrial microorganisms (12–16). Metabolic engineers particularly seek to foresee cases in which the up-regulation of a specific gene results in severely reduced fitness (17, 18). Such genes are often referred to as "toxic genes" (19–22).

Nonetheless, there have been only a handful of genome-wide studies addressing gene overexpression to date (2, 23, 24), and this stands in sharp contrast to the wealth of data available from numerous genome-wide KO or knock-down studies (1, 25). For these reasons, in silico approaches to model gene overexpression and studies that apply them genome-wide are highly desirable. Here, we conduct a genome-scale study of the deleterious consequences of inducing metabolic gene overexpression through an in silico algorithm for predicting Expression-Dependent Gene Effects (EDGE). First, we show that EDGE successfully predicts growth retardation phenotypes arising due to inappropriate gene activation in microorganisms, which are a prime concern for metabolic engineers. Second, we show EDGE's universal applicability by demonstrating across multiple organisms that genes predicted to be nondeleterious when highly expressed are indeed the ones with the highest expression levels. Third, we show EDGE's applicability to the study of human disease by demonstrating that some aspects of genetic reprogramming in cancer can be explained as an attempt to silence genes whose expression is detrimental toward cancer proliferation.

Our study is conducted within the realm of cellular metabolism, which is particularly amenable to in silico modeling on a genome scale. Genome-scale metabolic models (GSMMs) translate the capabilities of a metabolic network, dictated by an organism's genome, into a set of mathematical equations (26, 27). They offer a powerful tool for predicting the outcomes of genetic perturbations through testable mechanistic explanations. Notably, GSMMs have been quite successful in predicting phenotypes of loss-of-function mutants (28–30), and therefore hold considerable promise to successfully predict the outcomes of induced gene overexpression as well. A GSMM typically includes a stoichiometric matrix, which represents the network's topology, constraints (e.g., thermodynamic or environmental constraints) applied to it, and gene–protein associations. EDGE takes a GSMM

Significance

Biologists frequently overexpress genes to learn about their cellular functions, and biotechnologists do so to construct novel metabolic pathways that produce valuable chemical compounds. However, gene overexpression often leads to deleterious consequences whose cause is unclear. Here, we present a computational method named Expression-Dependent Gene Effects (EDGE) that can successfully predict the deleterious effects resulting from overexpression of either native or foreign (originating in another species) metabolic genes. EDGE relies on genome-scale metabolic models, an emerging computational paradigm for studying metabolism in silico. Beyond its biotechnological significance, gene overexpression also plays an important role in human disease. We show EDGE's applicability in the latter case by demonstrating its ability to detect toxic genes whose expression tends to be suppressed in cancer cells.

Author contributions: A.W., U.G., and E.R. designed research; A.W., L.R., and C.G. performed research; A.W., R.Z., L.R., R.S., U.G., and E.R. analyzed data; and A.W. and E.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ruppin@post.tau.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1312361110/-/DCSupplemental.

as its input and quantifies the benefits or losses that a cell incurs by activating a certain metabolic gene in a given environment (Fig. 1, where blue boxes represent EDGE's inputs and purple boxes represent computation outputs; *Materials and Methods*). For this, it relies on a hypothesized cellular objective [in this study, we always took the objective to be the commonly assumed maximization of biomass production (31, 32)]. Genes are then classified as (*i*) beneficial, i.e., contributing toward the realization of the objective (positive score); (*ii*) detrimental to the objective (negative score); or (*iii*) neutral with respect to the objective (zero score; *SI Appendix*).

Results

EDGE Algorithm. A complete description of EDGE is given in *Materials and Methods*. In short, the EDGE score of a gene, g, quantifies the utility of transcribing the gene in small levels compared with not using g at all. This notion is strongly related to analyzing the sensitivity of a mathematical program to a perturbation of its constraints [sometimes called "shadow pricing" (33)]. Such analyses are fundamental to mathematical programming in general and have previously been applied to GSMMs (27, 34–36). EDGE is an adaptation of these analyses to the special setting of flux balance analysis (FBA) (32). EDGE measures the sensitivity of the optimal objective value to the simultaneous perturbation of multiple constraints that are associated with a particular gene. EDGE scores are not dependent on a particular optimum, which is desirable in the case of GSMMs, whose solution space typically contains multiple optima (37).

EDGE simulates the expression of a given gene by enforcing a flux through reactions associated with it; reversible reactions are constrained to carry a flux through either direction. When a gene is associated with more than one reaction, EDGE finds the bottleneck reaction whose limitation has the greatest effect and determines the EDGE score by its utility with regard to the a priori objective function. In each test in this study, we excluded genes whose proteins catalyze blocked reactions (i.e., reactions that cannot carry a flux in steady state in the medium relevant to the test; *SI Appendix, Supplementary Methods*).

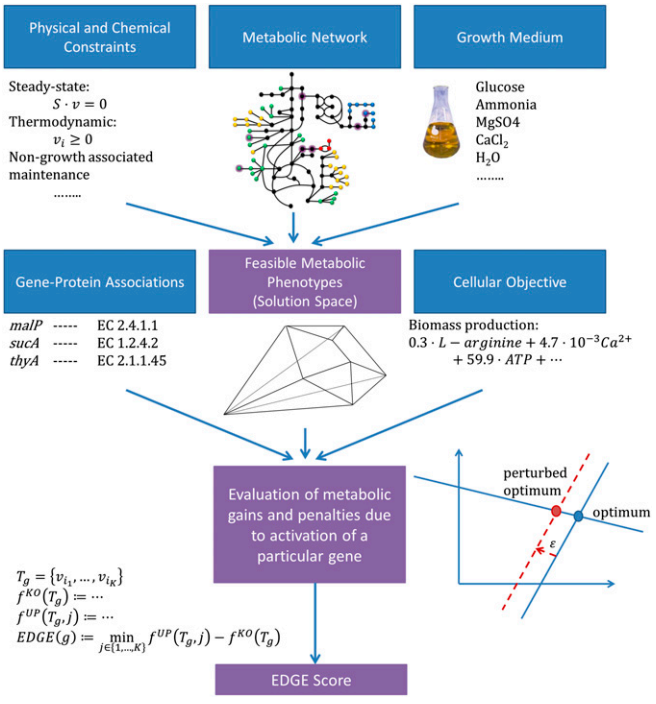

Fig. 1. Schematic workflow of EDGE analysis. Blue boxes represent inputs to the computation, and purple boxes represent computation outputs. EDGE relies on a GSMM, constraints applied to it, an objective function, and an in silico growth medium; all these define a space of feasible metabolic phenotypes. EDGE is formulated as a mixed integer linear program that quantifies the positive or negative effect that small changes in the expression level of a particular gene have on the cellular objective due to the flux rerouting that they induce. The image of the network under the box entitled Metabolic Network was adapted from the work of Montagud et al. (62).

EDGE Predicts Reduced Fitness Due to Overexpression of Native Genes. We used EDGE to predict which metabolic *Escherichia coli* genes will prove toxic when overexpressed during growth on glucose-supplemented M9 minimal medium and ranked them by the confidence level of the prediction (*SI Appendix*). Twenty-six high-ranking genes were chosen for subsequent growth experiments: 12 that were confidently predicted to be toxic and 14 that were confidently predicted not to be toxic (Dataset S1, Table S1). Plasmids (23) containing Isopropyl-β-D-thiogalactoside (IPTG)–inducible constructs of these genes were transformed into a WT K-12 MG1655 *E. coli* strain, and clones were grown in a minimal (M9) medium supplemented with glucose and 0–1 mM IPTG (*Materials and Methods*). The growth inhibition resulting from the genes' experimental overexpression was found to be highly correlated with the genes' EDGE scores (Spearman $\rho = 0.74$, $P < 7.56\text{e-}6$; Fig. 2 and *SI Appendix, Supplementary Notes* and Fig. S1; all P values reported in this paper are against one-sided alternatives unless noted otherwise). These results affirm EDGE's capability to flag potentially toxic genes, and thus facilitate the design of novel metabolic pathways.

We then turned to conducting a large-scale validation through two existing genome-scale overexpression libraries: the ASKA library for *E. coli* (23) and the yeast GST-tagged collection for *Saccharomyces cerevisiae* (24). Both studies carried out systematic, large-scale overexpression of ORFs and listed genes that were toxic (i.e., caused severe growth inhibition when overexpressed) (*SI Appendix, Supplementary Notes*). The predicted EDGE scores of experimentally toxic genes proved to be significantly lower than those of nontoxic genes ($P < 6.1\text{e-}4$ and $P < 7.8\text{e-}5$ for ASKA and yeast Gal-GST, respectively, for a rank sum test; *SI Appendix*, Fig. S2). In both datasets, experimentally toxic genes were significantly enriched among the in silico detrimental genes (hypergeometric $P < 1.5\text{e-}4$ and $P < 0.018$, respectively).

EDGE Predicts Reduced Fitness Due to Overexpression of Foreign Genes Within *E. coli*. So far, we were concerned with the overexpression of genes within their native host. Can EDGE predict in a similar manner the consequences of expressing foreign genes within an organism? It was previously observed that gaps in Sanger-based genome sequencing are often caused by toxic genes that cannot be expressed in an *E. coli* host (20). To study EDGE's ability to predict failed gene transfer between organisms due to toxic effects, we used the recently published PanDaTox dataset (21) of genes found to be unclonable into *E. coli*. We simulated in silico the process of gene transfer from 50 different Gammaproteobacteria into *E. coli*, and then used EDGE to predict which of the heterologous genes should be toxic to the *E. coli* host (*SI Appendix*). Comparing the results with PanDaTox's experimental data, we found that EDGE scores were highly predictive of the experimental outcome, with a median area under the ROC curve (AUC) of 0.71 when distinguishing unclonable genes from clonable genes based on their EDGE scores (median $P < 0.00048$; the result is significant for 42 of the Gammaproteobacteria using a 5% false discovery rate level; Dataset S1, Table S2). The focus on Gammaproteobacteria (the class to which *E. coli* belongs) was due to the nature of the PanDaTox data. When the source organism's gene promoters are not recognized by the *E. coli* transcription machinery, little or no gene product is produced; therefore, no toxicity (i.e., "inclonability") can be observed (20). EDGE, on the other hand, assumes by definition that the gene in question is successfully transcribed. Indeed, a clear inverse correlation was observed between EDGE's success rate and the phylogenetic

SYSTEMS BIOLOGY

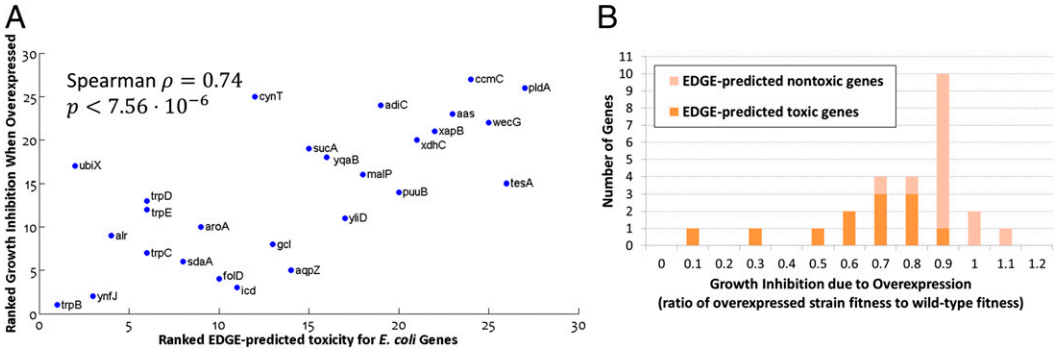

Fig. 2. EDGE predicts deleterious effects due to gene overexpression in *E. coli*. We conducted growth experiments in which we measured the growth inhibition resulting from the overexpression of 26 genes in glucose-supplemented M9 minimal medium. The genes were selected by the confidence level assigned by EDGE to the predicted outcome. Twelve genes were confidently predicted to be toxic, and 14 were confidently predicted to be nontoxic. (*A*) Magnitude of growth inhibition resulting from a gene's overexpression is strongly correlated with its EDGE-predicted toxicity score (Spearman $\rho = 0.74$, $P <$ 7.56e-6). (*B*) Histogram of the growth inhibitions resulting from overexpression of each of the 26 genes. Growth inhibition was quantified as the ratio of the fitness following IPTG-induced overexpression to the fitness with no induced overexpression (*SI Appendix*). Therefore, 0 denotes growth arrest and 1 denotes no inhibition compared with the WT. Three genes (*trpB*, *ynfJ*, and *icd*) obtained an "inhibition ratio" that was higher than 1.0, which means that the IPTG induction did not inhibit growth but rather contributed to it.

distance of the source organism from *E. coli* (Fig. 3; Spearman $\rho = -0.32$, $P <$ 9e-5), which reflects transcriptional compatibility. Nonetheless, even when considering all 138 bacteria in the PanDaTox dataset for which data were available (*SI Appendix*), EDGE's predictive power remained highly significant (combined Fisher $P <$ 5.5e-203, median AUC = 0.67, median P value <0.0085; median AUC is larger than expected by chance with $P <$ 1e-308; Dataset S1, Table S2). In conclusion, these results support the hypothesis that the transfer of metabolic genes (both by natural horizontal gene transfer and in metabolic engineering efforts) can be hindered by deleterious effects of transcribing them in the recipient organism, and that many of these events can be successfully predicted by EDGE.

Genes Whose Expression Impedes Growth (According to EDGE) Are Lowly Expressed Across a Wide Variety of Species and Tissues. Next, we expected genes that EDGE predicts to be disadvantageous in a given environment to be lowly expressed in vivo when the organism grows in that environment. We tested this hypothesis by comparing extensive microarray data for *E. coli* and *S. cerevisiae*, spanning multiple growth conditions whose medium composition was recreated in silico, with EDGE predictions. In all cases, EDGE-predicted detrimental genes were significantly lowly expressed compared with genes predicted to be beneficial (Fig. 4*A*; rank sum test, $P <$ 8.8e-18 for all 20 *E. coli* microarrays except for growth on LB with $P <$ 1.6e-5 and $P <$ 0.0027 for all seven yeast arrays except for one with $P <$ 0.0165; *SI Appendix*, *Supplementary Notes* and Dataset S1, Table S3). We believe that the relatively weaker result for *E. coli* growth on LB medium stems from the incomplete characterization of its chemical contents (leading to its inexact definition in silico) in comparison to the synthetic M9 medium that was used as the basis for all other *E. coli* arrays. Similarly, the yeast data were obtained on YP medium whose exact chemical composition is unknown.

The availability of transcriptomic data for human and plant cells allowed us to test EDGE's applicability in these cases as well. Whereas in the case of microorganisms biomass yield is a common approximation of the cellular objective, the objective of human or plant cells is far more complex and tissue-specific. Nonetheless, even when using biomass yield as a proxy cellular objective (reflecting the need to replenish metabolites continuously due to ongoing metabolic turnover), EDGE-predicted detrimental genes were significantly lowly expressed in comparison to genes predicted to be beneficial across 79 different human tissues, 6 of them cancerous and the rest healthy (38) (rank sum $P <$ 1.4e-5 in all arrays except for superior cervical ganglion, with $P = 0.141$; Dataset S1, Table S4), and across all the NCI60 cancer cell lines (39) (rank sum $P <$ 8.6e-9 in all cases; Dataset S1, Table S5). We then analyzed in a similar manner transcriptomic data of *Arabidopsis thaliana* measured in 79 biological contexts and spanning multiple organs and developmental phases of the plant (40), and found a similar trend (rank sum $P < 0.05$ for 72 of 79 microarrays, median $P <$ 3e-5 across all 79 arrays; *SI Appendix*, *Supplementary Notes* and Dataset S1, Table S6). Remarkably, the magnitude of this effect in human and plant transcriptomes is on the order of that we had previously observed in microorganisms (Fig. 4*B*). We conclude that genes that EDGE predicts to be detrimental toward proliferation are lowly expressed in diverse organisms, both in microbes and in multicellular species. The activation of these genes is thus likely to be highly undesirable and results in reduced fitness that EDGE successfully predicts. Taken together,

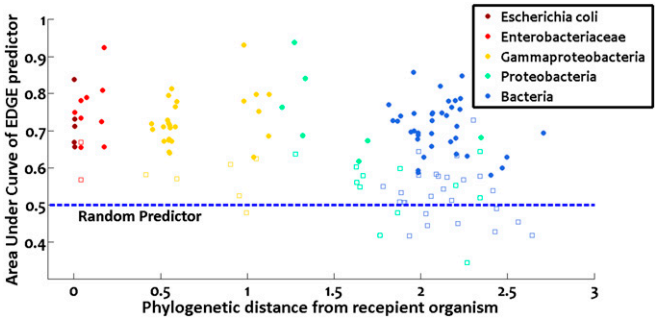

Fig. 3. EDGE predicts whether the transfer of foreign genes into *E. coli* would be successful. Each point represents one bacterial species. For each of its metabolic genes, EDGE predicted whether its transfer to *E. coli* would be successful, and the results were compared with previously published experimental data (20, 21). The predictor's quality was quantified by the area under the ROC curve (AUC) (*y* axis). A random predictor achieves AUC = 0.5. Filled circles and empty squares represent significant and nonsignificant results, respectively, following correction for multiple hypotheses using a 5% false discovery rate level. Color codes denote the phylogenetic relatedness of the bacterium in question to *E. coli* according to the National Center for Biotechnology Information's taxonomy tree (i.e., most recent common ancestor node of the bacterium and *E. coli* in that tree). Clearly, EDGE's predictive power is inversely correlated with the bacterium's phylogenetic distance (63) from *E. coli* (*x* axis; Spearman $\rho = -0.32$, $P <$ 9e-5; eight bacteria that are not part of this phylogenetic tree are omitted from the figure). This is to be expected, because if the transferred genes cannot be transcribed by the host cell due to promoter dissimilarity, they would be clonable even if they are toxic.

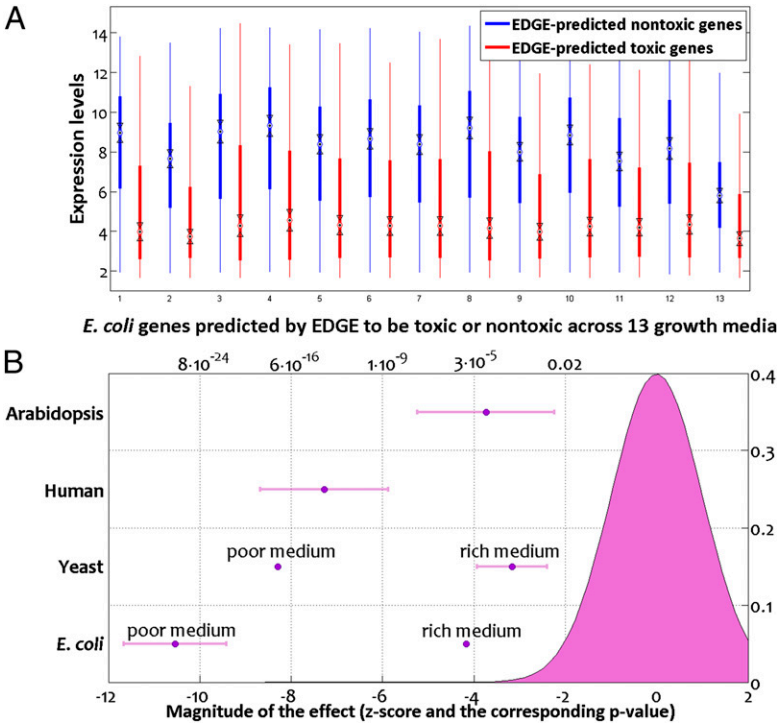

Fig. 4. (*A*) *E. coli* genes predicted by EDGE to be toxic when overexpressed are significantly down-regulated compared with genes predicted to be nontoxic. Results are presented for 13 different growth conditions (64), as detailed in Dataset S1, Table S3*A*. (*B*) Magnitude by which EDGE-predicted toxic genes are lowly expressed compared with EDGE-predicted nontoxic genes in the case of human and plant cells (*Upper* two rows) resembles that observed in microorganisms (*Lower* two rows). The magnitude of this effect is quantified by the statistic of the Wilcoxon rank sum test, normalized to a standard normal variable (*Lower x* axis). The *P* values that match those test statistics, when considering a one-sided alternative, are given as well (*Upper x* axis), along with a standard normal curve for a sense of scale. Each of the four rows in the *y* axis represents one organism for which we tested the aforementioned effect using transcriptome data and denotes the mean value and SD of the corresponding test statistics. In the case of microorganisms, we separate experiments conducted in rich and poor, defined media because the former's composition is uncertain, leading to inexact representation in silico, and thus to less accurate predictions.

these findings reflect a universal selection force keeping the expression of potentially deleterious genes in check.

Genetic Reprogramming in Cancer Suppresses the Expression of Proliferation-Inhibiting Genes. We next turned to study the ability of EDGE to predict genes whose activation impedes proliferation in cancer cells. We first analyzed a dataset of DNA somatic copy number alterations in cancer (41) and found that the set of genes reported to be deleted in at least one of the cancer types in that dataset is significantly enriched with genes that EDGE identifies as detrimental with respect to biomass yield (hypergeometric $P < 0.0013$), whereas genes that are beneficial with respect to that objective are underrepresented (hypergeometric $P < 2.5\text{e-}4$). We then analyzed three studies that compared mRNA levels of cancer cells and their healthy counterparts. We found that in all three cases, the genes that were down-regulated in the cancerous cells were enriched with genes that EDGE classified as detrimental toward proliferation (*SI Appendix* and datasets 1–3 in Dataset S1, Table S7). Intriguingly, EDGE was also able to reveal the down-regulation of genes that impede proliferation in histologically normal breast epithelia taken from women held to be at high risk for developing breast cancer (dataset 4 in Dataset S1, Table S7).

Finally, we returned to the aforementioned dataset of gene expression of 79 healthy and cancerous human tissues. In light of the better adherence of cancer cells to the proliferation cellular objective, it was not surprising that the effect we have previously observed (namely, down-regulation of EDGE-predicted toxic genes compared with genes predicted to be nontoxic) was particularly pronounced in cancer tissues. This yields a remarkable ability to separate between cancer and benign tissues in that data based only on the correlation between the gene expression measured in each tissue and EDGE scores (AUC = 0.96, $P < 1.64\text{e-}22$; *SI Appendix*). Furthermore, when ranking the samples by the magnitude of that effect, we found that the highest ranking healthy tissues were lymphoblasts, adipocytes, and bronchial epithelial cells, all of which are known to be proliferative. In contrast, three of the four lowest ranking tissues (including the only case in which the effect was not statistically significant) were nonproliferative neuronal ganglia cells, with the fourth being another nonproliferative source of rhythmic cells in the cardiac atrioventricular node (*SI Appendix*, Fig. S3 and Dataset S1, Table S4). We conclude that EDGE has a clear predictive power in identifying genes whose expression potentially impedes proliferation of human cells.

Discussion

Our results that concern gene expression patterns in *E. coli* are in line with the findings of Lewis et al. (42), who reported that inefficient metabolic pathways are down-expressed in two strains of evolved *E. coli*. Here, we show that this phenomenon is universal across multiple growth conditions and holds not only in bacteria but also in eukaryotes, including humans. In addition, Lewis et al. (42) showed that the evolution of those *E. coli* strains was facilitated by up-regulation of optimal genes and, in many cases, by down-regulation of suboptimal genes. We extend these observations and show that the genetic reprogramming in cancer can be partially explained in a similar manner, as alterations intended to down-regulate genes that are detrimental to proliferation. The magnitude of this down-regulation is large enough to enable an almost perfect separation between normal and cancerous microarray samples based solely on EDGE results. Although such separation can be achieved through various computational methods (43), the fact that it can be achieved while relying only on EDGE-based analysis shows that EDGE successfully captures a distinct proliferative signature in the transcriptome of human cells.

Beyond its contribution to obtaining better mechanistic insights into the way gene expression levels are controlled by their potential toxicity, EDGE bears considerable applicative value for biotechnologists: Genome-scale metabolic modeling has already been successfully applied to devise novel pathways for rational strain design (12, 44–51), and gene overexpression has been considered in this framework as a means to produce a desired chemical (52, 53). EDGE complements the existing computational methods by addressing a prime concern of metabolic engineers, who seek to foresee and mitigate the deleterious effects that often accompany the introduction of a foreign metabolic pathway into a host organism, or the overexpression of one of its native genes (17, 18). Because EDGE provides a mechanistic

insight for its predicted deleterious effects, it could be used to suggest further perturbations that can mitigate the expected deleterious effects of gene overexpression within an organism, or of gene transfer between organisms. These perturbations may include either media supplements, i.e., nutrients, that can abolish the rerouted fluxes and reinstate normal biomass production, or genetic perturbations (e.g., gene KOs) that may serve the same purpose.

EDGE operates on a GSMM (26, 27) (Fig. 1). As such, EDGE scores are computed only for genes that take part in the metabolic model, and their predictive power is dependent on the model's quality. Manually curated GSMMs, which have demonstrated their predictive power, exist for many model organisms, including industrial microorganisms (28, 54) and human (55), and undergo constant improvement; plant models have been recently published as well (56). The Model SEED platform was the first to generate GSMMs automatically (57). Although these models do not outperform manually curated models at the present time (57), the Model SEED platform enables studies that require metabolic reconstructions of multiple species, for which no manually curated GSMMs are available, and it was used in this study for testing EDGE against the PanDaTox data.

WT internal flux levels are generally unavailable and cannot be computed to a sufficient precision without additional inputs beyond a GSMM (28, 58, 59). For this reason, EDGE quantifies the difference (with respect to the objective function) between having the relevant fluxes carry epsilontic and zero fluxes, rather than the difference between the WT plus epsilon and WT fluxes. However, EDGE scores are highly predictive of overexpression phenotypes even for genes that are expressed in the WT, as we demonstrate in this paper.

We expect future work to further improve EDGE's predictive capability on three levels: first, by detecting and removing deleterious futile metabolic cycles (60); second, by taking into account the promiscuous enzymatic functions that are not normally manifested in the metabolic network but could be catalyzed by an overexpressed protein (2); and, third, by incorporating better-tuned, tissue-specific objective functions for human and plant cells. EDGE could actually facilitate the identification of such functions on account of its proven ability to detect correspondence between empirical transcriptome data and a hypothesized cellular objective.

Materials and Methods

Strain Construction. The ASKA background strain AG1 [*recA1 endA1 gyrA96 thi-1 hsdR17*($r_K^- m_K^+$) *supE44 relA1*] (23) is a thiamine auxotroph, which is unable to grow in M9 medium. We therefore transformed the plasmids into the MG1655 WT strain, enabling work in minimal M9 medium. This was also desirable because the *E. coli* metabolic model used in this study is based on the latter strain (28).

The ASKA GFP-tagged *E. coli* strain library (23) was obtained from the National Bioresource Project at the National Institute of Genetics, Japan. Plasmid DNA was isolated from strains overexpressing the required genes, and the GFP tag sequence was removed by *Not*I restriction followed by self-ligation. The plasmids were then transformed by electroporation to WT *E. coli* strain MG1655. Correct sequence of the plasmid was verified by amplifying the insert using pCA24N (ASKA backbone)-specific primers (forward: 5′-ATC ACC ATC ACC ATA CGG AT; reverse: 5′-CTG AGG TCA TTA CTG GAT CTA) and sequencing the product using the pCA24N reverse primer.

Growth Experiments. *E. coli* MG1655 clones harboring plasmids of choice were picked into LB medium supplemented with chloramphenicol (34 μg/mL) and grown to OD_{595} = 0.5. Cells were washed once in saline and resuspended at a dilution of 1:100 into M9-glucose-chloramphenicol minimal media (1× M9 salts, 2 mM $MgSO_4$, 0.1 mM CaCl, 0.4% glucose, 34 μg/mL chloramphenicol) supplemented with 0–1 mM IPTG. Growth measurements were performed in a 96-well plate incubated for 19–24 h at 37 °C in a temperature-controlled plate reader with continuous shaking (ELX808IU-PC; Biotek), and OD_{595} was monitored every 15 min. Each strain/medium combination was loaded into 2 duplicate wells. The entire growth experiment was repeated two to five times for each strain.

EDGE Algorithm. The EDGE score of a gene, g, quantifies the utility in transcribing the gene in epsilontic levels compared with not using g at all. This notion is strongly related to analyzing the sensitivity of a mathematical program to a perturbation of its constraints [sometimes called shadow pricing (33)]. Such analyses are fundamental to mathematical programming in general and have previously been applied to GSMMs (27, 34–36). EDGE is an adaptation of these analyses to the special setting of FBA (32). EDGE measures the sensitivity of the optimal objective value to epsilontic perturbations in a particular gene's expression, with each gene potentially affecting multiple reactions in the network, and thus multiple constraints. For this, EDGE relies on the gene–protein-reaction mapping embedded in the model. A gene's EDGE score is uniquely determined and is not dependent on a particular optimum, which is desirable in the case of GSMMs, whose solution space typically contains multiple optima (37).

EDGE simulates the expression of a given gene by enforcing minimal flux through reactions associated with it; reversible reactions are constrained to carry a minimal flux through either direction. When a gene is associated with more than one reaction, we find the bottleneck reaction, whose limitation has the greatest effect, and determine the EDGE score by its utility with regard to the a priori objective function.

We now present EDGE's full formulation. Given a gene g, let $T_g = \{v_{i_1}, \ldots, v_{i_K}\}$ denote the set of reactions in the network that are associated with g. We define:

$$\mathrm{EDGE}(g) := \min_{j\in\{1,\ldots,K\}} f^{UP}(T_g, j) - f^{KO}(T_g),$$

where f^{KO} is the optimal objective subject to silencing g. The minuend $\min_{j\in\{1,\ldots,K\}} f^{UP}(T_g, j)$ 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 the 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:

$$f^{KO}(T_g) := \max_{v\in\mathbb{R}^n} f(v),$$

subject to (*i*) $S\cdot v = 0$, (*ii*) $\forall i = 1, \ldots, n.\ \alpha_i \le v_i \le \beta_i$, and (*iii*) $\forall v_i \in T_g : v_i = 0$.

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

$$f^{UP}(T_g, j) := \max_{v\in\mathbb{R}^n, a\in\{0,1\}^K} f(v)$$

subject to (*i*) $S\cdot v = 0$, (*ii*) $\forall i = 1, \ldots, n.\ \alpha_i \le v_i \le \beta_i$, (*iii*) $\forall v_{i_k} \in T_g \backslash \{v_{i_j}\} : a_k = 1 \rightarrow v_{i_k} \ge \varepsilon$, $a_k = 0 \rightarrow v_{i_k} \le -\varepsilon$, (*iv*) $a_j = 1 \rightarrow v_{i_j} = \varepsilon$, and (*v*) $a_j = 0 \rightarrow v_{i_j} = -\varepsilon$, where ε is an infinitesimal constant chosen to reflect the smallest nonnegligible flux possible. However, ε cannot be arbitrarily small due to the finite precision of the floating-point representation. a_k variables 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 (61). Commercial solvers are sometimes able to branch explicitly on these constraints. We note that we described the algorithm as adding an 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 in *SI Appendix, Supplementary Methods*.

Genes were classified as toxic if they had a negative EDGE score and as nontoxic if they had a positive EDGE score. For the purpose of conducting growth experiments (*Results*), 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 (*SI Appendix, Supplementary Methods*).

ACKNOWLEDGMENTS. We thank Livnat Jerby, Martin Kupiec, Nathan E. Lewis, and Roded Sharan for many fruitful discussions. This research was supported by the Israeli Centers of Research Excellence program of the Israeli Planning and Budgeting Committee and the Israel Science Foundation (Grant 41/11) and by the Seventh Framework Programme of the European Union Microme and Infect Projects. A.W. was supported, in part, by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. R.S. was supported, in part, by National Institutes of Health Grant R01AI082376-01, Israeli Science Foundation Grant 1303/12, and European Research Council Starting Grants program Grant 260432.

1. Nichols RJ, et al. (2011) Phenotypic landscape of a bacterial cell. *Cell* 144(1):143–156.
2. Prelich G (2012) Gene overexpression: Uses, mechanisms, and interpretation. *Genetics* 190(3):841–854.
3. van't Veer LJ, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871):530–536.
4. Cooper-Knock J, et al. (2012) Gene expression profiling in human neurodegenerative disease. *Nat Rev Neurol* 8(9):518–530.
5. Oh SC, et al. (2012) Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. *Gut* 61(9):1291–1298.
6. Graham M, Shutter JR, Sarmiento U, Sarosi I, Stark KL (1997) Overexpression of Agrt leads to obesity in transgenic mice. *Nat Genet* 17(3):273–274.
7. Fleming SM, et al. (2004) Early and progressive sensorimotor anomalies in mice overexpressing wild-type human α-synuclein. *J Neurosci* 24(42):9434–9440.
8. Hallett PJ, McLean JR, Kartunen A, Langston JW, Isacson O (2012) α-Synuclein overexpressing transgenic mice show internal organ pathology and autonomic deficits. *Neurobiol Dis* 47(2):258–267.
9. Martinou J-C, et al. (1994) Overexpression of BCL-2 in transgenic mice protects neurons from naturally occurring cell death and experimental ischemia. *Neuron* 13(4): 1017–1030.
10. Parkes TL, et al. (1998) Extension of Drosophila lifespan by overexpression of human SOD1 in motorneurons. *Nat Genet* 19(2):171–174.
11. Betz AL, Yang G-Y, Davidson BL (1995) Attenuation of stroke size in rats using an adenoviral vector to induce overexpression of interleukin-1 receptor antagonist in brain. *J Cereb Blood Flow Metab* 15(4):547–551.
12. Lee JW, et al. (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat Chem Biol* 8(6):536–546.
13. Keasling JD (2010) Manufacturing molecules through metabolic engineering. *Science* 330(6009):1355–1358.
14. Martin VJJ, Pitera DJ, Withers ST, Newman JD, Keasling JD (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat Biotechnol* 21(7):796–802.
15. Ro D-K, et al. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440(7086):940–943.
16. Steen EJ, et al. (2010) Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* 463(7280):559–562.
17. Pitera DJ, Paddon CJ, Newman JD, Keasling JD (2007) Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metab Eng* 9(2):193–207.
18. Zhang F, Carothers JM, Keasling JD (2012) Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat Biotechnol* 30(4):354–359.
19. Saïda F, Uzan M, Odaert B, Bontems F (2006) Expression of highly toxic genes in E. coli: Special strategies and genetic tools. *Curr Protein Pept Sci* 7(1):47–56.
20. Sorek R, et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318(5855):1449–1452.
21. Kimelman A, et al. (2012) A vast collection of microbial genes that are toxic to bacteria. *Genome Res* 22(4):802–809.
22. Amitai G, Sorek R (2012) PanDaTox: A tool for accelerated metabolic engineering. *Bioengineered* 3(4):218–221.
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. Sopko R, et al. (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* 21(3):319–330.
25. Costanzo M, et al. (2010) The genetic landscape of a cell. *Science* 327(5964):425–431.
26. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 10(4):291–305.
27. Palsson BO (2006) *Systems Biology: Properties of Reconstructed Networks* (Cambridge Univ Press, Cambridge, UK).
28. 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.
29. 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(6): R69.
30. Szappanos B, et al. (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* 43(7):656–662.
31. Feist AM, Palsson BO (2010) The biomass objective function. *Curr Opin Microbiol* 13(3):344–349.
32. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* 28(3):245–248.
33. Bradley SP, Hax AC, Magnanti TL (1977) *Applied Mathematical Programming* (Addison–Wesley, Reading, MA).
34. Varma A, Boesch BW, Palsson BO (1993) Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. *Appl Environ Microbiol* 59(8):2465–2473.
35. Edwards JS, Ramakrishna R, Palsson BO (2002) Characterizing the metabolic phenotype: A phenotype phase plane analysis. *Biotechnol Bioeng* 77(1):27–36.
36. Reznik E, Mehta P, Segrè D (2013) Flux imbalance analysis and the sensitivity of cellular growth to changes in metabolite pools. *PLOS Comput Biol* 9(8):e1003195.
37. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 5(4):264–276.
38. Su AI, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101(16):6062–6067.
39. Lee JK, et al. (2007) A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc Natl Acad Sci USA* 104(32):13086–13091.
40. Schmid M, et al. (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet* 37(5):501–506.
41. Beroukhim R, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463(7283):899–905.
42. Lewis NE, et al. (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol* 6:390.
43. Fishel I, Kaufman A, Ruppin E (2007) Meta-analysis of gene expression data: A predictor-based approach. *Bioinformatics* 23(13):1599–1606.
44. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 84(6):647–657.
45. Ranganathan S, et al. (2012) An integrated computational and experimental study for overproducing fatty acids in Escherichia coli. *Metab Eng* 14(6):687–704.
46. Lee KH, Park JH, Kim TY, Kim HU, Lee SY (2007) Systems metabolic engineering of Escherichia coli for L-threonine production. *Mol Syst Biol* 3:149.
47. Park JH, Lee KH, Kim TY, Lee SY (2007) Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc Natl Acad Sci USA* 104(19):7797–7802.
48. Becker J, Zelder O, Häfner S, Schröder H, Wittmann C (2011) From zero to hero—Design-based systems metabolic engineering of Corynebacterium glutamicum for L-lysine production. *Metab Eng* 13(2):159–168.
49. Asadollahi MA, et al. (2009) Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering. *Metab Eng* 11(6):328–334.
50. Izallalen M, et al. (2008) Geobacter sulfurreducens strain engineered for increased rates of respiration. *Metab Eng* 10(5):267–275.
51. Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 5:320.
52. Pharkya P, Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab Eng* 8(1):1–13.
53. Kim J, Reed JL (2010) OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol* 4:53.
54. Mo ML, Palsson BO, Herrgård MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* 3:37.
55. Duarte NC, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* 104(6):1777–1782.
56. de Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol* 152(2):579–589.
57. Henry CS, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 28(9):977–982.
58. Sauer U (2006) Metabolic networks in motion: 13C-based flux analysis. *Mol Syst Biol* 2:62.
59. Suthers PF, et al. (2007) Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metab Eng* 9(5-6):387–405.
60. Pinchuk GE, et al. (2010) Constraint-based model of Shewanella oneidensis MR-1 metabolism: A tool for data analysis and hypothesis generation. *PLOS Comput Biol* 6(6):e1000822.
61. Bisschop J (2011) *AIMMS—Optimization Modeling* (Paragon Decision Technology, Haarlem, The Netherlands).
62. Montagud A, Navarro E, Fernández de Córdoba P, Urchueguía JF, Patil KR (2010) Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst Biol* 4:156.
63. Dehal PS, et al. (2009) MicrobesOnline: An integrated portal for comparative and functional genomics. *Nucleic Acids Res* 38(Database issue):D396–D400.
64. Lewis NE, Cho B-K, Knight EM, Palsson BO (2009) Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: Providing context for content. *J Bacteriol* 191(11):3437–3444.