## Title
A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism



## Journal
Cell Reports, 27(11)

## ISSN
2639-1856

## Authors
Schober, Andrew F
Mathis, Andrew D
Ingle, Christine
et al.

## Publication Date
2019-06-01

## DOI
10.1016/j.celrep.2019.05.030

Peer reviewed

HHS Public Access

Author manuscript

*Cell Rep*. Author manuscript; available in PMC 2019 July 12.

Published in final edited form as:

*Cell Rep*. 2019 June 11; 27(11): 3359–3370.e7. doi:10.1016/j.celrep.2019.05.030.

# A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism

**Andrew F. Schober**[1,2], **Andrew D. Mathis**[1], **Christine Ingle**[1], **Junyoung O. Park**[3,4,8], **Li Chen**[3,5], **Joshua D. Rabinowitz**[3,5], **Ivan Junier**[6], **Olivier Rivoire**[7], and **Kimberly A. Reynolds**[1,2,9,*]

[1]The Green Center for Systems Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

[4]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA

[5]Department of Chemistry, Princeton University, Princeton, NJ 08544, USA

[6]Centre National de la Recherche Scientifique, Université Grenoble Alpes, TIMC-IMAG, F-38000 Grenoble, France

[7]Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, PSL Research University, F-75005 Paris, France

[8]Present address: Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, Los Angeles, CA 90095, USA

[9]Lead Contact

## SUMMARY

Enzyme function and evolution are influenced by the larger context of a metabolic pathway. Deleterious mutations or perturbations in one enzyme can often be compensated by mutations to others. We used comparative genomics and experiments to examine evolutionary interactions with the essential metabolic enzyme dihydrofolate reductase (DHFR). Analyses of synteny and co-occurrence across bacterial species indicate that DHFR is coupled to thymidylate synthase (TYMS) but relatively independent from the rest of folate metabolism. Using quantitative growth rate measurements and forward evolution in *Escherichia coli*, we demonstrate that the two enzymes adapt as a relatively independent unit in response to antibiotic stress. Metabolomic

*Correspondence: kimberly.reynolds@utsouthwestern.edu.

AUTHOR CONTRIBUTIONS

Conceptualization, A.F.S., I.J., O.R., and K.A.R.; Methodology, A.F.S., A.D.M., J. D.R., I.J., O.R., and K.A.R.; Investigation, A.F.S., A.D.M., C.I., J.O.P., L.C., I.J., O.R., and K.A.R.; Writing – Original Draft, A.S. and K.A.R.; Writing – Review & Editing, A.S., A.D.M., C.I., J.O.P., J.D.R., I.J., O.R., and K.A.R.; Supervision K.A.R.

SUPPLEMENTAL INFORMATION

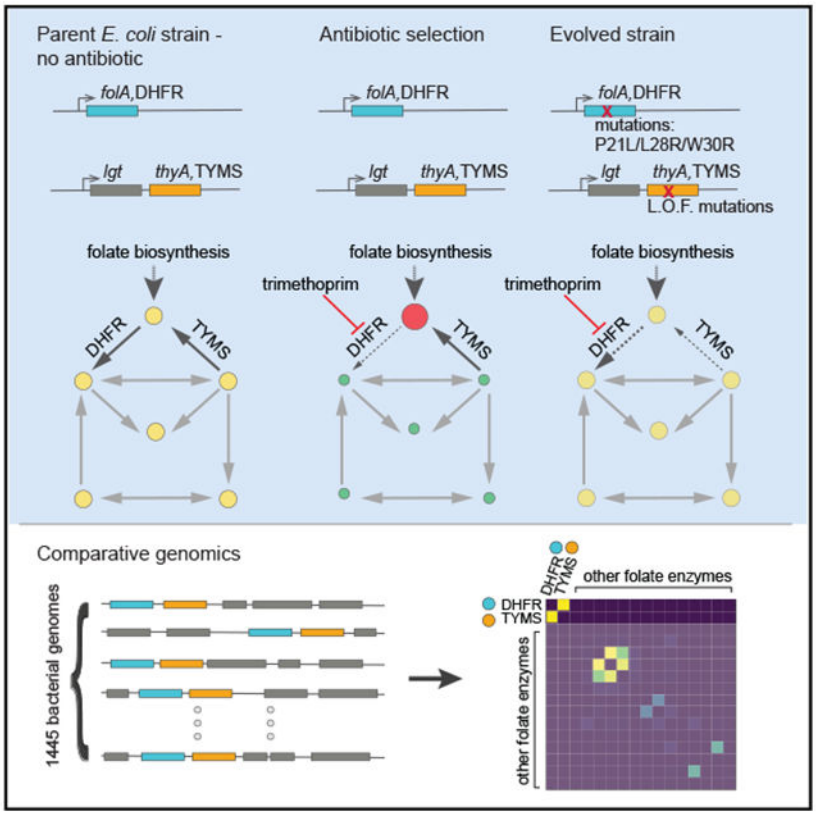
Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.05.030.

DECLARATION OF INTERESTS

The authors declare no competing interests.

profiling revealed that TYMS activity must not exceed DHFR activity to prevent the depletion of reduced folates and the accumulation of the intermediate dihydrofolate. Comparative genomics analyses identified >200 gene pairs with similar statistical signatures of modular co-evolution, suggesting that cellular pathways may be decomposable into small adaptive units.

## Graphical Abstract




## In Brief

Comparative genomics identified the enzymes DHFR and TYMS as an evolutionary module embedded within folate metabolism. Schober et al. show experimentally that these enzymes adapt as a unit in response to DHFR inhibition with an antibiotic. Extending comparative genomics analyses genome wide suggests >200 additional candidate adaptive units throughout bacterial metabolism.

## INTRODUCTION

The collective action of enzymes in metabolism produces the basic materials for cells to grow and divide. Biochemical pathway maps provide a rich description of the reactions and intermediates needed for metabolism, yet it remains difficult to predict how metabolic systems will respond to perturbation. That is, if the activity or expression level of a particular enzyme were changed, then which (if any) of the other enzymes in metabolism would require compensatory modification? Because the pattern of such adaptive interactions

between proteins remains largely unknown, our ability to rationally engineer new systems (Kim and Copley, 2012; Michener et al., 2014a, 2014b), understand how the cell responds to perturbations (Kim et al., 2010; Long et al., 2018), and quantify the relation between mutations and disease (Kondrashov et al., 2002; Zuk et al., 2012) is limited. An ability to globally map dependencies between proteins and to identify groups of enzymes within a pathway that adapt as distinct subunits would help render cellular systems more tractable and predictable.

To begin to address this problem, we used a combination of comparative genomics and experiments to measure the pattern of adaptive interactions with the folate metabolic enzyme dihydrofolate reductase (DHFR). DHFR is an essential enzyme involved in the synthesis of purine nucleotides, thymidine, and several amino acids (Green and Matthews, 2007). As such, it is a common target of antibiotics, antimalarials, and chemotherapeutics (Ducker and Rabinowitz, 2017; Gangjee et al., 2007). In recent years, it has become a prominent model system for understanding the evolution of drug resistance (Costanzo and Hartl, 2011; Ogbunugafor et al., 2016; Palmer et al., 2015; Rodrigues et al., 2016; Toprak et al., 2011), the evolution of protein conformational dynamics (Bhabha et al., 2013; Francis et al., 2013; Liu et al., 2013), and constraints on horizontal gene transfer (Bershtein et al., 2015; Bhattacharyya et al., 2016). However, it remains unclear how changes to other enzymes in folate metabolism influence the relation between DHFR function and cellular fitness. Understanding how mutations in other enzymes can compensate for the loss of DHFR function is important to questions of antibiotic resistance and the evolution of folate metabolism. More generally, DHFR provides a well-studied model system to test computational approaches for predicting adaptive interactions.

Examining patterns of evolutionary coupling across species presents a practical strategy for inferring genes that share an adaptive constraint. We used analyses of both synteny (conservation of chromosomal proximity) and gene co-occurrence across 1,445 bacterial genomes to create a statistical map of evolutionary couplings within folate metabolism. Although the folate biochemical pathway map is highly interconnected (many enzymes share a metabolic intermediate), our comparative genomics analyses suggest a sparse architecture of adaptive interactions in folate metabolism. In particular, the enzymes DHFR and thymidylate synthase (TYMS) are strongly evolutionarily coupled to each other, and less so with the remainder of the pathway. To experimentally examine adaptive dependencies, we used CRISPR interference (CRISPRi), and forward evolution in *Escherichia coli* to map couplings to both DHFR and TYMS. We demonstrate that these two enzymes are (1) coupled through a shared metabolite, (2) are relatively less coupled to the remainder of folate metabolism, and (3) can adapt independently from the remainder of the genome. Extending our statistical analyses of co-evolution genome-wide reveals additional gene pairs that co-evolve with one another, yet are relatively independent from the rest of the genome. This provides a set of hypotheses for future experimental testing, and suggests that the finding of small adaptive units within larger cellular systems may be more prevalent.

# RESULTS

## A Comparative Genomics-Based Map of Evolutionary Coupling in Folate Metabolism

Starting from DHFR (encoded by the *E. coli* gene *folA*), we selected 16 folate metabolic genes for study using information from the STRING (search tool for recurring instances of neighboring genes) database (STRINGdb; version 10.5; Szklarczyk et al., 2015) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps (Kanehisa et al., 2012). STRINGdb is an online database of known and predicted protein interactions that uses a combination of information, including synteny, co-occurrence, online databases, co-expression, high-throughput experiments (e.g., yeast two-hybrid data), and automated text mining, to produce an amalgamated confidence score (Szklarczyk et al., 2015). This amalgamated score is benchmarked to predict the likelihood that two enzymes are in the same KEGG pathway. We selected all highest-confidence interactions to DHFR (encoded by the *E. coli* gene *folA*). This resulted in 14 genes, all of which encode an enzyme sharing a product or substrate with DHFR. We added two more genes (*metF* and *folD*) to this set to complete the folate cycle on the basis of KEGG pathway information.

This selection of genes encompasses core folate metabolism, a set of reactions that interconvert various folate species and produce methionine, glycine, thymidine, and purine nucleotides in the process (Figure 1A; Table S1) (Green and Matthews, 2007). The input of the pathway is 7,8-dihydrofolate (DHF), produced by the bifunctional enzyme dihydrofolate synthase-folylpolyglutamate synthetase (FPGS) through the addition of L-glutamate to dihydropteroate. Once DHF is formed, it is reduced to 5,6,7,8-tetrahydrofolate (THF) by the enzyme DHFR using the reduced form of nicotinamide adenine dinucleotide phosphate (NADPH) as a co-factor. THF can then be modified by a diversity of one-carbon groups at both the N-5 and N-10 positions, and subsequently serves as a one-carbon donor in several critical reactions, including the synthesis of a few amino acids and purine nucleotides (Figure 1A, bottom square portion of the pathway). The only step that oxidizes reduced folate (5,10-methylene THF) back to DHF is catalyzed by the enzyme TYMS, which modifies uridine monophosphate (dUMP) to thymidine monophosphate (dTMP) in the process.

Consistent with the stated goal of the STRINGdb, to recover biochemical pathways, the STRINGdb confidence scores indicate coupling between nearly all pairs of enzymes in folate metabolism. Both STRINGdb and the associated pathway map suggest that folate metabolism is highly interconnected, with many enzymes interacting via a shared product or substrate (Figures 1A and S1A). To examine the extent to which these biochemical interactions lead to evolutionary couplings, we focused our analysis on measures that report on evolutionary selection across organisms. For a dataset of 1,445 bacterial genomes, we computed (1) synteny, the conservation of chromosomal proximity between genes, and (2) co-occurrence, the coordinated loss and gain of genes across species (Junier and Rivoire, 2013, 2016; Rivoire, 2013). Both approaches have been established as reliable indicators of protein functional relations (Beck et al., 2018; Huynen et al., 2000; Janga et al., 2005; Junier and Rivoire, 2016; Kim and Price, 2011; Pellegrini et al., 1999; Snel et al., 2002), and technical variations on these methods are components of the amalgamated STRINGdb score.

These approaches are expected to report on any interaction important to evolutionary fitness and can thus indicate gene pairs that interact through various mechanisms, including physical binding or biochemical constraints.

In contrast to the amalgamated STRINGdb score, analyzing these two signals individually results in a sparse pattern of evolutionary couplings wherein most folate metabolic genes are relatively independent from one another (Figures 1B, 1C, S1B, and S1C). Consistent with expectations, we observe evolutionary couplings between physically interacting gene products: the glycine cleavage system proteins H, P, and T (*gcvH*, *gcvP*, and *gcvT* in *E. coli*) form a macromolecular complex (Okamura-Ikeda et al., 1993). We also see evolutionary couplings for the following three enzyme pairs: (1) DHFR-TYMS, (2) methionine synthase (MS) with methionine tetrahydrofolate reductase (MTHFR), and (3) the purine biosynthesis proteins PGT and IMPS. While these enzyme pairs are not known to physically bind, each pair of enzymes shares a metabolic intermediate, suggesting that a biochemical mechanism underlies their evolutionary coupling. However, most enzymes that catalyze neighboring reactions do not show statistical correlation (Figures 1 and S1). Instead, synteny and co-occurrence provide a different representation of folate metabolism in which most genes are relatively independent and a few interact to form evolutionarily constrained units.

What factors could contribute to the observed sparsity of evolutionary couplings? One potential contributor is limited sensitivity of our methods. False-negatives, gene pairs that are evolutionarily coupled but without a strong synteny or co-occurrence signal, are expected due to limited statistical power. Thus, perhaps the apparent modularity of the DHFR-TYMS pair within folate metabolism simply reflects an incomplete pattern of predicted interactions. We find only 6 evolutionary couplings of 120 possible pairwise interactions, even though 83 enzyme pairs interact biochemically (by sharing a product or substrate). Thus, a high false-negative rate would be necessary to completely explain the difference between the pattern of biochemical interactions and the pattern of evolutionary couplings.

More generally, the pattern of adaptive interactions between enzymes need not strictly resemble the pattern of biochemical interactions, given the non-linear relation between enzyme activities, metabolite concentrations, and fitness. The observed sparsity may then reflect a modular organization of adaptive constraints between enzymes, shaped by evolution. The evolutionary couplings are a statistical inference across thousands of genomes, and are not expected to capture idiosyncratic couplings specific to a particular organism or environmental condition. Instead, we hypothesize that the evolutionary coupling map presents predictions of the major adaptive couplings (and of the degree of independence). If true, then this would suggest a route to begin decomposing metabolic pathways into meaningful, smaller multi-gene units, in a way that does not depend strongly on the choice of model organism or environment.

As a first step toward testing this broader hypothesis, we set out to experimentally test whether the statistical pattern of evolutionary coupling observed for DHFR and TYMS corresponds to the pattern of adaptive coupling in an individual organism, namely *E. coli*. In particular, are DHFR and TYMS more coupled to each other than to other enzymes in the

pathway, as the statistical results suggest? And can perturbations in TYMS suffice to compensate for perturbations in DHFR (without the need to modify other enzymes)? These proteins are not known to physically interact, providing an opportunity to better understand how biochemical constraints may drive co-evolution. Although the genes encoding DHFR and TYMS are proximal along the chromosome of many bacterial species, they are approximately 2.9 Mb apart in *E. coli*. Thus, our experiments will determine whether adaptive coupling between genes found in synteny persists even when the genes are distant on the chromosome.

## The Growth Rate Defect of DHFR Knockdown Is Partially Rescued by Changes in the Expression of TYMS, but Not Other Folate Metabolic Enzymes

As a first measurement of adaptive coupling between DHFR, TYMS, and the rest of folate metabolism, we performed targeted measurements of growth rate dependency. More specifically, we decreased DHFR (or TYMS) expression, measured the effect on growth rate, and then examined whether the effect on growth rate was modified in the background of a second gene knockdown. The null hypothesis is that the growth rate effect of reducing expression for one gene should be the same regardless of whether the second gene is knocked down, unless the two genes are coupled.

To make these measurements, we used CRISPRi and a next-generation sequencing-based measurement of relative growth rate. In CRISPRi, a single-guide RNA (sgRNA) and a catalytically dead Cas9 enzyme are used to target and repress gene expression (Qi et al., 2013). Pairs of sgRNAs can be cloned into a single vector, permitting knock down of two genes at once (Figure S2A). Using this approach, we constructed pairwise knockdowns of DHFR and TYMS with every other gene in the folate metabolic pathway, as described in Figure 1, with the exception of *lpdA* and *fmt*. We excluded *lpdA* and *fmt* from our experiments because they are involved in other critical cellular processes and are expected to have pleiotropic effects (*fmt* is necessary to produce the initiator methionine tRNA required for all protein translation; *lpdA* binds to the glycine decarboxylase complex [GDC] but is also part of two other physical complexes in the cell). We also measured the effects of knocking down each gene individually by pairing each sgRNA with an sgRNA lacking the homology region needed for targeting a gene (referred to as *none*). In total, this results in a small library of 14 single and 25 double gene knockdowns (Tables S2 and S3). The knockdown activity of individual sgRNAs was verified by comparing the growth rate effect of knockdown to similar data for gene knockouts (data from the Keio collection; Figure S2C), and by qPCR (Figure S2D; Table S2).

To measure growth rate effects for the entire library, we transformed the library into MG1655 K12 *E. coli* and grew the cells in a turbidostat, a device for continuous culture that ensures that (1) the cells remain in the exponential growth phase for the duration of the experiment and (2) the media conditions are constant. We used M9 minimal media containing 0.4% glucose, 0.2% amicase, and 5 μg/mL thymidine for growth. These conditions were selected to be similar to previous forward evolution experiments (Toprak et al., 2011), with the addition of thymidine, and provide some measure of buffering for variation in the expression of folate metabolic enzymes. We took time points during the

course of 12 h, and then computed relative frequencies for each knockdown in the population by using next-generation sequencing to "count" the number of each sgRNA pair in the population (Figure S2B). By fitting a line to the plot of relative frequency over time, we obtain a relative growth rate (slope), indicating the per-hour decrease or increase in the abundance of a particular knockdown relative to wild type (carrying the targetless "*none*" sgRNA).

As expected, we observed that knocking down DHFR expression was highly deleterious. This was true whether alone or in combination with any other gene, with the exception of TYMS. Knocking down DHFR and TYMS together uniquely led to a partial rescue of growth rate (Figure 2A). Knocking down TYMS expression alone was also deleterious (although the effect is partly buffered by the presence of thymidine), and the deleterious effect was not improved by knocking down any other gene (Figure 2B). This indicates asymmetry in the coupling between DHFR and TYMS in this condition–knocking down TYMS buffers DHFR, but not vice versa. These results are consistent with prior attempts to generate *ΔfolA* strains of *E. coli*; when attempting to delete the gene encoding DHFR, a TYMS mutation is spontaneously acquired (Howell et al., 1988). Overall, the data show that among folate metabolic enzymes, decreasing TYMS expression uniquely compensates for reductions in DHFR expression.

## Forward Evolution Indicates Adaptive Independence of DHFR and TYMS from the Rest of the Genome

The above data are consistent with the idea that DHFR and TYMS are more adaptively coupled to each other than to other genes in folate metabolism, but they leave open the possibility that DHFR and TYMS interact with other genes in the genome. Moreover, the comparative genomics analyses of synteny and co-occurrence suggest that the coupling between DHFR and TYMS leads them to co-evolve as a unit. To look for compensation by other enzymes genome-wide and test the idea that DHFR and TYMS can adapt as an independent unit, we conducted a global suppressor screen. We inhibited DHFR with the common antibiotic trimethoprim and then examined the pattern of compensatory mutations with whole-genome sequencing. If DHFR and TYMS indeed represent a quasi-independent adaptive unit, then suppressor mutations should be found within these two genes (*E. coli folA* and *thyA*), with minimal contribution from other sites.

To facilitate forward evolution in the presence of trimethoprim, we used a specialized device for continuous culture called a morbidostat (Chevereau et al., 2015; Toprak et al., 2011, 2013). The morbidostat dynamically adjusts trimethoprim concentration in response to bacterial growth rate and total optical density, thereby providing steady selective pressure as resistance levels increase (Figures S3A–S3C). The basic principle is that cells undergo regular dilutions with fresh media until they hit a target optical density (OD = 0.15); once this density is reached, they are diluted with media containing trimethoprim until the growth rate is decreased. This approach makes it possible to obtain long trajectories of adaptive mutations in the genome with sustained phenotypic adaptation (Toprak et al., 2011). For example, in a single 13-day experiment, we observe resistance levels in our evolving bacterial populations that approach the trimethoprim solubility limit in minimal (M9) media.

As for the CRISPRi-based growth rate measurements, we conducted the forward evolution experiments in M9 minimal media supplemented with 0.2% amicase. We evolved three populations in parallel; in each case, the media was supplemented with a different concentration of thymidine (5, 10, and 50 μg/mL) chosen to range from a condition in which TYMS loss-of-function is deleterious to one that fully rescues TYMS activity. By alleviating selective pressure on the entire pathway, we sought to expose a larger range of adaptive mutations without biasing the pathway toward a particular result; this is similar to the common practice of conducting second site suppressor screens for essential genes under relatively permissive conditions (Forsburg, 2001). In this context, an absence of mutation outside the two-gene pair becomes more significant.

During 13 days of evolution, we observed that the 3 populations steadily increased in trimethoprim resistance (Figures S3D and S4). To identify the mutations underlying this phenotype, we selected 10 single colonies from the endpoint of each of the 3 experimental conditions for phenotypic and genotypic characterization (30 strains in total). For each strain, we measured the trimethoprim half-maximal inhibitory concentration ($IC_{50}$), the growth rate dependence on thymidine, and conducted whole-genome sequencing (Figure 3; Tables S4, S5, S6, and S7). All 30 strains were confirmed to be highly trimethoprim resistant (Figure 3; Table S4), with the exception of strains 4, 5, and 10 from the 50 μg/mL thymidine condition. These 3 strains grew very slowly at all concentrations of trimethoprim tested (up to 1,500 μg/mL). Because the dependence of growth on trimethoprim did not fit a sigmoidal function for these three strains, we do not report an $IC_{50}$. Furthermore, strains from all three thymidine-supplemented conditions became dependent on exogenous thymidine for growth, indicating a loss of function in the *thyA* gene that encodes TYMS (Figure 3). We confirmed that this loss of function is not a simple consequence of neutral genetic variation in the presence of thymidine; cells grown in 50 μg/mL thymidine and 0.2% amicase in the absence of trimethoprim retained TYMS function over similar timescales (Figure S5).

Whole-genome sequencing for all 30 strains revealed that isolates from all 3 conditions acquired coding-region mutations in both DHFR and TYMS, or even solely in TYMS (Figure 3). The sequencing data for strains 4, 5, 7, and 10 (in 50 μg/mL thymidine) did not indicate any mutations or amplifications to DHFR or the promoter. For all of the other strains, the mutations identified in DHFR reproduce those observed in an earlier morbidostat study of trimethoprim resistance (Toprak et al., 2011). The mutations in TYMS–two insertion sequence elements, a frameshift mutation, loss of two codons, and a non-synonymous active site mutation–are consistent with loss of function. Thus, the observed pattern of mutation is consistent with our CRISPRi experiment: reduced TYMS activity can buffer the inhibition of DHFR.

However, are mutations in DHFR and TYMS sufficient to explain the resistance phenotype of the evolved strains? We did not observe mutations in any other folate metabolic genes during the course of the experiment, but we did observe that a few other genes mutated under multiple thymidine conditions (Figure 3; Table S6). These mutations may be neutral, contribute to trimethoprim resistance, or be adaptive for growth in continuous culture. Of particular interest were recurring mutations in *gadX*, a regulator of the acid response system, and the insertion sequences (IS1) in *dusB*. Trimethoprim is known to induce an acid

response in *E. coli* that results in the upregulation of the *gadX* target genes *gadB/C* (Mitosch et al., 2017), and *dusB* is located in an operon upstream of the global *E. coli* transcriptional regulator *fis* (Bradley et al., 2007). Thus, we wanted to test whether mutations in DHFR and TYMS alone were sufficient to explain the resistance phenotype. We introduced several of the observed DHFR and TYMS mutations into a clean wild-type *E. coli* MG1655 background (referred to as strains R1–R4) and measured the $IC_{50}$. Strain R1 contained only a TYMS mutation (a deletion of 6 nucleotides, removing residues 25–26). As for strains 4, 5, and 10 (in 50 μg/mL thymidine), strain R1 grew very slowly at all concentrations of trimethoprim measured (up to 1,800 μg/mL), preventing the determination of a reliable $IC_{50}$. Strains containing mutations in both DHFR and TYMS (R2–R4) showed resistance phenotypes that are comparable to the evolved strains (Figure 3). For example, the DHFR L28R/TYMS Δ6(64–69) double mutant in a clean genetic background (strain R4) was more resistant to trimethoprim than all of the evolved strains containing these two mutations (Figure 3A). For comparison, we also measured the $IC_{50}$ for 3 previously constructed strains containing the DHFR single mutants (P21L, L28R, and W30R) (Palmer et al., 2015). The DHFR mutations alone were uniformly less trimethoprim resistant than the corresponding DHFR-TYMS double mutants. We therefore conclude that the paired DHFR and TYMS mutations are both necessary and sufficient to generate the observed resistance phenotype. Consistent with this, one of the evolved strains contains only mutations in DHFR and TYMS (strain 1 in 50 μg/mL thymidine; Figure 3C).

TYMS loss-of-function mutations have been observed in trimethoprim-resistant clinical isolates for multiple genera of bacteria (King et al., 1983; Kriegeskorte et al., 2014). This provides a degree of confidence that the laboratory evolution experiment reflects the natural conditions of adaptation. The data support the idea that DHFR and TYMS represent a two-gene adaptive unit, with compensatory mutations contained within the unit.

## DHFR and TYMS Are Coupled by Constraints on Metabolite Concentration

Given the above data and no evidence for the physical association of DHFR and TYMS in bacteria, what is the mechanism underlying their adaptive coupling? The CRISPRi experiments and second site suppressor screen indicate that reductions in DHFR activity, either from reducing expression, or inhibiting catalysis, are buffered by decreases in TYMS activity. This suggests a simple hypothesis—that their coupling arises from the need to balance the concentration of key metabolites in the folate metabolic pathway. In particular, both the depletion of reduced folates (THF species) and the accumulation of DHF are expected to be detrimental to cell growth (Kwon et al., 2008, 2010). Therefore, we wanted to test the relation between the activity of DHFR, the activity of TYMS, metabolite abundance, and growth.

We started with a panel of 10 DHFR mutations selected to span a range of *in vitro* catalytic activities (Reynolds et al., 2011) and measured the effect of these mutations on *E. coli* growth in the background of either wild-type (WT) TYMS or TYMS R166Q, a catalytically inactive variant. Again, we used a next-generation sequencing-based assay to measure the relative fitness of all of the possible mutant combinations (20 total) (Reynolds et al., 2011). In this system, DHFR and TYMS were expressed from a single plasmid that contains two

DNA barcodes, one associated with DHFR and one with TYMS, that uniquely encode the identity of each mutant (Figure S6A). The full library of mutants was transformed into the thymidine auxotroph strain *E. coli* ER2566 *ΔfolA ΔthyA*, and grown as a mixed population in a turbidostat. Counting the frequency of the barcodes over time permits the estimation of a relative growth rate for each mutant in the population (Figure S6B).

Relative growth rates were measured across two thymidine supplementation conditions: one in which TYMS R166Q produces a growth defect (5 μg/mL thymidine) and one that constitutes a full rescue of TYMS activity (50 μg/mL). In the background of WT TYMS (gray points, Figures 4A and 4B), the growth rate was relatively insensitive to an ~10-fold decrease in DHFR activity. As DHFR activity was diminished further, the relation between catalytic power and growth rate became monotonic; decreases in DHFR activity result in slower growth. In the background of TYMS R166Q, decreases in DHFR catalytic activity were buffered. More precisely, we observed that the TYMS R166Q mutation is deleterious or neutral in the context of WT DHFR but is beneficial in the context of low-activity DHFR mutants (e.g., F31Y.G121V). In high thymidine conditions, this effect was very pronounced. The TYMS R166Q mutation was sufficient to rescue growth to WT-like levels for DHFR mutants with orders of magnitude less activity (Figure 4B).

From this set of mutants, we selected 10 DHFR-TYMS pairs for liquid chromatography-mass spectrometry (LC-MS) profling of folate pathway metabolites. The experiment was carried out for log-phase cultures in M9 glucose media supplemented with 0.1% amicase and 50 μg/mL thymidine, conditions in which the selected DHFR mutations displayed significant growth defects individually but in which the corresponding DHFR-TYMS double mutants are restored to near-WT growth. Current mass spectrometry methods allow discrimination between the full diversity of folate species, which differ in oxidation, one-carbon modification, and polyglutamylation states, permitting a broad metabolic study of the effects of mutations (Lu et al., 2007).

The data confirm that for DHFR loss-of-function mutants, intracellular DHF concentration increases (Figure 4C, bottom four rows). In addition, we observe the depletion of reduced polyglutamated folates (Glu ≥ 3), while several mono- and diglutamated THF species accumulate (particularly for THF, methylene THF, and 5-methyl THF). Prior work found that DHF accumulation results in the inhibition of the upstream enzyme folylpoly-γ-glutamate synthetase (FP-γ-GS) (Kwon et al., 2008). FP-γ-GS catalyzes the polyglutamylation of reduced folates, an important modification that increases folate retention in the cell and promotes the use of reduced folates as substrates in a number of downstream reactions (McGuire and Bertino, 1981). The observed pattern of changes in the reduced folate pool (depletion of polyglutamated THF forms) is consistent with inhibition of FP-γ-GS by DHF.

In the background of the corresponding TYMS loss-of-function mutant, the metabolite profiles become more WT-like; the depletion of reduced folates is less severe, and the accumulation of DHF is more moderate (Figure 4C). Overall, the changes in metabolite concentration are consistent with the observed growth rate defects in the DHFR loss-of-function mutants; mutants with more severe DHF accumulation and THF depletion grow

more slowly (Figures 4D and S7). Thus, coordinated decreases in the activity of DHFR and TYMS serve to maintain balance in the DHF and THF metabolite pools, a condition that is associated with optimal growth. This provides a plausible biochemical mechanism for the strong statistical association of DHFR and TYMS by synteny and co-occurrence across thousands of bacteria. In this sense, the bifunctional fused form of DHFR-TYMS found in protists and plants could be regarded as an extreme case that guarantees stoichiometric expression (Beverley et al., 1986; Lazar et al., 1993).

## Genome-wide Analyses of Co-evolution Identify Additional Adaptive Pairs

To examine the possibility of small adaptive units such as DHFR and TYMS elsewhere in metabolism, we extended our analyses of gene synteny and co-occurrence to include all of the genes that are represented in *E. coli*. To ensure good statistics, we filtered the orthologs analyzed to those that co-occur in a sufficiently large number of genomes (2,095 Clusters of Orthologous Groups [COGs], ~500,000 pairs in total) (see also Quantitative and Statistical Analysis). We also compared our analysis to the following: (1) metabolic annotations from KEGG (Kanehisa et al., 2012) and (2) the set of high-confidence binding interactions in *E. coli* reported by the STRINGdb (Szklarczyk et al., 2015). Consistent with intuition and prior work, co-evolving gene pairs show enrichment for physical complexes, enzymes in the same metabolic pathway, and, more specifically, enzymes with a shared metabolite (Figures 5A and 5B).

To identify candidate adaptive units composed of two genes, we constructed a scatterplot indicating the strength of coupling within a gene pair (as a relative entropy, along the x axis) versus the strongest coupling outside the pair (along the y axis) (Figures 5C and 5D). This analysis is limited to the identification of two gene units, and our analysis will need to be extended to identify larger communities within the complete network of co-evolutionary relations (Newman, 2010). It presents a simple graphical method for identifying gene pairs that are more tightly coupled to each other than to any other gene in the dataset. These are the set of points below the diagonal (Table S8). This results in 258 two-gene co-evolving units by synteny and 194 by co-occurrence, including the DHFR-TYMS pair. Many of the predicted adaptive units are two-protein physical complexes, while others share a metabolic intermediate such as DHFR-TYMS.

In this plot, distance from the diagonal indicates the degree of modularity of a pair. The most modular pairs occupy the space in the lower right corner of the graph. By this criterion, DHFR and TYMS show an exceptionally strong signature of modularity. Considering the regime in which $D_{ij}^{\text{intra}} > 1.0$ and $D_{ij}^{\text{exter}} < 0.5$ on the synteny plot, we observe a few other pairs with experimental evidence for adaptive coupling–for example, the gene pair *accB/accC*, which encodes two of the four subunits of acetyl-coenzyme A (CoA) carboxylase, the first enzymatic step in fatty acid biosynthesis. Overexpression of either *accB* or *accC* individually causes reductions in fatty acid biosynthesis, but overexpressing the two genes in stoichiometric amounts rescues this defect (Abdel-Hamid and Cronan, 2007; Janßen and Steinbüchel, 2014). Constraints on relative expression have also been noted for the *selA/selB* and *tatB/tatC* gene pairs (Bolhuis et al., 2001; Rengby et al., 2004). The *tatB/tatC* genes encode components of the TatABCE twin-arginine translocation complex, while *selA/selB*

are not known to bind but are both involved in selenoprotein biosynthesis. The full set of gene pairs below the diagonal now serve as a starting place for a deeper understanding of the hierarchical pattern of evolutionary couplings inside cellular pathways and for testing the prevalence of adaptive modules more generally (Table S8).

## DISCUSSION

Comparative genomics analyses of synteny and co-occurrence are well-established tools for the prediction of interactions between genes (Huynen et al., 2000; Janga et al., 2005; Junier and Rivoire, 2016; Kim and Price, 2011; Pellegrini et al., 1999; Snel et al., 2002). However, they are not typically used to infer independence. Here, we effectively ask whether, for evolutionary couplings between genes, the absence of evidence can be evidence for absence. Conceptually, this is analogous to recent work in proteins, in which co-evolution was used to map quasi-independent adaptive and functional residue groups within a single protein (Halabi et al., 2009). In our case study of folate metabolism, comparative genomics suggested a sparse pattern of evolutionary coupling between folate metabolic enzymes, with DHFR and TYMS showing significant coupling to each other but less coupling to the remainder of the folate metabolic pathway. Experimental analyses support the interpretation that these two enzymes form a relatively independent adaptive unit. In this case, therefore, comparative genomics identified a meaningful adaptive unit that is much smaller than the scale of the entire biochemical pathway. In light of these findings, it becomes less obvious that shared membership in a KEGG metabolic pathway should be a default gold standard for interaction prediction approaches; instead, the goals of the method and type of interaction being predicted (shared biochemical intermediate, physical complex, or adaptive interaction) should be carefully considered.

DHFR and TYMS represent a first case study, and comparative genomics approaches are not expected to predict all of the possible adaptive interactions. Analyses of synteny and co-occurrence are an inference across thousands of species, encapsulating hundreds of millions of years of evolutionary divergence. In the context of specific environmental conditions (including conditions never encountered in evolutionary history) or other model organisms, additional adaptive interactions likely exist. We expect that the adaptive units predicted by synteny and co-occurrence represent well-conserved, core, adaptive interactions that can sometimes be elaborated on under particular environmental conditions or in particular species. We propose that if a perturbation is made in one gene of the adaptive unit, the first and most common adaptive mutations will also occur within the unit (either within the same gene or in the other gene in the pair). Furthermore, we expect that mutations within the pair should largely suffice to restore function, with mutations outside the pair having more subtle and/or idiosyncratic (e.g., environment or species specific) effects.

If it is generally possible to decompose metabolic pathways into smaller, relatively independent pieces, then this would suggest a route to identify units for biosynthetic engineering and provide fundamental insights into how cells maintain homeostasis and adapt in the face of changing environments. For example, thymidine synthesis is the rate-limiting step for DNA synthesis in eukaryotic cells, and transcription of the TYMS and DHFR genes is greatly upregulated (via a common transcription factor) at the $G_1$/S cell-cycle transition

(Bjarnason et al., 2001). In computational models of eukaryotic folate metabolism, computationally increasing the activities of DHFR and TYMS 100-fold results in increased thymidine synthesis, but it only modestly changes the concentration of other folates. Thus, decoupling the DHFR-TYMS pair from the remainder of the folate metabolism may be important for ensuring modularity in different metabolite pools. Substantial further work is needed to go beyond the DHFR-TYMS pair, to comprehensively test the relation between the units identified by comparative genomics, patterns of adaptive dependencies, and adaptation to environmental changes. The data presented here provide the computational hypotheses, technical approaches, and motivation for now doing so.

## STAR★METHODS

### CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Kimberly Reynolds (kimberly.reynolds@utsouthwestern.edu).

### EXPERIMENTAL MODEL AND SUBJECT DETAILS

The parent strain of the forward evolution experiment was *E. coli* MG1655 modified by phage transduction to encode green fluorescent protein (*egfp*) and chloramphenicol resistance (*cat*) at the P21 attachment site. This strain was also used in the construction of 'reconstitution' genotypes and the control experiment of cell culture in the turbidostat without trimethoprim selection. Additional *E. coli* strains and recombinant DNA are described in the relevant Method Details sections.

### METHOD DETAILS

**sgRNA library construction—**CRISPRi sgRNAs were designed to target 14 genes in *E. coli* central metabolism (*folA*, *thyA*, *folC*, *glyA*, *gcvH*, *gcvP*, *gcvT*, *metH*, *metF*, *folD*, *purU*, *purN*, *purT*, and *purH*). All the genes except for *gcvP*, *gcvT*, and *gcvH* are in different operons, ensuring separate targeting. For each gene, we selected a 20bp homology region (preceding a PAM site) on the non-template strand of the 5′ end of the gene for targeting. We used BLAST to confirm that each guide lacks significant homology to any other locus in the genome. See Table S3 for sequences and PAM sites. The individual sgRNAs were cloned into a modification of the pCRISPR plasmid (Addgene plasmid #42875) (Jiang et al., 2013). The modified pCRISPR plasmid was made by removing the original sgRNA sequences by restriction digest and replacing them with a small DNA sequence containing two BsaI sites for golden gate cloning, creating pAM-111. The sgRNA (and corresponding promotor) from the pgRNA vector (Addgene plasmid #44251) was then inserted into pAM-111 using golden gate cloning (creating pAM-112). From this point forward, making a new sgRNA was done by inserting a new homology region into pAM-112 using iPCR as described (Hawkins et al., 2015).

To achieve pairwise knockdowns, we used golden gate assembly to combine the sgRNA inserts for folA (or thyA) with all other genes in folate metabolism or with a negative control guide (none) which lacks a targeting homology region. The primers GG-sgRNA1-F/R or GG-sgRNA2-F/R were used to amplify the sgRNA inserts with BsaI sites from pAM-112

for the golden gate reaction (Table S2). The inserts were then ligated into three vectors (pAM-288, pAM-289, and pAM-290) using three separate golden gate reactions. pAM-288, 289, 290 are nearly identical to pAM-111 with two important differences. First, orthogonal primer sites were inserted into this vector flanking the golden gate site. These orthogonal primers allow specific amplification of plasmids that have the golden gate BsaI sites and do not amplify the pAM-112 parent plasmid which carries over from the sgRNA insert amplification. Second, pAM-288, 289, 290 each contain a unique 6N barcode (CTTTCA, ATCATG, GCATGG) directly downstream of the sgRNAs. Thus, following assembly, we obtain three small libraries of 39 unique knockdowns, each with a unique barcode. Because the sgRNAs can be cloned in either order in the plasmid, we end up with six total replicates of each gRNA pair in our full library (three barcodes in two orders). This allows us to calculate the error across internal replicates in a single experiment, and potentially exclude so-called “escapers.” The final library was transformed into MG1655 *E. coli* with dCas9 in the chromosome (gift from Bikard lab) and glycerol stocked (gAM-350).

**Fitness assay and next generation sequencing (NGS) of sgRNA library—** Immediately following transformation, the cells were grown over night at 37°C shaking in a SOB LB mix containing 35 μg/mL Kanamycin, folA mix (38 μg/mL glycine, 75.5 μg/mL methionine, 1 μg/mL pantothenate, 20 μg/mL adenosine), and 50 μg/mL thymidine. These additives relieve selection pressure on genes targeted by CRISPRi during outgrowth in case of leaky dCas9 expression. In the morning, the culture was centrifuged and washed 2 times with M9 minimal media pH 6.5, 0.4% glucose, 35 μg/mL Kanamycin, 5 μg/mL thymidine, and 0.2% amicase (Sigma, Cat#65072-00-6). The cells were adapted to this media for 12 hours at 30°C shaking. The cultures were then back diluted to $OD_{600}$ = 0.05 and used to inoculate 15 mL cultures in our turbidostat at 30°C. Optical density was clamped at an OD600 of 0.15. Turbidostat input media was M9 minimal media pH 6.5,0.4% glucose, 35 μg/mL Kanamycin, 5 μg/mL thymidine, and 0.2% amicase. The culture was further adapted in the turbidostat for 8 hours, then CRISPRi was induced by the addition of 50 ng/mL anhydrous tetracycline to the media (aTC). Culture samples (1 mL each) for NGS were taken 3 hours, 5 hours, 7 hours, and 11 hours after CRISPRi induction. These samples pelleted by centrifugation, decanted, and frozen at −20°C. All samples were prepared for amplicon sequencing using two rounds of PCR to add standard Illumina TruSeq and i5/i7 sequencing primers. Each time point was assigned a unique i5/i7 combination for demultiplexing after sequencing. PCR amplicons for each sample (time point) were quantified using a picogreen assay and then mixed in equal stoichiometric ratio. This mix was run on an agarose gel and the band corresponding to the expected size (~550 BP) was excised. The resulting amplicon mixture was sequenced on an Illumina MiSeq using a 300 cycle V2 kit (Illumina, Cat#MS-102-2002) configured to read 175 base pairs (read 1) and 125 base pairs (read 2) to sequence the barcode and both sgRNA homology regions. MiSeq quality: reads passing Q30 = 96.4%, 818K/$mm^2$ cluster density, and 15,384,044 reads passing filters.

**q-PCR methods for CRISPRi sgRNA verification—**CRISPRi gene knockdowns were induced in MG1655 *E. coli* growing in M9 minimal media (0.4% glucose, 35 μg/mL Kanamycin) at 37°C with 50 ng/μl aTC for four hours prior to RNA extraction. Total RNA

extractions were performed using the Purezol reagent (Bio-Rad, Cat#7326890) following manual instructions. RT-qPCR was performed using the Luna Universal One-Step RT-qPCR kit (NEB, Cat#E3005S) and a CFX384 Real-Time System, both following manual instructions. The hcaT gene was used as the house keeping control. Every reaction was performed in technical triplicate except the glyA knockdown measurement, which was performed in duplicate. Melt curves were performed for all qPCR primer pairs to confirm single product amplification. However, if in later experiments one of the technical replicates had a multi-modal melt curve, it was removed. Relative mRNA expression was calculated using the $2^{-\Delta\Delta Ct}$ method, and standard error of the mean from technical replicates was propagated using a first order Taylor Series expansion.

**Forward evolution of trimethoprim resistance in the morbidostat—**The morbidostat/turbidostat apparatus was constructed as described by Toprak et al. (2013). The parent strain was *E. coli* MG1655 with a chromosomal *egfp/cat* resistance cassette. Throughout the forward evolution experiment, cells were grown at 30°C in M9 media supplemented with 0.4% glucose and 0.2% amicase; 30 μg/mL of chloramphenicol was added for positive selection.

To begin the experiment, the parent strain was cultured overnight at 37°C in Luria Broth (LB) + 30 μg/mL Cam. This culture was washed twice with M9, and back diluted into M9 + 30 μg/mL Cam supplemented with 5, 10, or 50 μg/mL thymidine (thy) for overnight adaptation in culture tubes at 30°C. The next day (henceforth referred to as day 0; day 1 is the end of the first day of adaptation), these overnight cultures were streaked onto LB agar plates: two colonies per condition were chosen for whole genome sequencing (WGS) in order to obtain an accurate sequence for the parent strain. The remainder of the overnight cultures was used to inoculate three morbidostat tubes at containing M9 media with varying thymidine supplementation (5, 10, and 50 μg/mL thy). The starting optical density was approximately 0.005. Initial antibiotic concentrations were 0, 11.5 and 57.5 μg/mL trimethoprim for media stocks A, B, and C respectively. Each culture grew unperturbed until it surpassed an $OD_{600}$ of 0.06, at which point it underwent periodic dilutions with fresh media. The dilution rate is given by the formula

$$r_{dil} = f \ln \frac{V}{V + \Delta V}$$

where $V = 15ml$ is the culture volume, and $\Delta V = 3ml$ is volume added. We chose a dilution frequency $f = 3h^{-1}$, to give $r_{dil} = 0.55$. Above the $OD_{600} = 0.15$, these dilutions are used to introduce TMP into the culture (see also Figure S3). This allows controlled inhibition of DHFR activity in response to growth rate. Cycles of growth and dilution continued for a period of ~22 hours, at which point the run was paused to make glycerol stocks, replenish media, and update TMP stock concentrations. Culture vials for the next day of evolution were filled with fresh media and inoculated using 300 μL from the previous culture. Complete trajectories of $OD_{600}$ and drug concentration are shown in Figure S4. Endpoint cultures were streaked onto LB agar plates supplemented with 30 μg/mL of Cam and 50 μg/mL thymidine to obtain isolated colonies for whole genome sequencing.

**Genome preparation and sequencing—**Two isolates were selected from each adapted day 0 culture, and ten clonal isolates (colonies) were randomly selected from the endpoint of each evolution condition, totaling 36 strains. Isolation of genomic DNA was performed using the QIAamp DNA Mini Kit (QIAGEN, Cat#51304). The Nextera XT DNA Library Prep Kit (Illumina, Cat#FC-131-1096) was used to fragment and label each genome for paired-end sequencing using a v2 300-cycle MiSeq kit (Illumina). Average read length and coverage can be found in Table S5.

**Measurement of growth in the absence of thymidine—**All strains were grown overnight in LB + 5 μg/mL thy, with the exception of those evolved in the 50 μg/mL thy condition, which were supplemented with 50 μg/mL thy to ensure viability. Overnight cultures were pelleted and washed twice into M9 media supplemented with 0.4% glucose and 0.2% amicase. They were then inoculated at $OD_{600}$ = 0.005 into 96-well plates containing the same media with either 0 μg/mL or 50 μg/mL thymidine. $OD_{600}$ was monitored in a Victor X3 plate reader at 30°C over a period of 20 hours.

**Measurement of growth in the presence of trimethoprim—**All strains were grown overnight in LB + 5 μg/mL thy, with the exception of the strains evolved in the 50 μg/mL thy, which were supplemented with 50 μg/mL thy to ensure viability. Each strain was pelleted and washed twice into M9 media supplemented with 0.4% glucose, 0.2% amicase, and the thymidine concentration matching their respective forward evolution condition (5, 10, and 50 μg/mL thy). The washed cells were back-diluted 1:10 and adapted for 5.5 hours at 30°C. The resulting cultures were then used to inoculate 96-well plates containing the same media along with serial dilutions of TMP (in triplicate). Three replicates were innoculated for each combination of strain and trimethoprim concentration at a starting $OD_{600}$ = 0.005. $OD_{600}$ was monitored using a Tecan Infinite M200 Pro microplate reader and Freedom Evo robot at 30°C over a period of at least 12 hours.

**Turbidostat culture without trimethoprim selection in 50 μg/mL thymidine—**The parent strain for this experiment was identical to that used for evolution of trimethoprim resistance. Throughout the experiment, cells were grown at 30°C in M9 media supplemented with 0.4% glucose and 0.2% amicase; 30 μg/mL of chloramphenicol (Michener et al.) was added for positive selection. To begin the experiment, the parent strain was cultured overnight at 37°C in Luria Broth (LB) + 30 μg/mL Cam. This culture was washed twice with M9, and back diluted into M9 supplemented with 50 μg/mL thymidine (thy) for overnight adaptation in culture tubes at 30°C. The next day (henceforth referred to as day 0; day 1 is the end of the first day of continuous culture), the overnight culture was used to inoculate three turbidostat tubes containing 17mL of M9 supplemented with 50 thy. The starting optical density was approximately 0.005. Each culture grew unperturbed until it reached an $OD_{600}$ of 0.15, at which point it was diluted with 2.4 mL of fresh media. These cycles of growth and dilution persisted for a period of ~22 hours, at which point the run was paused to make glycerol stocks and replenish media. Culture vials for each following day of evolution were filled with fresh media and inoculated using 300 μL from the previous culture.

**Fitness assay and next generation sequencing (NGS) of DHFR/TYMS point mutants—**All relative growth rate measurements were performed in the *E. coli* folate auxotroph strain ER2566 *ΔfolA ΔthyA* (Lee et al., 2008). DHFR (*folA*) and TYMS (*thyA*) are provided on the plasmid pACYC-Duet1 (in MCS1 and MCS2, respectively) and are each under control of a T7 promoter. For these experiments, we use leaky expression (no IPTG induction). Each mutant plasmid (20 in total) is marked with a genetic barcode in a non-coding region between the two genes. Plasmids were transformed into the auxotroph strain, and each mutant was grown overnight in separate LB +30 μg/mL Cam +50 μg/mL thy cultures. Then, cultures were washed 2x in M9 media supplemented with 0.4% glucose and 0.2% amicase and 30 μg/mL Cam, and adapted overnight at 30°C. All mutants were mixed in equal ratios based on $OD_{600}$ and inoculated at a starting $OD_{600}$ = 0.1 in the turbidostat. Growth rates were measured under two conditions: 5 thy and 50 thy, with three replicates each. The turbidostat clamps the culture to a fixed $OD_{600}$ = 0.15 by adding fresh dilutions of media. Every 2 hours over the course of 12 hours a 1 mL sample was removed, pelleted and frozen for next-generation sequencing. Amplicons containing the barcoded region with appropriate sequencing adaptors (350 basepairs in total size) were generated by two sequential rounds of PCR with Q5 polymerase. The barcoded region was sequenced with a single-end MiSeq run using a v2 50 cycle kit (Illumina, Cat#MS-102-2001). We obtained 14,348,937 reads.

**Constructing DHFR/TYMS mutants in a clean genetic background—**We followed the protocol for scarless genome integration using the modified λ-red system developed by Tas et al. (2015). In this method, a tetracycline (Tet) resistance cassette (“landing pad”) is first integrated at the site targeted for mutagenesis. Then, the landing pad is excised by the endonuclease I-SceI, and replaced with the desired mutation by λ-red mediated recombination. $NiCl_2$ is used to counterselect against cells that retain the tetracycline cassette. Tas et al. provides a detailed protocol; here we give the specifics necessary for our experiments. For the λ-red machinery, we transformed the plasmid pTKRED (Addgene plasmid #41062) (Kuhlman and Cox, 2010) into electrocompetent *E. coli* MG1655 with a chromosomal *egfp/cat* resistance cassette (the forward evolution parent strain). For the Δ25-26 TYMS mutation, we introduced the *tetA* landing pad between genome positions 2,964,900 and 2,965,201 (genome NC000913) corresponding to the N terminus of the *thyA* gene. For the DHFR mutations (L28R, W30R, and P21L), the landing pad was recombined between genome positions Addgene plasmid #4106249,684 and 49,990 (genome NC000913). In order to replace the Tet cassette, cells were induced with 2mM IPTG and 0.4% arabinose, and then transformed with 100ng of dsDNA PCR product containing the mutation of interest (with appropriate homology arms). This reaction experienced 3 days of outgrowth at 30°C in rich defined media (Teknova, Cat#M2105) with glucose substituted for 0.5% v/v glycerol. The media was supplemented with 6 mM or 4mM $NiCl_2$ for counterselection against *tetA* at the *thyA* locus or *folA* locus respectively. The outgrowth culture was streaked onto agar plates and screened daily for the mutant of interest using LB supplemented with 50 μg/mL thy, 30 μg/mL spectinomycin, and ± 5-10 μg/mL Tet. All mutations were confirmed by Sanger sequencing of the complete *folA* and *thyA* open reading frame; for *folA* the promoter region was also sequenced.

**LC-MS Metabolite Measurements—**Cells were cultured in M9 0.2% glucose media containing 0.1% amicase, 50 μg/mL thy, and 30 μg/mL Cam at 30°C for metabolite analysis. In mid-log phase at $OD_{600}$ ~0.2, *E. coli* culture (3 mL for nucleotide measurement and 7 mL for folate measurement) was filtered on a nylon membrane (0.2 μm), and the residual medium was quickly washed away by filtering warm saline solution (200 mM NaCl at 30'C) over the membrane loaded with cells to exclude non-desirable extracellular metabolites from LC-MS analysis. The membrane was immediately transferred to a 6 cm Petri dish containing 1 mL cold extraction solvent (−20°C 40:40:20 methanol/acetonitrile/water; for folate stability, 2.5 mM sodium ascorbate and 25 mM ammonium acetate in folate extraction solvent (Lu et al., 2007)) to quench metabolism. After washing the membrane, the cell extract solution was transferred to a microcentrifuge tube and centrifuged at 13,000 rcf. for 10 min. The supernatant was transferred to a new microcentrifuge tube. Folate samples were prepared with an additional extraction: the pellet was resuspended in the cold extraction solvent and sonicated for 10 min in an ice bath. After the second extraction and centrifugation, the supernatant was combined with the initial supernatant. The metabolite extracts were dried under nitrogen flow and reconstituted in HPLC-grade water for LC-MS analysis. Metabolites were measured using stand-alone orbitrap mass spectrometers (ThermoFisher Exactive and Q-Exactive) operating in negative ion mode with reverse-phase liquid chromatography (Lu et al., 2010). Exactive chromatographic separation was achieved on a Synergy Hydro-RP column (100 mm × 2 mm, 2.5 μm particle size, Phenomenex) with a flow rate of 200 μl/min. Solvent A was 97:3 $H_2O$/MeOH with 10 mM tributylamineand 15 mM acetic acid; solvent B was methanol. The gradient was 0 min, 5% B; 5 min, 5% B; 7 min, 20% B; 17 min, 95% B; 20 min, 100% B; 24 min, 5% B; 30 min, 5% B. Q-Exactive chromatographic separation was achieved on an Poroshell 120 Bonus-RP column (150 × 2.1 mm, 2.7 μm particle size, Agilent) with a flow rate of 200 μl/min. Solvent A is 10mM ammonium acetate + 0.1% acetic acid in 98:2 water:acetonitrile and solvent B is acetonitrile. The gradient was 0 min, 2% B; 4 min, 0% B; 6 min, 30% B; 11 min, 100% B; 15 min, 100% B; 16 min, 2% B; 20 min, 2% B. LC-MS data were analyzed using the MAVEN software package (Clasquin et al., 2012).

## QUANTIFICATION AND STATISTICAL ANALYSIS

**Comparative genomics dataset—**Calculations of both synteny and co-occurrence require a collection of genomes where individual genes are assigned into orthology classes. The Clusters of Orthologous Groups of proteins (COGs) defined by Koonin and colleagues provide one well-established set of ortholog annotations (Galperin et al., 2015). The results presented here use all complete and COG-annotated bacterial genomes available in the NCBI database as of March 2015 (1445 genomes and 4764 COGs, this dataset is also used in Junier and Rivoire (2016). A genome may contain more than one gene in the same COG, but for clarity, we start by presenting the calculations assuming that every orthology class maps to at most one gene in each genome.

**Counting pairs in co-occurrence—**To quantify co-occurrence, we began by counting genomes where orthology classes i and j co-occur. As previously published (Junier and Rivoire, 2016), we corrected for the uneven phylogenetic distribution of sequenced genomes (strains) by introducing genome weights. To this end, we computed a distance between each

pair of strains, based on the sequence similarity of a few conserved genes ($\delta_{gh} = 1 - S_{gh}$, where $S_{gh}$ is average sequence similarity). The weight $w_s$ of each strain $s$ is defined as $1/n_s$ where $n_s$ is the number of strains within a given distance $\delta$ of $s$. Varying $\delta$ can provide information at different "phylogenetic depths" (Junier and Rivoire, 2013), but here we fixed $\delta = 0.3$, our results being generally invariant to this value. The effective number of strains where orthology classes $i$ and $j$ co-occur is formally given by:

$$M_{ij} = \sum_s w_s \mathbb{1}[i \cap s \neq \varnothing] \mathbb{1}[j \cap s \neq \varnothing]$$

where the sum is over the strains $s$ and where $\mathbb{1}[X]$ is a generic indicator function with $\mathbb{1}[X] = 1$ if and only if X is true. Hence, $\mathbb{1}[i \cap s \neq \varnothing] = 1$ if $i$ is represented in strain $s$ and 0 otherwise.

**Defining gene proximity (for synteny)—**We measure the distance $d(i,j)$ between the midpoint of two genes $i$ and $j$ in base pairs (and set $d(i,j) = \infty$ if they are on different chromosomes). Given a circular chromosome of length $L$, the greatest possible distance between genes is $L/2$ (on opposite sides of the circle). Thus, given a null model in which genes are randomly distributed along the chromosome, the probability of finding the gene pair within a genomic proximity $d^*$ is just the normalized value $p^* = d^*/(L/2)$.

**Counting pairs in synteny—**The value $p^*$ provides a measure of significance for finding two genes at a distance $d^*$ in one genome. To determine the conservation of proximity across many species, we began by counting the effective number of strains in which i and j are within a given distance $d^*$:

$$X_{ij} = \sum_s w_s \mathbb{1}[d(i,j) < d^*]$$

However, because $p^* = (2d)/L$, the probability of finding two genes within distance $d^*$ depends on the chromosome length $L$, which varies between strains. In order to define a null model that is common to all strains, we instead considered the normalized distance and computed:

$$X_{ij} = \sum_s w_s \mathbb{1}[2d(i,j) / L_s < p^*]$$

For strains that contain multiple chromosomes, we took for $L_s$ the sum of the lengths of its different chromosomes. This corresponds to a null model where the genes are randomly shuffled within and between chromosomes (or, up to boundary effects, to concatenating all the chromosomes into a single one). We took $p^* = 0.02$, corresponding to $d = 50$ kb in the context of a chromosome of length 5 Mb. This cutoff was chosen to represent a length scale longer than those typical for gene co-expression and synteny, so that the choice of cutoff does not determine the results. Further, the results are robust with respect to the choice of $p^*$.

Finally, to account for the possibility that a single strain may contain multiple pairs of genes in two given orthology classes $i$ and $j$, we corrected $X_{ij}$ averaging over all these pairs:

$$X_{ij} = \sum_s \frac{1}{|\, i \cap s \,|\;|\, j \cap s \,|} \sum_{g_i \in i \cap s,\, g_j \in j \cap s} w_s \mathbb{1}[2d(g_{ji}, g_j) / L_s < p^*]$$

where $i \cap s$ is the set of genes in orthology class $i$ and in strain s and $|i \cap s|$ is the size of this set. This formula is simpler than the one used in Junier and Rivoire (2016) but leads to similar results.

**Measuring significance—**To assess the significance of finding a pair of genes in proximity $X_{ij}$ times out of $M_{ij}$ (given $p^* = 0.02$), we used the binomial distribution (the same as the "coin toss" problem in standard statistics):

$$\pi_{ij} = \sum_{K \geq X_{ij}} \binom{M_{ij}}{K} (p^*)^K (1 - p^*)^{M'_{ij} - K} = I(X_{ij}, M_{ij} - X_{ij} + 1, p^*)$$

where $I(a, b, x)$ is the regularized incomplete beta function. This relatively naive null model (which assumes a uniform distribution of genes along the chromosome, and treats weighted genomes as independent trials) provides a good description of the data for the majority of orthology class pairs - indicating that most gene pairs have no significant conservation of chromosomal proximity (Junier and Rivoire, 2016). A subset of pairs nevertheless deviate from the statistical expectations of the null model; these are the syntenic pairs of interest.

Finally, analysis of any large dataset inevitably leads to spurious false positives that simply occur by random chance. To account for this, we applied the Bonferroni principle – we set a threshold of significance to $\pi_{ij} = 2/N(N-1) \simeq 10^{-7}$ where $N = 4764$ is the number of orthology classes defined by COGs. We chose a cutoff such that we should not have found any significantly syntenic gene pair "by random" among all $10^7$ possible gene pairs. This criterion is very stringent, and may be relaxed to set instead a false discovery rate (Junier and Rivoire, 2016).

**Degree of synteny—**The p values $\pi_{ij}$ depend on the number of genomes in the dataset. It is more meaningful to define a measure of conservation the depends only on rescaled variables, here the frequencies $f_{ij} = X_{ij}/M_{ij}$. To ensure sufficient sample size, we restricted our analysis to cases where the number $M_{ij}$ of genomes where $i$ and $j$ is large. Here, we considered pairs of COGs with $M_{ij} \geq 100$. The degree of synteny is then given by the relative entropy:

$$D(f_{ij} \mid p^*) = f_{ij} \ln \frac{f_{ij}}{p^*} + (1 - f_{ij}) \ln \frac{1 - f_{ij}}{1 - p^*}$$

In the limit of large $M_{ij}$, $e^{-M_{ij}D_{ij}}$ approximates the first term of the sum in the equation and therefore $M_{ij}D_{ij}$ correlates with $-\ln \pi_{ij}$. The maximal value of $D_{ij}$ is set by $p^*$: as $p^* = 0.02$

corresponds to $-\ln p^* \simeq 4$, the range of values for $D_{ij}$ is thus [0, 4]. Finally, since $M_{ij} \geq 10^2$ and $\pi^* = 10^{-7}$, any value of $D_{ij}$ larger than $D^* = -(\ln 10^{-7})/10^2 \simeq 0.025$ reports significant synteny.

**Degree of co-occurrence—**Following the same logic, we define a similar measure for co-occurrence. We computed the probability of finding a given gene pair *i,j* together (in co-occurrence) $X_{ij}$ or more times in set of $M_{ij}$ genomes, given the null model that occurrence of each gene is independent ($f_{ij} = f_i f_j$). Starting from the binomial distribution function and taking Stirling's approximation, one obtains the mutual information as a measure of co-occurrence between genes:

$$D(f_{ij} \mid f_i f_j) = f_{ij} \ln \frac{f_{ij}}{f_i f_j} + (1 - f_{ij}) \ln \frac{1 - f_{ij}}{1 - f_i f_j}$$

**Application of comparative genomics analyses to *E. coli* genome—**To analyze synteny and co-occurrence relationships relevant to *E. coli*, we kept only the COGs *i* that are represented in its genome, and analyzed COG pairs for which $M_{ij} \geq 100$ (2095 COGs in total). In Figure 5, we plot for each pair *ij* of these COGs their degree of synteny (or co-occurrence) $D_{ij}$ (x axis) against their maximal degree of synteny (or co-occurrence) with any other COG $\max_{k \neq i,j} (D_{ik} D_{jk})$ (y axis). In this figure, we annotate physical interactions using the STRING "actions" database for *E. coli* (511145.protein.actions.v10.5.txt – this file reports the subset of STRING interaactions which correspond to physical binding) taking a threshold of 700 ("high confidence") and the largest score when multiple paralogs are present. To annotate genes within the same pathways and enzymes pairs sharing an intermediate, we used KEGG pathways/reactions. Because a number of the KEGG pathways are extremely general, we removed the four most inclusive pathways in consideration of "shared pathway" status: "Metabolic Pathways," "Biosynthesis of secondary metabolites," "Microbial metabolism in diverse environments," and "Biosynthesis of antibiotics." For the annotation of shared metabolites, we also removed all KEGG compounds that occur in 50 reactions or more ($H_2O$, ATP, ADP, Orthophosphate, Diphosphate, $H^+$, NAD+, NADH, NADPH, NADP+, $CO_2$, $NH_3$, AMP).

**Quantifying relative growth from next generation sequencing (NGS)—**MiSeq data from the sgRNA library and DHFR/TYMS point mutant fitness assays were processed using a custom data analysis script written in Python 2.7. This script counts how often each plasmid 'genotype' *g* and barcode combination were found at each time point. Count data was normalized to the 'wild-type' (WT) control at each time point (to control for mixing variation). In the case of the sgRNA library, the targetless none|none sgRNA pair is used for normalization. In the case of the DHFR/TYMS point mutants, the plasmid encoding wild-type sequences for DHFR and TYMS are used. The relative frequency of each plasmid genotype at time zero is subtracted, to yield the following equation for relative frequency over time:

$$f(t) = \log_{10}\left(\frac{n_g}{n_{WT}}\right)_t - \log_{10}\left(\frac{n_g}{n_{WT}}\right)_{t=0}$$

Where $n_g$ is the number of NGS counts of a given plasmid genotype $g$ at a time $t$. We then conducted a linear fit of the relative frequency versus time to determine a relative growth rate for each plasmid genotype. Refer to Figures S2 and S6 for illustration.

***E. coli* genome assembly—**Genome assembly and mutation prediction was performed using *breseq* (Deatherage and Barrick, 2014; Langmead and Salzberg, 2012). The reference sequence was a modification of the *E. coli* MG1655 complete genome (accession no. NC_000193), edited to include the GFP marker and chloramphenicol resistance cassette in our parent strain.

**Quantification of growth for thymidine dependence and $IC_{50}$ estimation—**Growth was quantified using the positive integral of $OD_{600}$ over time. This measure captures mutational or drug-induced changes in the duration of lag phase as well as perturbations in growth rate (Toprak et al., 2011). For each strain, we identified a start-time ($t_0$) at the end of lag-phase for the fully-rescued 50 μg/mL thy condition. We chose each $t_0$ computationally as the last point before monotonic growth above the limit of detection. The log($OD_{600}$) versus time curves for all conditions are then vertically shifted ('background-subtracted'), such that the function value at this start-time is zero. This curve is then numerically integrated from $t_0$ to $t_0$+15 hours using the trapezoid method.

**$IC_{50}$ estimation—**The trimethoprim resistance of each strain was quantified by its absolute $IC_{50}$, the drug concentration (μg/mL) at which growth is half-maximal. The relationship between growth and trimethoprim inhibition is modeled using the four parameter logistic function:

$$Y = \frac{a - d}{1 + (X / c)^b} + d$$

where $Y$ is growth, $X$ is TMP concentration, $a$ is the asymptote for uninhibited growth, $d$ is the limit for inhibited growth, $c$ provides the concentration midway between $a$ and $d$, and $b$ captures sensitivity (Sebaugh, 2011). Growth versus TMP concentration was fit to the above model using MATLAB. $IC_{50}$ was calculated as the concentration $X^*$ for which growth $Y(X^*) = a/2$.

## DATA AND SOFTWARE AVAILABILITY

Statistical analysis and growth rate calculations were conducted using Python 2.7 and MATLAB. Genome assembly was conducted using *breseq* (Deatherage and Barrick, 2014; Langmead and Salzberg, 2012). The *E. coli* MG1655 with chromosomal *cat/egfp* reference sequence and forward evolution strain genomic sequencing reads have been deposited in the NCBI BioProject database under accession number BioProject: PRJNA378892. LC-MS data was analyzed using the MAVEN software package (Clasquin et al., 2012).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

We thank members of the Reynolds lab for their review of the manuscript, E. Toprak for *E. coli* MG1655 folA single mutants and extensive advice on morbidostat construction and operation, S. Benkovic for the ER2566 *ΔfolA ΔthyA* strain, D. Bikard for the *E. coli* MG1655 strain containing dCas9, T. Kuhlman for molecular biology reagents used in genome editing, S. Thompson for the *folA* supplementation mix recipe, and T. Bergmiller for the GFP and chloramphenicol resistance marker incorporated into our forward evolution parent strain. This research was funded by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through grant GBMF4557 (to K.R.) and by the Green Center for Systems Biology at UT Southwestern Medical Center. A.S. was supported in part by NIH training grant 5T32GM8203-28. I.J. is supported by an ATIP-Avenir grant (Centre National de la Recherche Scientifique).

## REFERENCES

Abdel-Hamid AM, and Cronan JE (2007). Coordinate expression of the acetyl coenzyme A carboxylase genes, accB and accC, is necessary for normal regulation of biotin synthesis in Escherichia coli. J. Bacteriol 189, 369–376. [PubMed: 17056747]

Beck C, Knoop H, and Steuer R (2018). Modules of co-occurrence in the cyanobacterial pan-genome reveal functional associations between groups of ortholog genes. PLoS Genet. 14, e1007239. [PubMed: 29522508]

Bershtein S, Serohijos AW, Bhattacharyya S, Manhart M, Choi JM, Mu W, Zhou J, and Shakhnovich EI (2015). Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. PLoS Genet. 11,e1005612. [PubMed: 26484862]

Beverley SM, Ellenberger TE, and Cordingley JS (1986). Primary structure of the gene encoding the bifunctional dihydrofolate reductase-thymidylate synthase of Leishmania major. Proc. Natl. Acad. Sci. USA 83, 2584–2588. [PubMed: 3458220]

Bhabha G, Ekiert DC, Jennewein M, Zmasek CM, Tuttle LM, Kroon G, Dyson HJ, Godzik A, Wilson IA, and Wright PE (2013). Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat. Struct. Mol. Biol 20, 1243–1249. [PubMed: 24077226]

Bhattacharyya S, Bershtein S, Yan J, Argun T, Gilson AI, Trauger SA, and Shakhnovich EI (2016). Transient protein-protein interactions perturb *E. coli* metabolome and cause gene dosage toxicity. eLife 5, e20309. [PubMed: 27938662]

Bjarnason GA, Jordan RC, Wood PA, Li Q, Lincoln DW, Sothern RB, Hrushesky WJ, and Ben-David Y (2001). Circadian expression of clock genes in human oral mucosa and skin: association with specific cell-cycle phases. Am. J. Pathol 158, 1793–1801. [PubMed: 11337377]

Bolhuis A, Mathers JE, Thomas JD, Barrett CM, and Robinson C (2001). TatB and TatC form a functional and structural unit of the twin-arginine translocase from Escherichia coli. J. Biol. Chem 276, 20213–20219. [PubMed: 11279240]

Bradley MD, Beach MB, de Koning AP, Pratt TS, and Osuna R (2007). Effects of Fis on Escherichia coli gene expression during different growth stages. Microbiology 153, 2922–2940. [PubMed: 17768236]

Chevereau G, Dravecká M, Batur T, Guvenek A, Ayhan DH, Toprak E, and Bollenbach T (2015). Quantifying the Determinants of Evolutionary Dynamics Leading to Drug Resistance. PLoS Biol. 13, e1002299. [PubMed: 26581035]

Clasquin MF, Melamud E, and Rabinowitz JD (2012). LC-MS data processing with MAVEN: a metabolomic analysis and visualization engine. Curr. Protoc. Bioinformatics, Chapter 14, Unit14.11.

Costanzo MS, and Hartl DL (2011). The evolutionary landscape of antifolate resistance in Plasmodium falciparum. J. Genet 90, 187–190. [PubMed: 21869466]

Deatherage DE, and Barrick JE (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol. Biol. 1151, 165–188. [PubMed: 24838886]

Ducker GS, and Rabinowitz JD (2017). One-carbon metabolism in health and disease. Cell Metab. 25, 27–42. [PubMed: 27641100]

Forsburg SL (2001). The art and design of genetic screens: yeast. Nat. Rev. Genet 2, 659–668. [PubMed: 11533715]

Francis K, Stojkovic V, and Kohen A (2013). Preservation of protein dynamics in dihydrofolate reductase evolution. J. Biol. Chem 288,35961–35968. [PubMed: 24158440]

Galperin MY, Makarova KS, Wolf YI, and Koonin EV (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43, D261–D269. [PubMed: 25428365]

Gangjee A, Jain HD, and Kurup S (2007). Recent advances in classical and non-classical antifolates as antitumor and antiopportunistic infection agents: part I. Anticancer. Agents Med. Chem. 7, 524–542. [PubMed: 17896913]

Green JM, and Matthews RG (2007). Folate biosynthesis, reduction, and polyglutamylation and the interconversion offolate derivatives. EcoSal Plus 2.

Halabi N, Rivoire O, Leibler S, and Ranganathan R (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell 138, 774–786. [PubMed: 19703402]

Hawkins JS, Wong S, Peters JM, Almeida R, and Qi LS (2015). Targeted Transcriptional Repression in Bacteria Using CRISPR Interference (CRISPRi). Methods Mol. Biol. 1311, 349–362. [PubMed: 25981485]

Howell EE, Foster PG, and Foster LM (1988). Construction of a dihydrofolate reductase-deficient mutant of Escherichia coli by gene replacement. J. Bacteriol 170, 3040–3045. [PubMed: 2838456]

Huynen M, Snel B, Lathe W 3rd, and Bork P (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10, 1204–1210. [PubMed: 10958638]

Janga SC, Collado-Vides J, and Moreno-Hagelsieb G (2005). Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. Nucleic Acids Res. 33, 2521–2530. [PubMed: 15867197]

Janßen HJ, and Steinbüchel A. (2014). Fatty acid synthesis in Escherichia coli and its applications towards the production of fatty acid based biofuels. Biotechnol. Biofuels 7, 7. [PubMed: 24405789]

Jiang W, Bikard D, Cox D, Zhang F, and Marraffini LA (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol 31, 233–239. [PubMed: 23360965]

Junier I, and Rivoire O (2013). Synteny in bacterial genomes: inference, organization and evolution. arXiv, arXiv:13074291. https://arxiv.org/abs/1307.4291v1.

Junier I, and Rivoire O (2016). Conserved units of co-expression in bacterial genomes: an evolutionary insight into transcriptional regulation. PLoS One 11, e0155740. [PubMed: 27195891]

Kanehisa M, Goto S, Sato Y, Furumichi M, and Tanabe M (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40,D109–D114. [PubMed: 22080510]

Kim J, and Copley SD (2012). Inhibitory cross-talk upon introduction of a new metabolic pathway into an existing metabolic network. Proc. Natl. Acad. Sci. USA 109, E2856–E2864. [PubMed: 22984162]

Kim PJ, and Price ND (2011). Genetic co-occurrence network across sequenced microbes. PLoS Comput. Biol. 7, e1002340. [PubMed: 22219725]

Kim J, Kershner JP, Novikov Y, Shoemaker RK, and Copley SD (2010). Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5′-phosphate synthesis. Mol. Syst. Biol 6, 436. [PubMed: 21119630]

King CH, Shlaes DM, and Dul MJ (1983). Infection caused by thymidine-requiring, trimethoprim-resistant bacteria. J. Clin. Microbiol. 18, 79–83. [PubMed: 6604070]

Kondrashov AS., Sunyaev S, and Kondrashov FA. (2002). Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. USA 99, 14878–14883. [PubMed: 12403824]

Kriegeskorte A, Block D, Drescher M,Windmüller N, Mellmann A, Baum C, Neumann C, Lorè NI, Bragonzi A, Liebau E, et al. (2014). Inactivation of thyA in Staphylococcus aureus attenuates virulence and has a strong impact on metabolism and virulence gene expression. MBio 5, e01447–e14. [PubMed: 25073642]

Kuhlman TE, and Cox EC (2010). Site-specific chromosomal integration of large synthetic constructs. Nucleic Acids Res. 38, e92. [PubMed: 20047970]

Kwon YK, Lu W, Melamud E, Khanam N, Bognar A, and Rabinowitz JD (2008). A domino effect in antifolate drug action in Escherichia coli. Nat. Chem. Biol 4, 602–608. [PubMed: 18724364]

Kwon YK, Higgins MB, and Rabinowitz JD (2010). Antifolate-induced depletion of intracellular glycine and purines inhibits thymineless death in E. coli. ACS Chem. Biol. 5, 787–795. [PubMed: 20553049]

Langmead B, and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. [PubMed: 22388286]

Lazar G, Zhang H, and Goodman HM (1993). The origin of the bifunctional dihydrofolate reductase-thymidylate synthase isogenes of Arabidopsis thaliana. Plant J. 3, 657–668. [PubMed: 8374616]

Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, Russ WP, Benkovic SJ, and Ranganathan R (2008). Surface sites for engineering allosteric control in proteins. Science 322, 438–442. [PubMed: 18927392]

Liu CT, Hanoian P, French JB, Pringle TH, Hammes-Schiffer S, and Benkovic SJ (2013). Functional significance of evolving protein sequence in dihydrofolate reductase from bacteria to humans. Proc. Natl. Acad. Sci. USA 110, 10159–10164. [PubMed: 23733948]

Long CP, Gonzalez JE, Feist AM, Palsson BO, and Antoniewicz MR (2018). Dissecting the genetic and metabolic mechanisms of adaptation to the knockout of a major metabolic enzyme in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 115,222–227. [PubMed: 29255023]

Lu W, Kwon YK, and Rabinowitz JD (2007). Isotope ratio-based profiling of microbial folates. J. Am. Soc. Mass Spectrom. 18, 898–909. [PubMed: 17360194]

Lu W, Clasquin MF, Melamud E, Amador-Noguez D, Caudy AA, and Rabinowitz JD (2010). Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. Anal. Chem 82, 3212–3221. [PubMed: 20349993]

McGuire JJ, and Bertino JR (1981). Enzymatic synthesis and function of folylpolyglutamates. Mol. Cell. Biochem 38, 19–48. [PubMed: 7027025]

Michener JK, Camargo Neves AA, Vuilleumier S, Bringel F, and Marx CJ (2014a). Effective use of a horizontally-transferred pathway for dichloromethane catabolism requires post-transfer refinement. eLife 3, e04279.

Michener JK, Vuilleumier S, Bringel F, and Marx CJ (2014b). Phylogeny poorly predicts the utility of a challenging horizontally transferred gene in Methylobacterium strains. J. Bacteriol 196, 2101–2107. [PubMed: 24682326]

Mitosch K, Rieckh G, and Bollenbach T (2017). Noisy response to antibiotic stress predicts subsequent single-cell survival in an acidic environment. Cell Syst. 4, 393–403.e5. [PubMed: 28342718]

Newman MEJ (2010). Networks: An Introduction (Oxford University Press).

Ogbunugafor CB, Wylie CS, Diakite I, Weinreich DM, and Hartl DL (2016). Adaptive landscape by environment interactions dictate evolutionary dynamics in models of drug resistance. PLoS Comput. Biol. 12, e1004710. [PubMed: 26808374]

Okamura-Ikeda K, Ohmura Y, Fujiwara K, and Motokawa Y (1993). Cloning and nucleotide sequence of the gcv operon encoding the Escherichia coli glycine-cleavage system. Eur. J. Biochem 216, 539–548. [PubMed: 8375392]

Palmer AC, Toprak E, Baym M, Kim S, Veres A, Bershtein S, and Kishony R (2015). Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. Nat. Commun 6, 7385. [PubMed: 26060115]

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA 96, 4285–4288. [PubMed: 10200254]

Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, and Lim WA (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152, 1173–1183. [PubMed: 23452860]

Rengby O, Johansson L, Carlson LA, Serini E, Vlamis-Gardikas A, Kårsnäs P, and Arnér ES (2004). Assessment of production conditions for efficient use of Escherichia coli in high-yield heterologous recombinant selenoprotein synthesis. Appl. Environ. Microbiol 70, 5159–5167. [PubMed: 15345395]

Reynolds KA, McLaughlin RN, and Ranganathan R (2011). Hot spots for allosteric regulation on protein surfaces. Cell 147, 1564–1575. [PubMed: 22196731]

Rivoire O (2013). Elements of coevolution in biological sequences. Phys. Rev. Lett 110, 178102. [PubMed: 23679784]

Rodrigues JV, Bershtein S, Li A, Lozovsky ER, Hartl DL, and Shakhnovich EI (2016). Biophysical principles predict fitness landscapes of drug resistance. Proc. Natl. Acad. Sci. USA 113, E1470–E1478. [PubMed: 26929328]

Sebaugh JL (2011). Guidelines for accurate EC50/IC50 estimation. Pharm. Stat 10, 128–134. [PubMed: 22328315]

Snel B, Bork P, and Huynen MA (2002). The identification of functional modules from the genomic association of genes. Proc. Natl. Acad. Sci. USA 99, 5890–5895. [PubMed: 11983890]

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–D452. [PubMed: 25352553]

Tas H, Nguyen CT, Patel R, Kim NH, and Kuhlman TE (2015). An integrated system for precise genome modification in Escherichia coli. PLoS One 10, e0136963. [PubMed: 26332675]

Toprak E, Veres A, Michel JB, Chait R, Hartl DL, and Kishony R (2011). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat. Genet 44, 101–105. [PubMed: 22179135]

Toprak E, Veres A, Yildiz S, Pedraza JM, Chait R, Paulsson J, and Kishony R (2013). Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. Nat. Protoc 8, 555–567. [PubMed: 23429717]

Zuk O, Hechter E, Sunyaev SR, and Lander ES (2012). The mystery of missing heritability: genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA 109, 1193–1198. [PubMed: 22223662]

## Highlights

- Comparative genomics reveals that the enzymes DHFR and TYMS form an evolutionary module
- Loss of TYMS activity is sufficient to compensate for reductions in DHFR activity
- Coordinated reduction of DHFR and TYMS activity maintains folate metabolite pools
- Genome-wide analyses identify other enzyme pairs as candidate adaptive units

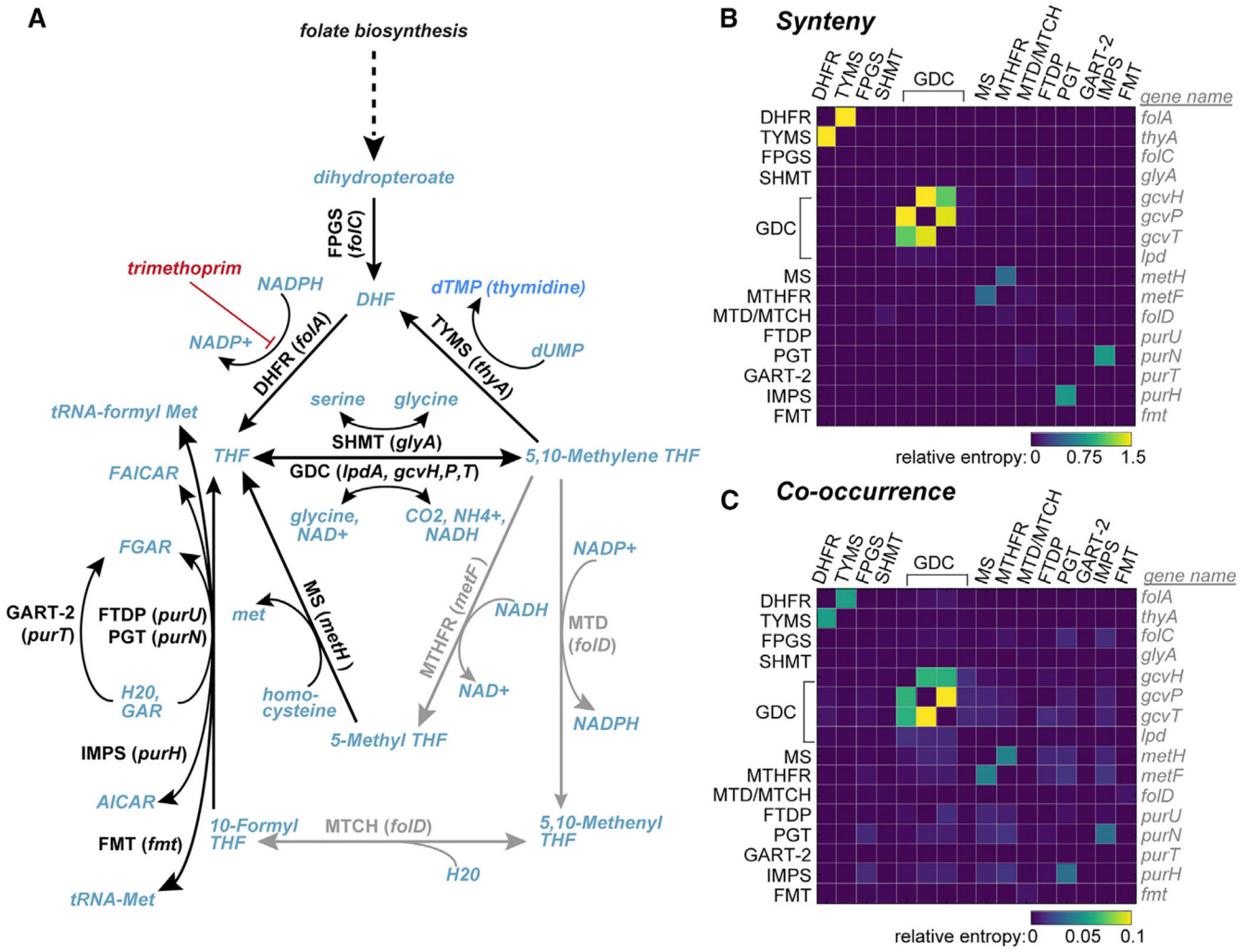



**Figure 1. Biochemical and Statistical Representations of Folate Metabolism**

(A) Biochemical pathway map of folate metabolism. Blue font indicates metabolites; abbreviated enzyme names are in black or gray text. Black text and lines correspond to enzymes annotated as highest-confidence interactions for DHFR (*folA*) in STRINGdb version 10.5. See Table S1 for a more complete description of each enzyme.

(B and C) Heatmaps of evolutionary coupling between gene pairs in folate metabolism as evaluated by gene synteny (B) and co-occurrence (C), and indicated as a relative entropy, $D_{ij}^{intra}$. Enzyme names are indicated on the left and top of the matrix in black text, gene names are given at right in gray italics. In *E. coli*, a single gene (*folD*) encodes a bifunctional enzyme that catalyzes both the methylene tetrahydrofolate dehydrogenase (MTD) and methenyltetrahydrofolate cyclohydrolase (MTCH) reactions.

See also Figure S1.

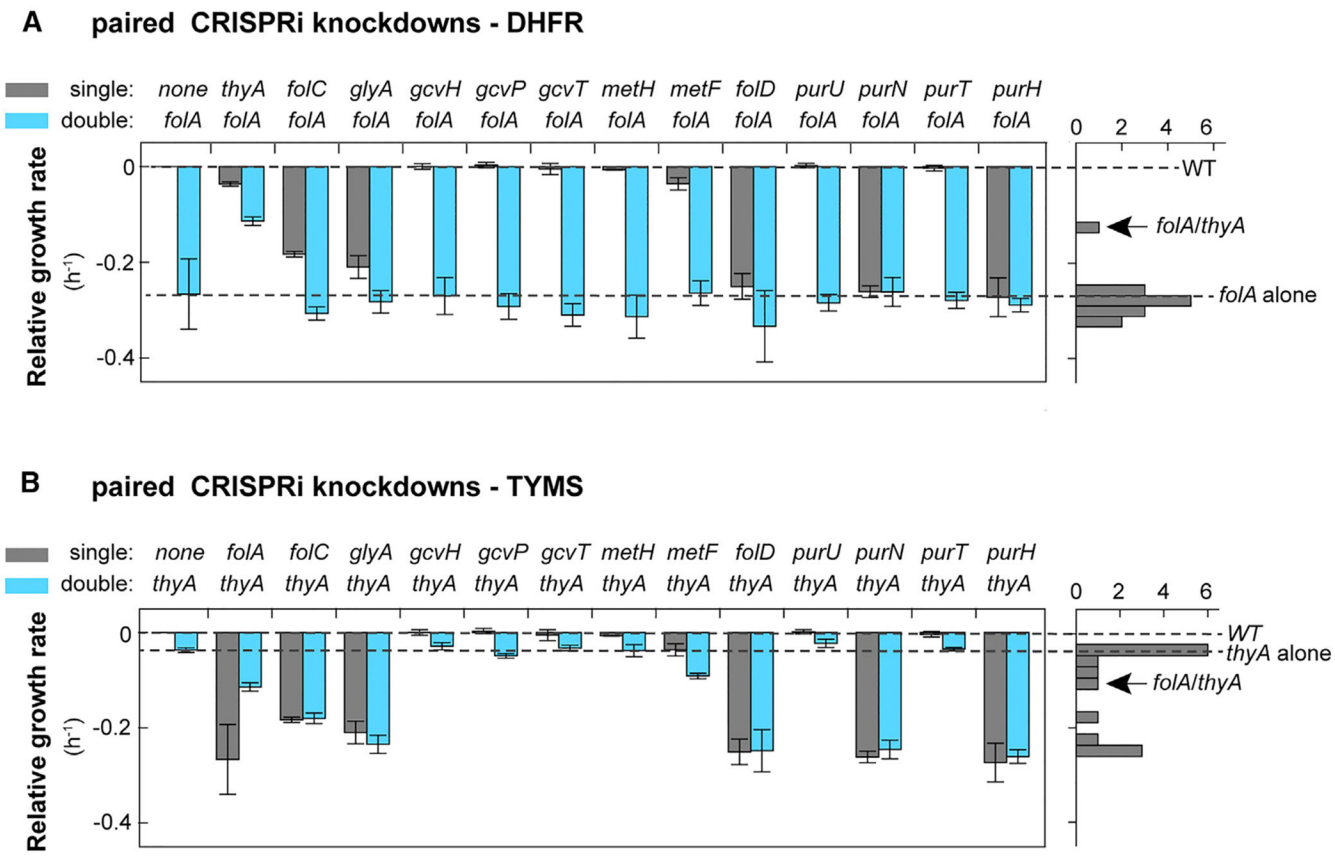



**Figure 2. CRISPRi-Based Measurements of Growth Rate Dependency for DHFR and TYMS**

(A and B) Relative growth rates for CRISPRi knockdowns of folate genes paired with either DHFR (A) or TYMS (B). Growth rates are calculated relative to the fitness of a strain carrying an sgRNA with no target homology region (*none*). Gray bars indicate the growth rate effect of a single mutant; blue bars indicate the effect of the double mutant. Error bars correspond to a 95% confidence interval (CI) across at least three internal technical replicates (see also Method Details). The DHFR knockdown is significantly rescued by TYMS knockdown ($p < 0.005$ by Student's t test), but not by other gene knockdown ($p > 0.20$). As a reference point, the absolute doubling time of the *none* strain in turbidostat monoculture is 0.83 ± 0.09 (95% CI) h. From this, we estimate that a relative growth rate of −0.2 in these mixed population measurements corresponds to a doubling time of approximately 1.8 h, and −0.4 is approximately zero growth rate (dead). To the right of each bar graph, we also plot a histogram of the growth rate effects for all knock downs of DHFR (A) or TYMS (B).

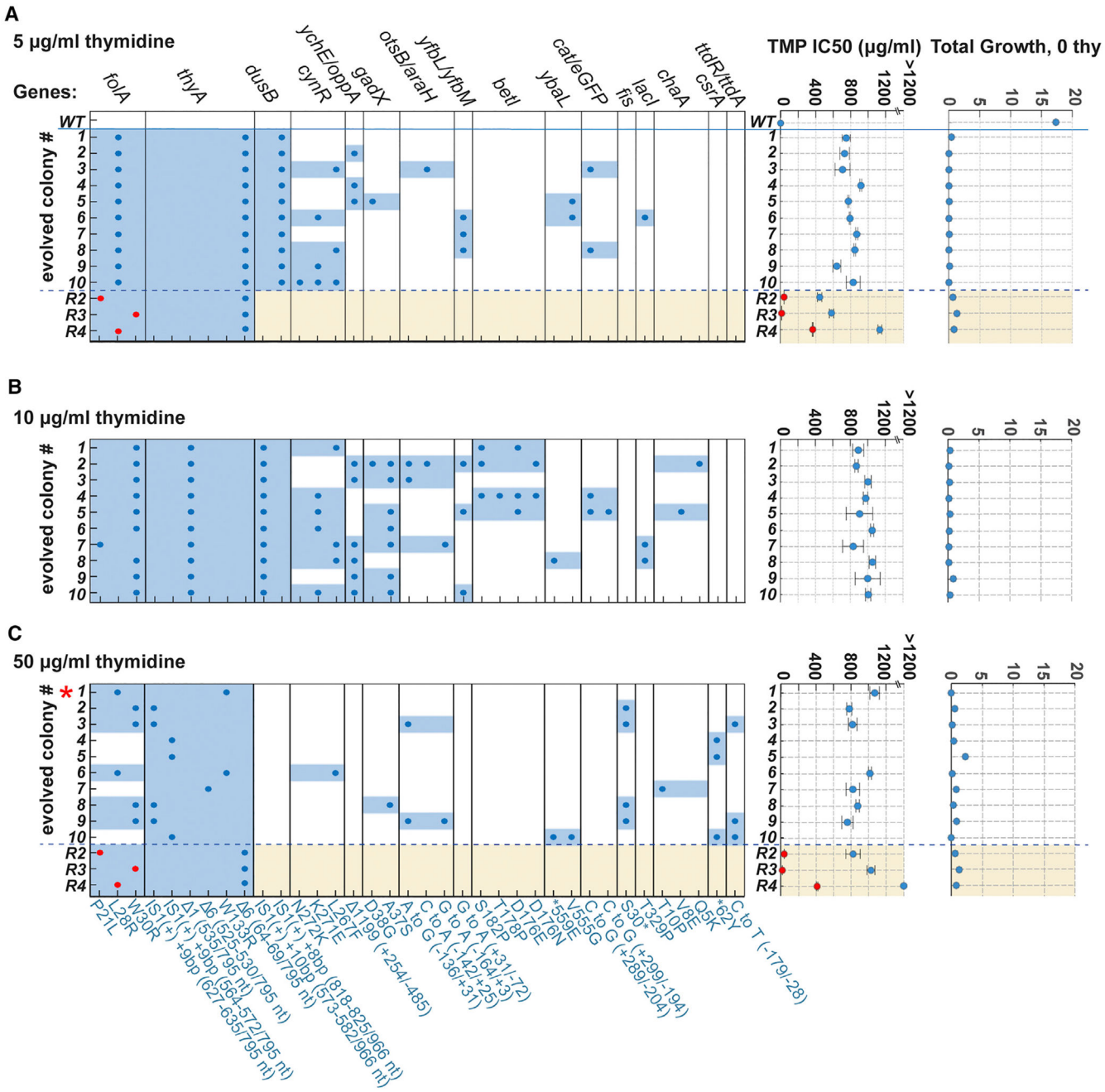



**Figure 3. Genotype and Phenotype of Selected Strains from Three Evolved Populations**

(A–C) Ten single colonies (strains) were selected at the endpoint of each forward evolution condition in 5 μg/mL (A), 10 μg/mL (B), or 50 μg/mL (C) thymidine for genotyping and phenotyping (30 in total). The figure indicates the mutations observed in each strain sampled from that evolution condition. Genes that were mutated in two or fewer strains across all conditions are excluded, as are synonymous mutations (see Table S5 for sequencing statistics and Table S6 for a complete list of non-synonymous mutations). Gene names are labeled along the top edge of the map, with the corresponding residue or nucleotide change(s) denoted along the bottom. If a strain acquires any mutation in a particular gene,

then the column section corresponding to that gene is shaded blue. All but four strains acquired mutations in both *folA* and *thyA*, encoding DHFR and TYMS, with the few exceptions lacking a *folA* mutation. A small red star indicates one strain with mutations in only DHFR and TYMS. To the right of each mutation map are trimethoprim (TMP) $IC_{50}$ and thymidine dependence measurements for each strain. Error bars represent SE over triplicate measurements. Evolved strains 4, 5, and 10 (50 μg/mL thymidine) grew very slowly at all concentrations of trimethoprim measured, and we were unable to determine an $IC_{50}$ by sigmoidal fit. Table S4 contains exact $IC_{50}$ values and errors. Thymidine dependence is represented as area under the $\log(OD_{600})$ curve in 0 μg/mL thymidine over 10 h. Evolved strains are no longer viable in the absence of extracellular thymidine, indicating a loss-of-function mutation in TYMS. Three "reconstitution strains" featuring representative *folA* and *thyA* mutations recombined into a clean genetic background have been included in (A) and (C) for comparison (denoted R2–R4). Red dots in the rows of reconstituted genotypes R2–R4 indicate the $IC_{50}$ of a strain containing only the corresponding *folA* mutation.

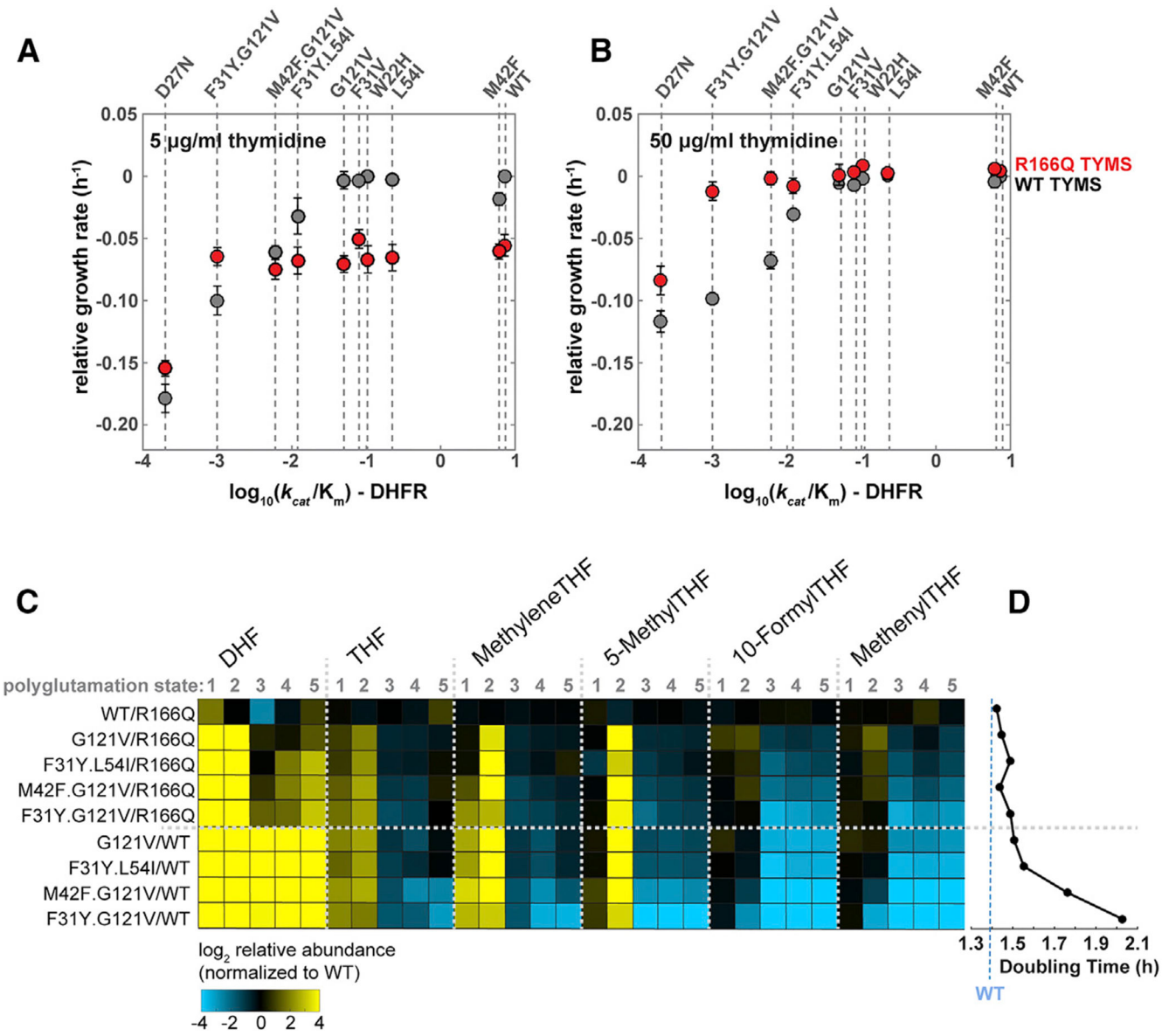



**Figure 4. A Loss-of-Function Mutation in TYS Buffers Metabolic Changes from Decreased DHFR Activity**

(A and B) Scatterplots of relative growth rate for DHFR mutants spanning a range of catalytic specificities ($k_{cat}/K_m$), and either a wild-type (WT, gray points) or catalytically dead (R166Q, red points) TYMS. Measurements were performed in M9 media supplemented with 0.2% amicase and either 5 μg/mL (A) or 50 μg/mL (B) thymidine. Error bars correspond to SE across triplicate measurements.

(C) Liquid chromatography-mass spectrometry profiling of intracellular folate species. Rows reflect mutant DHFR-TYMS combinations, columns correspond to metabolites. Data represent the mean of three replicates; see also Figure S7 for associated errors. Each folate species can be modified by the addition of 1–5 glutamates. The color of each square denotes the $\log_2$ abundance of each species relative to WT.

(D) The corresponding doubling time for each mutant, as measured in batch culture (conditions identical to A).

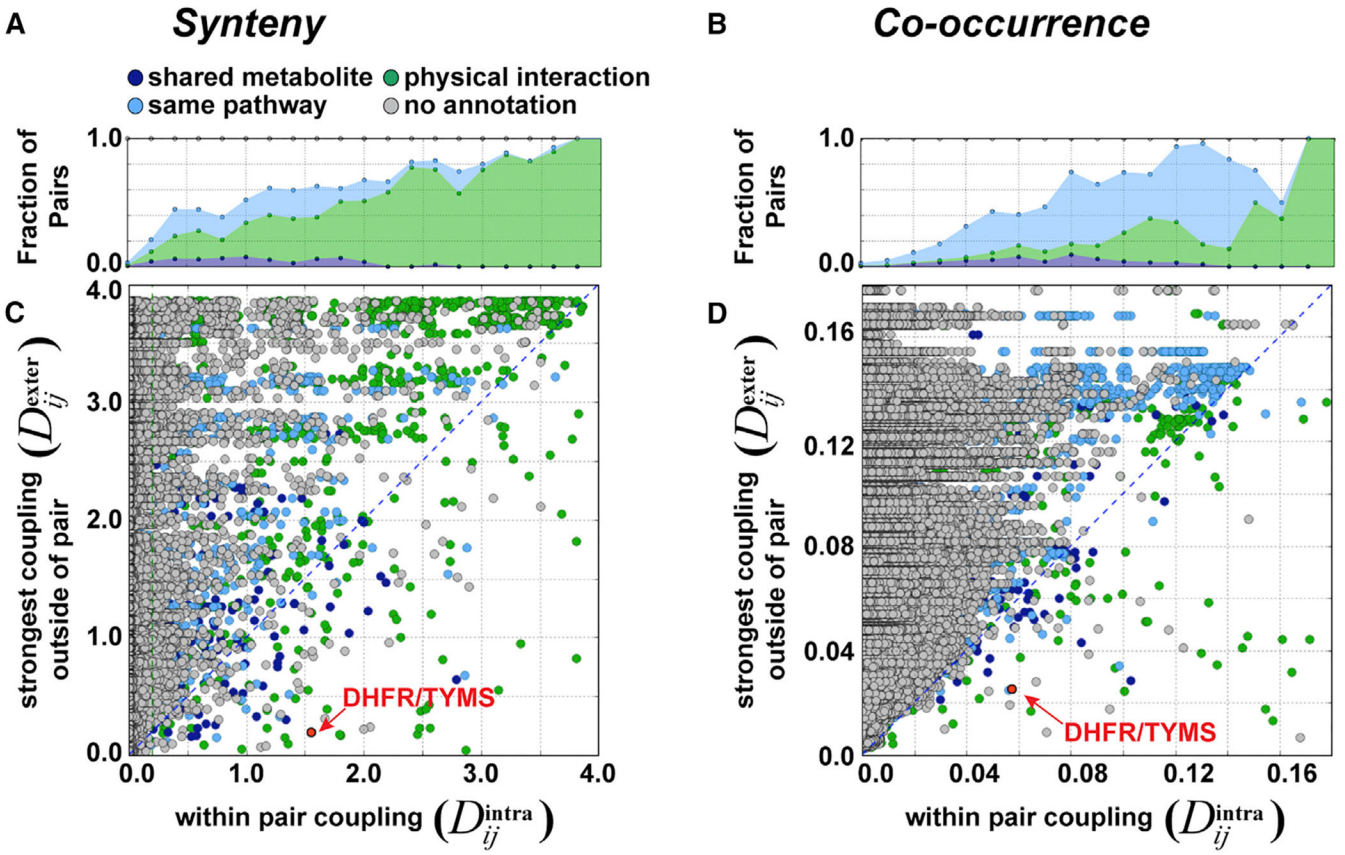



**Figure 5. Genome-wide Analysis of Co-evolution in *E. coli***

(A and B) Enrichment of physical and metabolic interactions as a function of synteny coupling (A) or co-occurrence coupling (B).

(C) A scatterplot of synteny-based coupling for all of the analyzed gene pairs. Each point represents a pair of Clusters of Orthologous Groups (COGs); coupling within the pair is shown on the x axis, and the strongest coupling outside the pair is shown on the y axis. Color-coding reflects annotations from the STRING database (physical interactions) or the KEGG database (metabolic pathways): green indicates binding, while pairs in dark blue or light blue are not annotated as physical interactions but are found in the same metabolic pathway. Dark blue gene pairs share a metabolic intermediate. The DHFR-TYMS pair is highlighted in red. See Table S8 for an annotated list of gene pairs below the diagonal.

(D) Scatterplot of coupling by co-occurrence for all analyzed gene pairs (same format as C).

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Bacterial and Virus Strains | | |
| *E. coli* K12 MG1655 | CGSC | CGSC#:7740 |
| MG1655 with chromosomal dcas9 | Gift from David Bikard | N/A |
| ER2566 *ΔfolA ΔthyA* | Gift from S. Benkovic | N/A |
| Chemicals, Peptides, and Recombinant Proteins | | |
| Purezol | Bio-Rad | Cat#7326890 |
| Amicase | Sigma | Cat#65072-00-6 |
| MOPS EZ Rich Defined Medium | Teknova | Cat#M2105 |
| Critical Commercial Assays | | |
| 300 cycle Illumina MiSeq kit V2 | Illumina | Cat#MS-102-2002 |
| Luna Universal One-Step RT-qPCR kit | NEB | Cat#E3005S |
| QIAamp DNA Minit Kit | QIAGEN | Cat#51304 |
| Nextera XT DNA Library Prep kit | Illumina | Cat#FC-131-1096 |
| 50 cycle Illumina MiSeq kit V2 | Illumina | Cat#MS-102-2001 |
| Deposited Data | | |
| Evolved MG1655 genome sequencing reads | NCBI BioProject | PRJNA378892 |
| Experimental Models: Organisms/Strains | | |
| MG1655 with chromosomal dcas9 and sgRNA library (gAM-350) | This study | N/A |
| Evolved MG1655 strains (5thy condition) | This study | N/A |
| Evolved MG1655 strains (10thy condition) | This study | N/A |
| Evolved MG1655 strains (50thy condition) | This study | N/A |
| MG1655 reconstitution strains | This study | N/A |
| MG1655 *folA* single mutants | Gift from E. Toprak (Palmer et al., 2015) | N/A |
| ER2566 *ΔfolA ΔthyA* pACYC *folA/thyA* library | This study | N/A |
| Oligonucleotides | | |
| Table S2: Primers used for CRISPRi and qPCR | This study | N/A |
| Recombinant DNA | | |
| pAM-111 | This study | N/A |
| pAM-112 | This study | N/A |
| pAM-288/290 | This study | N/A |
| pCRISPR | Jiang et al., 2013 | Addgene plasmid #42875 |
| pgRNA | Qi et al., 2013 | Addgene plasmid #44251 |

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| pTKRED | Kuhlman and Cox. 2010 | Addgene plasmid #41062 |
| Software and Algorithms | | |
| breseq | Deatherage and Barrick. 2014 | http://barricklab.org/twiki/bin/view/Lab/ToolsBacterialGenomeResequencing |
| Bowtie2 | Langmead and Salzberg, 2012 | http://bowtie-bio.sourceforge.net/bowtie2/index.shtml |
| MAVEN | Clasquin et al., 2012 | http://genomics-pubs.princeton.edu/mzroll/index.php |
| Other | | |
| Morbidostat/turbidostat | Toprak et al., 2013 | N/A |