An evolutionary module in central metabolism

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

[1]The Green Center for Systems Biology, and [2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[3]Lewis-Sigler Institute for Integrative Genomics, [4]Department of Chemical and Biological Engineering, and [5]Department of Chemistry, Princeton University, Princeton, NJ 08544, USA

[6]Centre National de la Recherche Scientifique (CNRS), and [7]Université Grenoble Alpes, TIMC-IMAG, F-38000 Grenoble, France

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

*Please address correspondence to:

Kimberly Reynolds
Green Center for Systems Biology
UT Southwestern Medical Center
6001 Forest Park Rd.
Dallas, TX 75390-8597

Ph: 214-645-6157
kimberly.reynolds@utsouthwestern.edu

The ability to predict cell behavior is complicated by an unknown pattern of functional interdependence among genes. Here, we use the conservation of gene proximity across species (synteny) to infer functional couplings between genes. For the folate metabolic pathway, we observe a sparse, modular architecture of interactions, with two small groups of genes coevolving in the midst of others that evolve independently. For one such module –dihydrofolate reductase and thymidylate synthase – we use epistasis measurements and forward evolution to demonstrate both internal functional coupling and independence from the remainder of the genome. Mechanistically, the coupling is driven by a constraint on their relative activities, which must be balanced to prevent accumulation of a metabolic intermediate. The results indicate an organization of cellular systems not apparent from inspection of biochemical pathways or physical complexes, and support the strategy of using evolutionary information to decompose cellular systems into functional units.

Keywords:

co-evolution, statistical genomics, folate metabolism, dihydrofolate reductase (DHFR), epistasis, experimental evolution

Introduction

The activity of one gene is often modified by the activity of other genes in the genome. This functional coupling between genes makes it difficult to predict cellular behavior as a whole from measurements of each gene (or protein) taken independently. As a consequence, our ability to rationally engineer new metabolic systems (Kim and Copley, 2012; Michener et al., 2014a; Michener et al., 2014b), and quantify the relationship between mutations and disease (Kondrashov et al., 2002; Zuk et al., 2012) is limited. Further, this interdependency amongst genes makes it non-trivial to understand how complex cellular systems are possible through an evolutionary process of stepwise variation with selection (Breen et al., 2012; Wagner and Altenberg, 1996; Weinreich et al., 2013). Thus, an ability to globally map functional couplings between genes and subsequently decompose cellular systems into quasi-independent modules - each module consisting of several genes engaged in cooperative function - would help render biological systems tractable and predictable.

However, it remains unclear if such a modular decomposition is possible, and if so, what the general strategy should be for finding it. A fundamental aspect of this problem is to distinguish functional couplings associated with core, conserved processes from those couplings that reflect species and/or environment specific adaptations. In this sense, we seek a general description of genetic interactions that can serve as a basis for guiding targeted experiments and modeling cellular systems. Here, we develop a map of pairwise gene interactions through statistical analysis of co-evolution across thousands of bacterial genomes. The central premise is that functional couplings between proteins drive co-evolution of the associated genes, regardless of details of the interaction mechanism. This co-evolution then leaves a set of detectable statistical signatures in extant genome sequences. Comparative study of natural genetic (co-)

variation across genomes should then reveal fundamental functional interactions important to core cellular processes under evolutionarily relevant conditions, rather than those specific to particular species or environments.

Co-evolution can be manifested in different ways – correlations in amino acid sequence variation, coordinated loss and gain of genes across species, or constraints on relative chromosomal location. In this work, we focus on synteny, the conservation of chromosomal proximity between genes (Overbeek et al., 1999; Tamames, 2001). Synteny is a reliable indicator of functional relationships (Huynen et al., 2000; Janga et al., 2005; Overbeek et al., 1999; Rogozin et al., 2002), and the co-expression of genes (Junier and Rivoire, 2016; Korbel et al., 2004). As in prior work, we thus use synteny to infer functional couplings between genes. In addition, we also use the *absence* of synteny as a measure of independence, with the goal of decomposing cellular systems into groups of genes that co-evolve with each other, but are relatively independent from the rest of the genome.

We begin with a focused study of an experimentally powerful model system: folate metabolism. The folate metabolic pathway involves several interlocking enzymatic loops that catalyze the reactions necessary for synthesis of purine nucleotides, thymidine and a few amino acids. Analysis of gene synteny indicates that this pathway can be decomposed into small modules of one to three genes. Using quantitative measurements of epistasis and forward evolution, we present the first critical tests of these predictions: (1) epistatic coupling within a module and (2) adaptive independence of the module from the remainder of the genome. Motivated by these findings, we carry out a genome wide analysis of pairwise functional couplings between genes (2095 genes, 551,198 gene pairs), which recapitulates and extends the basic findings of our evolutionary analysis for the folate metabolic network. The results indicate

a modular organization of genes into groups that is not obvious given knowledge of the underlying biochemistry or physical complexes. We suggest that such evolutionary modules might represent basic units of function within the cell.

Results

An evolution-based map of functional coupling in folate metabolism

The core one-carbon folate metabolic pathway consists of thirteen enzymes that interconvert various folate species and produce methionine, serine, thymidine and purine nucleotides in the process (Fig. 1A and Table S1) (Green and Matthews, 2013). 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 dihydrofolate reductase (DHFR) using 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 serine and purine nucleotides (Fig. 1A, bottom square portion of pathway). The only step that oxidizes THF back to DHF is catalyzed by the enzyme thymidylate synthase (TYMS), which modifies uridine monophosphate (dUMP) to thymidine monophosphate (dTMP) in the process. This pathway is well conserved across organisms, ensuring good statistics for our analysis. Further, the function of folate metabolism can be readily assessed through quantitative growth rate measurements (Reynolds et al., 2011), and due to the central role of these metabolites in cell growth and division, folate metabolism is a target of several well-known antibiotics (trimethoprim and sulfamethoxazole), and chemotherapeutics (methotrexate and 5-fluorouracil) (Ducker and Rabinowitz, 2016; Gangjee et al., 2007). These factors enable experimental

strategies to measure gene function and epistatic coupling *in vivo*. Thus, folate metabolism provides a good model system to examine the use of synteny in identifying functional modules.

We studied the pattern of couplings between genes in the folate pathway through a quantitative analysis of synteny over 1445 bacterial genomes. The basic operation is to compute the frequency at which a particular pair of orthologous genes occur within a given distance along the chromosome across all genomes, and then calculate the significance of this observation (as a p-value) given a null model in which genes are randomly and uniformly distributed along the chromosome (Junier and Rivoire, 2016). Previous work has shown that this analysis identifies stretches of genes larger than single operons that tend to be co-expressed (Junier and Rivoire, 2016). Here, we convert the synteny p-value between any two genes i and j into a relative entropy D_{ij}. This provides a measure of synteny that is independent of the number of genomes analyzed (see Supplemental Experimental Procedures for details).

Examining synteny for genes comprising the folate pathway reveals a sparse pattern of evolutionary coupling in which most genes are relatively independent from each other (Fig. 1B). Consistent with intuition and expectations from prior work, we observe coupling between physically interacting genes: the glycine cleavage system proteins H, P and T (*gcvH*, *gcvP*, and *gcvT* in *E. coli*). Together with lipoamide dehydrogenase (*lpdA*), these enzymes form the glycine decarboxylase complex (GDC), a macromolecular complex that reversibly catalyzes either the degradation or biosynthesis of glycine (Okamura-Ikeda et al., 1993). Notably, lipoamide dehydrogenase also functions as part of both the 2-oxoglutarate dehydrogenase and pyruvate dehydrogenase multienzyme complexes (Carothers et al., 1989); this generality of function may underlie its evolutionary independence from *gcvH*, *gcvP,* and *gcvT*.

Interestingly, we also see evolutionary coupling of enzyme pairs with no evidence for physical interaction: 1) DHFR/TYMS and 2) methionine synthase (MS) and methionine tetrahydrofolate reductase (MTHFR). Indeed, DHFR and TYMS comprise the most strongly coupled gene pair in the folate cycle. Both pairs of enzymes catalyze consecutive reactions in folate metabolism, suggesting a possible mechanistic basis for functional coupling. However, we note that biochemical proximity of reactions is not a sufficient criterion for evolutionary coupling; many gene pairs that are locally linked in the biochemical network do not show statistical correlation (Fig. 1). Thus, our synteny analysis does not simply recapitulate the connections in a standard biochemical network map. Instead, it provides a different representation in which many genes are near independent and a few interact to form modular units. These interacting genes behave as *evolutionary modules* – they coevolve with one another, but are relatively independent from the rest of the metabolic pathway.

The DHFR/TYMS enzyme pair provides a good test case for the hypothesis that evolutionary modules represent near-independent functional units. The genes are highly coupled by co-evolution, but this coupling is not explained by the formation of a physical complex. Further, though the genes encoding DHFR and TYMS are proximal along the chromosome of many bacterial species, they are approximately 2.9 megabases apart in *E. coli* (~4.6 Mbp total genome size). So, experiments in this model system provide an opportunity to test if statistical modularity over an ensemble of genomes corresponds to functional modularity even in the absence of chromosomal proximity in the selected instance.

Coupling between DHFR and TYMS depends on enzyme activity

Does the coevolution of DHFR and TYMS correspond to functional coupling in the folate metabolic network? To address this, we conducted quantitative measurements of genetic

epistasis for a library of ten well-characterized DHFR mutants in the background of either WT TYMS or TYMS R166Q, a catalytically inactive variant (twenty constructs in total). We used a previously validated next-generation sequencing based assay to measure the relative fitness of all possible mutant combinations (20 total) in a single internally controlled experiment (Reynolds et al., 2011). In this system, DHFR and TYMS are 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 (Fig. 2A). The full library of mutants is transformed into the auxotroph strain *E. coli* ER2566 *ΔfolA ΔthyA*, and grown as a mixed population in a turbidostat. The turbidostat allows us to maintain the cell population at a fixed density in exponential phase for the duration of the experiment, with excellent control over media conditions. We then sampled time points over a twelve-hour period, and used next-generation sequencing to compute allele frequencies at each time point (Fig. 2B). By fitting a slope to the plot of allele frequencies versus time, we obtain a relative growth rate for each mutant in the population (Fig. 2B and Fig. S1). The advantage of this approach is that we obtain a quantitative measure of growth rate variation for many mutations in parallel, and thus establish a more complete picture of how epistasis varies with the magnitude of the perturbation.

Analysis of the DHFR mutants in the background of WT TYMS (grey points, Fig. 2C-D) shows that growth rate depends monotonically on DHFR catalytic activity: decreasing DHFR activity corresponds to slower growth. The TYMS R166Q mutant is non-viable, unless the media is supplemented with thymidine (the product of TYMS). In the context of WT DHFR, TYMS R166Q results in a growth rate defect in media supplemented with low amounts of thymidine (5 μg/ml), and no growth rate defect in the presence of 50 μg/ml thymidine. However, in the context of the low activity DHFR mutants, TYMS R166Q has the counter-intuitive

consequence of partly (Fig. 2C) or even fully restoring growth rate (Fig. 2D). That is, the TYMS R166Q mutation decreases fitness in the background of a high-activity DHFR, but increases fitness when paired with a low-activity DHFR. This epistasis – in which loss of function in TYMS buffers decreases in the catalytic activity of DHFR – is consistent with the evolutionary coupling of this enzyme pair.

Mechanism of DHFR/TYMS coupling

Given no evidence for the physical association of DHFR and TYMS in bacteria, what is the mechanism underlying their coupling? The finding that epistasis between the two genes depends on enzyme activity suggests a simple hypothesis: the coupling arises from the need to balance the concentration of key metabolites in the folate metabolic pathway. Support for this idea comes from prior work showing that treatment of *E. coli* with the DHFR inhibitor trimethoprim results in intracellular accumulation of DHF, which inhibits 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). Thus, DHF accumulation results in off-target enzyme inhibition and cellular toxicity, an explanation for the growth rate defect observed in hypomorphic DHFR alleles (Fig. 4C-D). Because DHF is a product of TYMS, it is logical that loss-of-function mutations in TYMS might rescue growth in DHFR hypomorphs by preventing the accumulation of DHF.

To test this hypothesis, we carried out liquid chromatography-mass spectrometry (LC-MS) profiling of folate pathway metabolites in DHFR/TYMS mutant combinations. Specifically, we selected five DHFR variants that span a range of catalytic activities (WT,

G121V, F31Y.L54I, M42F.G121V, and F31Y.G121V), and measured the relative abundance of intracellular folates in the background of either wild-type or R166Q TYMS. 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 display significant growth defects individually, but in which the corresponding DHFR/TYMS double mutants are restored to near wild-type 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 (Fig. 3A, bottom four rows). In addition, we find evidence for a depletion of reduced polyglutamated folates (Glu >= 3), while several mono- and di-glutamated THF species accumulate (particularly for THF, Methylene THF and 5-Methyl THF). This pattern of changes in the reduced folate pool is consistent with inhibition of FP-γ-GS by DHF accumulation (Fig. 3A, Fig. S2). It is also consistent with the observed growth rate defects in the DHFR loss-of-function mutants (Fig. 3B). How does the metabolite profile look in the background of the corresponding TYMS loss-of-function mutant? As predicted, we find clear evidence that the metabolite profile is corrected in the background of TYMS R166Q. Indeed, the concentrations of the reduced polyglutamated folates are restored to near-wild-type levels for most of the DHFR alleles (Fig 3A, top four rows). These data show that coordinated decreases in the activity of DHFR and TYMS maintain balance in key intracellular metabolites, a condition associated with optimal growth. Thus the coupling of DHFR and TYMS can be explained by a

joint constraint on their catalytic activities – a biochemical mechanism for the coevolution of the DHFR/TYMS gene pair.

Forward evolution reveals independence of DHFR and TYMS from the rest of the genome

The analysis of coevolution presented in Fig. 1 goes beyond just the prediction of epistatic coupling between DHFR and TYMS. The lack of coupling to other folate metabolic genes suggests that they might act as a near-independent evolutionary module within folate metabolism. To test this, we carried out a genome-wide suppressor screen in which we make perturbations to one component of the two-gene unit and examine the pattern of compensatory mutations. If DHFR and TYMS act as a quasi-independent unit, then suppressor mutations should be found within the genetic loci encoding this pair of enzymes with minimal contributions from other sites. Practically, this experiment entails making a perturbation within the DHFR/TYMS module that reduces organismal growth rate, conducting forward evolution to generate an adaptive response, and performing whole genome sequencing of the output.

As a perturbation, we grew wildtype *E. coli* cells (strain MG1655) in the presence of trimethoprim, a common antibiotic and inhibitor of many prokaryotic DHFRs. To facilitate the evolution of resistance to trimethoprim, we used a morbidostat, a specialized device for continuous culture (Toprak et al., 2012; Toprak et al., 2013) (Fig. 4A-C). The morbidostat dynamically adjusts the trimethoprim concentration in response to bacterial growth rate and total optical density, thereby providing steady selective pressure as resistance levels increase (see Fig. 4 legend for details). 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 growth rate is decreased. This approach makes it possible to obtain long trajectories of adaptive mutations in the genome with good statistics and

sustained phenotypic adaptation (Toprak et al., 2012). 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. We carried out evolutionary trajectories in four different media conditions, in which the concentration of exogenous thymidine was varied from none to an amount sufficient to rescue the knockout of TYMS (0, 5, 10, and 50 μg thymidine). All conditions were also supplemented with amicase, a source of free amino acids. As shown in Fig. 5, these different environments can buffer genetic variation in the folate metabolic pathway to different extents. This offers a means to expose a larger range of adaptive mutations than one would observe under a single environment; in this context an absence of mutations outside of the two-gene module becomes more significant.

Over the 13 days of evolution, we estimate the trimethoprim resistance of each evolving population by computing the median drug concentration in the culture vial from the first dilution with drug to the end of the day. Following the median drug concentration, we see that populations supplemented with thymidine evolve trimethoprim resistance more rapidly (Fig. 4D and Fig. S3), suggesting that addition of thymidine to the media accelerates the acquisition of resistance, possibly by opening up new evolutionary paths. To identify the mutants causally related to trimethoprim resistance, we selected 10 single colonies from the endpoint of each of the four experimental conditions for phenotypic and genotypic characterization (40 strains in total, Fig. 5). For each strain, we measured the trimethoprim IC50, growth rate dependence on thymidine, and conducted whole genome sequencing. Consistent with the dynamic estimates of trimethoprim resistance, strains isolated from thymidine-supplemented conditions attained trimethoprim IC50s two orders of magnitude higher than their un-supplemented counterparts (Fig. 5A and Table S2). We were unable to measure IC50 values for strains 4, 5 and 10 from the

50 μg/ml thymidine condition: these three strains grew very slowly but were completely insensitive to trimethoprim. Further, strains from all three thymidine supplemented conditions now depend on exogenous thymidine for growth, indicating a loss of function in the *thyA* gene that encodes TYMS (Fig. 5B and Fig. S4). 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 in the absence of trimethoprim retain TYMS function over similar time scales (Fig. S5).

Whole genome sequencing for all 40 strains reveals a striking pattern of mutation (Fig. 5C and Tables S3-S4). Consistent with a previous morbidostat-based study of trimethoprim resistance, under conditions of no thymidine we observe a mutation in the promoter of the *folA* gene that encodes DHFR, but no mutations in TYMS (Toprak et al., 2012). This mutation was previously shown to enhance trimethoprim resistance by increasing DHFR expression (Flensburg and Skold, 1987). In comparison, isolates from all three thymidine-supplemented conditions acquire coding-region mutations in both DHFR and TYMS, or even just in TYMS. For example, strains 4, 5, 7, and 10 in the 50 μg/ml thymidine condition contain mutations in TYMS but not DHFR – showing that one route to resistance is the acquisition of mutations in a gene not directly targeted by antibiotic. All mutations isolated in DHFR reproduce those observed in the earlier morbidostat study of trimethoprim resistance (Toprak et al., 2012). The mutations in TYMS – two insertion sequence elements, a frame shift mutation, loss of two codons, and a non-synonymous active site mutation - are consistent with loss of function. Thus, the mutations in DHFR and TYMS are consistent with the proposed mechanism of coupling: reduced TYMS activity can buffer inhibition of DHFR.

Consistent with the evolutionary independence of the DHFR/TYMS pair, we observe no other mutations in folate metabolism genes (Fig. 5C and Table S4). More generally, few other

mutations occur elsewhere in the genome, and the majority of these are not systematically observed across clones. This result implies that they may be spurious variations not associated with the adaptive phenotype. One of the evolved strains contains only mutations in DHFR and TYMS (strain 1 in 50 μg/ml thymidine, Fig. 5C), indicating that variation in the DHFR/TYMS genes is sufficient to produce resistance. To establish this, we introduced several of the observed DHFR and TYMS mutations into a clean wild-type *E. coli* MG1655 background and measured the IC50. These data show that the DHFR/TYMS mutations are sufficient to reproduce the resistance phenotype measured for the evolved strains (Fig. S5). Thus, DHFR and TYMS show a capacity for adaptation through compensatory mutation that is contained within the two-gene unit. Consistent with the laboratory findings reported here, loss-of-function mutations in TYMS have been observed in a subset of trimethoprim-resistant gram-negative clinical isolates (including *E. coli*), indicating that resistance from modulation of the DHFR/TYMS gene pair is also relevant in a natural environment (King et al., 1983). From this, we conclude that DHFR and TYMS act as a quasi-independent adaptive module.

A global statistical analysis of modular synteny pairs in bacteria

Our focused study of the folate metabolic pathway shows that gene synteny can reveal functionally meaningful evolutionary modules within a cellular system. To examine the modular structure of the entire genome, we conducted a global analysis of pairwise synteny relationships amongst genes represented in *E. coli*. Following from previous work, we use clusters of orthologous groups of proteins (COGs) to define orthologs across species (Galperin et al., 2015). To ensure good statistics, we limit the COGs analyzed to those that co-occur in at least 100 effective genomes (2095 COGs, ~500,000 pairs in total) (see also Supplemental Experimental Procedures). In Figure 6A, we show a scatterplot of gene pairs, indicating the strength of

coupling within each pair (as a relative entropy, along the x-axis) versus the strongest coupling outside of the pair (along the y-axis). In this plot, points fall below the diagonal if the genes in the pair are more tightly coupled to each other than any other gene in the dataset (see Table S5 for a list of pairs). One of these points (in red) corresponds to the DHFR/TYMS pair. Thus, these two enzymes are not only decoupled from folate metabolism, but from all other genes in the genome-wide analysis.

These data reinforce observations made at the single pathway scale. Just like for the folate pathway (Fig. 1B), the pattern of coupling between genes at the genome scale is sparse, as demonstrated by the high density of points with weak coupling (on the left of the graph, along the y-axis). Analysis of the maximum coupling for each gene shows that 906 genes (43%) do not have significant coupling to any other gene in the genome ($\max(D_{ij}) < 0.025$), suggesting that many genes might behave as single gene modules (Fig. S6A). To understand the relationship between gene pairs coupled by synteny and functional or physical interaction of the associated gene products, we compared our analysis to metabolic annotations from KEGG (Kanehisa et al., 2012) and the set of high-confidence binding interactions in *E. coli* reported by the STRING database (Szklarczyk et al., 2015). As expected, coupled gene pairs show enrichment for physical complexes, enzymes in the same metabolic pathway, and more specifically, enzymes with a shared metabolite (Fig. 6A,C). But, like for the folate pathway, the vast majority of sequential reactions are not coupled. In general, the statistical analysis does not simply recover the local biochemical relationships in the metabolic pathway diagram. Instead, it identifies couplings between a subset of enzyme pairs.

A general definition for evolutionary modules depends on both strong internal coupling within a module (D_{ij}^{intra}) and weak external coupling (D_{ij}^{exter}) to other genes in the genome.

Though it remains a matter for future work to experimentally test the relationship between both of these values and functional modularity, it is instructive to examine other gene pairs with patterns of evolutionary coupling similar to DHFR and TYMS. For illustrative purposes, we consider a simple definition of modular pairs based on empirical cutoffs for internal and external coupling (D_{ij}^{intra}> 1.0 and D_{ij}^{exter}< 0.5) (dashed orange box in Fig. 6B). In this set, we observe enrichment for known functional and physical interactions beyond that for coupled gene pairs (Fig. 6C). Table S6 shows that this enrichment does not depend strongly on the choice of cutoff.

The connection between synteny and co-expression (Junier and Rivoire, 2016; Korbel et al., 2004) leads to a natural interpretation of these evolutionary modules as groups of genes whose activity or expression is constrained relative to each other, but that are more independent from the rest of the pathway or system. The DHFR and TYMS pair is consistent with this interpretation – the cell can tolerate reductions in DHFR activity if they are accompanied by loss of function in TYMS. Study of other evolutionary modules from our analysis provides further support for this idea. For example, the gene pair *accB*/*accC* encodes two of the four subunits of acetyl-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 Steinbuchel, 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).

Though the analysis presented here focuses on pairs, the concept of evolutionary modules extends to larger groups of genes. In this regard, we expect that some of the genes near the diagonal are in fact part of larger gene modules (e.g. the highly coupled ribosomal gene pair *rpsC* and *rpmC,* see also Table S5). Beyond a mere partition into independent modules, the

evolutionary analysis in fact leads to a richer representation: a weighted network of synteny relationships. This network awaits further computational analysis and comprehensive testing, following from the approaches developed in this work.

Discussion and Conclusions

Metabolic constraints as an origin for co-evolution and modularity

Much prior work has demonstrated that physical protein interactions can drive co-evolution, particularly via the acquisition of complementary interface mutations (Aakre et al., 2015; Hopf et al., 2014; Ovchinnikov et al., 2014; Podgornaia and Laub, 2015). Our analysis of the DHFR/TYMS pair demonstrates a different mechanism for coevolution: constraints on metabolite concentration can drive coordinated changes in enzyme activity. For the DHFR/TYMS pair, coupling appears to be driven by the need to constrain intracellular levels of the intermediate DHF. As a consequence, we see that treatment with trimethoprim experimentally can result in coordinated evolution of both genes. Additionally, recent work has shown that growth rate defects due to overexpression of *E. coli* DHFR can be partly rescued by increasing TYMS expression, consistent with a general constraint on the relative activities of these two genes (Bhattacharyya et al., 2016). Thus, coevolution is not limited to physical complexes, but more generally reflects the coupling of gene activities regardless of mechanism (Huynen et al., 2000; Snel et al., 2002).

While the mechanism of DHFR/TYMS coupling seems reasonably clear, how and why this pair is decoupled from the rest of metabolism is less obvious. Mathematical models of the folate cycle based on standard biochemical kinetics provide several useful insights (Leduc et al., 2007; Nijhout et al., 2004). First, in eukaryotic cells, thymidine synthesis is the rate-limiting step for DNA synthesis, 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). Computationally increasing the activities of DHFR and TYMS 100-fold results in increased thymidine synthesis but only modestly changes the concentration of folate pools. Secondly, the bacterium *R. capsulatus* lacks both *thyA* (TYMS) or *folA* (DHFR) homologs, and instead produces thymidine via *thyX*, a thymidylate synthase that generates THF (rather than DHF) in the process of thymidine production. When *thyX* is deleted from *R. capsulatus*, growth can only be complemented by the addition of both *thyA* and *folA* from *R. sphaeroides*; the *thyA* gene alone is insufficient. Computational simulation shows that in the absence of a high-activity DHFR (*folA*), *thyA* rapidly depletes reduced folate pools by converting them to DHF. The results of these two computational studies are consistent with the idea that relative activities of DHFR and TYMS should be matched. Further, the results suggest that decoupling DHFR and TYMS from the remainder of folate metabolism provides a general strategy to maintain homeostasis independent of physiological or evolutionary variation in these two genes. That is, modularity might allow for adaptive variation in DHFR and TYMS activity while enabling robustness in the remainder of the pathway.

Using evolutionary statistics to decompose cellular systems

The central premise of this work is to use evolutionary statistics to infer couplings between genes, and identify near-independent adaptive modules. Prior work has largely focused on mapping functional couplings between genes in metabolic systems either computationally via flux balance analysis (Deutscher et al., 2006; He et al., 2010; Segre et al., 2005) or experimentally through high-throughput, quantitative assays of cell growth and epistasis (Babu et al., 2011; Collins et al., 2010; Costanzo et al., 2016; Typas et al., 2008). Though important, such studies cannot generally separate the species- or experiment-specific constraints between genes

from the conserved constraints that represent the fundamental aspects of genome function. We propose that quantitative analysis of statistical relationships over an ensemble of diverse genomes can provide general models that serve to focus experimental study on the core processes of cellular systems.

Comparison of the genome-scale synteny analysis to existing large datasets (KEGG and STRING) provides encouraging validation of this approach – many of the coupled gene pairs identified by our analysis are consistent with known interactions, including physical complexes and consecutive reactions in metabolism. However, the data reported here shows that existing databases of metabolic structure, physical interactions or gene expression should not be seen as "gold standards" for validating and interpreting co-evolutionary data. Indeed, since coevolution can be driven by different mechanisms, the patterns of epistasis we deduce could extend beyond known physical or metabolic interactions to yield new principles of genome organization and function. Thus, a meaningful test of evolutionary statistical analyses requires new types of experiments that can test both the functional coupling of genes, and the independence of proposed multi-gene modules. Large-scale measurements of gene epistasis begin to address this, but in many cases, are limited to the extreme case of total gene knockout. The epistasis measurements for DHFR and TYMS illustrate how mutations across a range of perturbations to catalytic activity (and growth rate phenotype) can provide additional insight into the nature of gene interaction. The experimental methods developed here provide a clear technical framework for testing and guiding development of co-evolution based approaches.

The experimental data for DHFR and TYMS establish that statistical analysis of synteny across genomes has the capacity to identify functional modules in metabolism. However, other signals of co-evolution exist and should be considered. For example, correlations in gene

presence (or absence) across bacterial species have been used to predict functional interactions (Pellegrini et al., 1999), and to identify modules of evolutionarily coupled genes (Kim and Price, 2011). In the case of the folate metabolic pathway, the pattern of coupling obtained by gene presence/absence echoes the modular decomposition observed by synteny (Fig. S6B,C). Again, we observe an overall sparse pattern of coupling, and the DHFR/TYMS gene pair forms an isolated evolutionary module. So, in this instance, the modularity of the DHFR/TYMS pair is identifiable by two distinct measures. More generally, further study is required to more carefully understand the relationship between different co-evolutionary signals, but it is possible that different measures may inform us about distinct aspects of the underlying biology. In summary, our results suggest the existence of a rich intermediate organizational layer between individual genes and complete pathways, consisting of multi-gene modules. This work establishes a viable path to decompose the genome into such functionally and evolutionarily meaningful gene groups using evolutionary information.

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

Acknowledgements We thank members of the Reynolds lab for review of the manuscript, E. Toprak for extensive advice on morbidostat construction and operation, S. Benkovic for the ER2566 *ΔfolA ΔthyA* strain, T. Kuhlman for molecular biology reagents used in genome editing, T. Bergmiller for the GFP/Chloramphenicol resistance marker incorporated into our founder

strain, and R. Ranganathan for discussions. This research was funded in part 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

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell *163*, 594-606.

Abdel-Hamid, A.M., and Cronan, J.E. (2007). Coordinate expression of the acetyl coenzyme A carboxylase genes, accB and accC, is necessary for normal regulation of biotin synthesis in Escherichia coli. Journal of bacteriology *189*, 369-376.

Babu, M., Gagarinova, A., and Emili, A. (2011). Array-based synthetic genetic screens to map bacterial pathways and functional networks in Escherichia coli. Methods in molecular biology *781*, 99-126.

Bhattacharyya, S., Bershtein, S., Yan, J., Argun, T., Gilson, A.I., Trauger, S.A., and Shakhnovich, E.I. (2016). Transient protein-protein interactions perturb E. coli metabolome and cause gene dosage toxicity. eLife *5*.

Bjarnason, G.A., Jordan, R.C., Wood, P.A., Li, Q., Lincoln, D.W., Sothern, R.B., Hrushesky, W.J., and Ben-David, Y. (2001). Circadian expression of clock genes in human oral mucosa and skin: association with specific cell-cycle phases. The American journal of pathology *158*, 1793-1801.

Bolhuis, A., Mathers, J.E., Thomas, J.D., Barrett, C.M., and Robinson, C. (2001). TatB and TatC form a functional and structural unit of the twin-arginine translocase from Escherichia coli. The Journal of biological chemistry *276*, 20213-20219.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature *490*, 535-538.

Carothers, D.J., Pons, G., and Patel, M.S. (1989). Dihydrolipoamide dehydrogenase: functional similarities and divergent evolution of the pyridine nucleotide-disulfide oxidoreductases. Archives of biochemistry and biophysics *268*, 409-425.

Clasquin, M.F., Melamud, E., and Rabinowitz, J.D. (2012). LC-MS data processing with MAVEN: a metabolomic analysis and visualization engine. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *Chapter 14*, Unit14 11.

Collins, S.R., Roguev, A., and Krogan, N.J. (2010). Quantitative genetic interaction mapping using the E-MAP approach. Methods in enzymology *470*, 205-231.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D.*, et al.* (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science (New York, NY) *353*.

Deatherage, D.E., and Barrick, J.E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods in molecular biology *1151*, 165-188.

Deutscher, D., Meilijson, I., Kupiec, M., and Ruppin, E. (2006). Multiple knockout analysis of genetic robustness in the yeast metabolic network. Nature genetics *38*, 993-998.

Ducker, G.S., and Rabinowitz, J.D. (2016). One-Carbon Metabolism in Health and Disease. Cell metabolism.

Flensburg, J., and Skold, O. (1987). Massive overproduction of dihydrofolate reductase in bacteria as a response to the use of trimethoprim. European journal of biochemistry / FEBS *162*, 473-476.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic acids research *43*, D261-269.

Gangjee, A., Jain, H.D., and Kurup, S. (2007). Recent advances in classical and non-classical antifolates as antitumor and antiopportunistic infection agents: part I. Anti-cancer agents in medicinal chemistry *7*, 524-542.

Green, J.M., and Matthews, R.G. (2013). Folate Biosynthesis, Reduction, and Polyglutamylation and the Interconversion of Folate Derivatives. EcoSal Plus.

He, X., Qian, W., Wang, Z., Li, Y., and Zhang, J. (2010). Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nature genetics *42*, 272-276.

Hopf, T.A., Scharfe, C.P., Rodrigues, J.P., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife *3*.

Huynen, M., Snel, B., Lathe, W., 3rd, and Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome research *10*, 1204-1210.

Janga, S.C., 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 research *33*, 2521-2530.

Janssen, H.J., and Steinbuchel, A. (2014). Fatty acid synthesis in Escherichia coli and its applications towards the production of fatty acid based biofuels. Biotechnology for biofuels *7*, 7.

Junier, I., and Rivoire, O. (2013). Synteny in Bacterial Genomes: Inference, Organization and Evolution. arXiv:13074291.

Junier, I., and Rivoire, O. (2016). Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. PloS one *11*, e0155740.

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 research *40*, D109-114.

Kim, J., and Copley, S.D. (2012). Inhibitory cross-talk upon introduction of a new metabolic pathway into an existing metabolic network. Proceedings of the National Academy of Sciences of the United States of America *109*, E2856-2864.

Kim, P.J., and Price, N.D. (2011). Genetic co-occurrence network across sequenced microbes. PLoS computational biology *7*, e1002340.

King, C.H., Shlaes, D.M., and Dul, M.J. (1983). Infection caused by thymidine-requiring, trimethoprim-resistant bacteria. Journal of clinical microbiology *18*, 79-83.

Kondrashov, A.S., Sunyaev, S., and Kondrashov, F.A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. Proceedings of the National Academy of Sciences of the United States of America *99*, 14878-14883.

Korbel, J.O., Jensen, L.J., von Mering, C., and Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nature biotechnology *22*, 911-917.

Kwon, Y.K., Lu, W., Melamud, E., Khanam, N., Bognar, A., and Rabinowitz, J.D. (2008). A domino effect in antifolate drug action in Escherichia coli. Nature chemical biology *4*, 602-608.

Leduc, D., Escartin, F., Nijhout, H.F., Reed, M.C., Liebl, U., Skouloubris, S., and Myllykallio, H. (2007). Flavin-dependent thymidylate synthase ThyX activity: implications for the folate cycle in bacteria. Journal of bacteriology *189*, 8537-8545.

Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and Ranganathan, R. (2008). Surface sites for engineering allosteric control in proteins. Science (New York, NY) *322*, 438-442.

Lu, W., Clasquin, M.F., Melamud, E., Amador-Noguez, D., Caudy, A.A., and Rabinowitz, J.D. (2010). Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. Analytical chemistry *82*, 3212-3221.

Lu, W., Kwon, Y.K., and Rabinowitz, J.D. (2007). Isotope ratio-based profiling of microbial folates. Journal of the American Society for Mass Spectrometry *18*, 898-909.

McGuire, J.J., and Bertino, J.R. (1981). Enzymatic synthesis and function of folylpolyglutamates. Molecular and cellular biochemistry *38 Spec No*, 19-48.

Michener, J.K., Camargo Neves, A.A., Vuilleumier, S., Bringel, F., and Marx, C.J. (2014a). Effective use of a horizontally-transferred pathway for dichloromethane catabolism requires post-transfer refinement. eLife *3*.

Michener, J.K., Vuilleumier, S., Bringel, F., and Marx, C.J. (2014b). Phylogeny poorly predicts the utility of a challenging horizontally transferred gene in Methylobacterium strains. Journal of bacteriology *196*, 2101-2107.

Nijhout, H.F., Reed, M.C., Budu, P., and Ulrich, C.M. (2004). A mathematical model of the folate cycle: new insights into folate homeostasis. The Journal of biological chemistry *279*, 55008-55016.

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. European journal of biochemistry / FEBS *216*, 539-548.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife *3*, e02030.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America *96*, 2896-2901.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America *96*, 4285-4288.

Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science (New York, NY) *347*, 673-677.

Rengby, O., Johansson, L., Carlson, L.A., Serini, E., Vlamis-Gardikas, A., Karsnas, P., and Arner, E.S. (2004). Assessment of production conditions for efficient use of Escherichia coli in high-yield heterologous recombinant selenoprotein synthesis. Applied and environmental microbiology *70*, 5159-5167.

Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hotspots for allosteric regulation on protein surfaces. Cell *147*, 1564-1575.

Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., and Koonin, E.V. (2002). Connected gene neighborhoods in prokaryotic genomes. Nucleic acids research *30*, 2212-2223.

Sebaugh, J.L. (2011). Guidelines for accurate EC50/IC50 estimation. Pharmaceutical statistics *10*, 128-134.

Segre, D., Deluna, A., Church, G.M., and Kishony, R. (2005). Modular epistasis in yeast metabolism. Nature genetics *37*, 77-83.

Snel, B., Bork, P., and Huynen, M.A. (2002). The identification of functional modules from the genomic association of genes. Proceedings of the National Academy of Sciences of the United States of America *99*, 5890-5895.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P.*, et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research *43*, D447-452.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. Genome biology *2*, RESEARCH0020.

Tas, H., Nguyen, C.T., Patel, R., Kim, N.H., and Kuhlman, T.E. (2015). An Integrated System for Precise Genome Modification in Escherichia coli. PloS one *10*, e0136963.

Toprak, E., Veres, A., Michel, J.B., Chait, R., Hartl, D.L., and Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nature genetics *44*, 101-105.

Toprak, E., Veres, A., Yildiz, S., Pedraza, J.M., 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. Nature protocols *8*, 555-567.

Typas, A., Nichols, R.J., Siegele, D.A., Shales, M., Collins, S.R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B.L.*, et al.* (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. Nature methods *5*, 781-787.

Wagner, G.P., and Altenberg, L. (1996). Perspective: Complex adaptations and the evolution of evolvability. Evolution *50*, 967-976.

Weinreich, D.M., Lan, Y., Wylie, C.S., and Heckendorn, R.B. (2013). Should evolutionary geneticists worry about higher-order epistasis? Current opinion in genetics & development *23*, 700-707.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America *109*, 1193-1198.

Figure Legends

Figure 1 Two representations of folate metabolism. **A**, Biochemical pathway map of folate metabolism. See Table S1 for a more complete description of each enzyme. **B,** Heatmap of synteny couplings between gene pairs in folate metabolism. Pixel intensity shows a measure of significance for the conservation of physical proximity between genes (given as a relative entropy, D_{ij}), assuming a null model in which genes are randomly and uniformly distributed across the chromosome. In *E. coli*, a single gene (*folD*) encodes a bifunctional enzyme that catalyzes both the methylene tetrahydrofolate dehydrogenase (MTD) and methenyltetrahydrofolate cyclohydrolase (MTCH) reactions in the biochemical pathway as shown in A. The majority of gene pairs show little coevolution in terms of gene synteny (dark purple pixels).

Figure 2 Epistatic coupling between DHFR and TYMS. **A,** Genetic barcoding scheme for deep sequencing. Each plasmid contains two barcodes uniquely encoding the identity of *folA* and *thyA* genes. Sequencing of both barcodes enables determination of relative allele frequencies within a population as they vary with time and experimental condition. **B,** Relative allele frequency versus time for a growth competition assay carried out in 50 μg/ml thymidine. The relative fitness of each allele pair is given by the linear slope m. See Fig. S1 for all growth rate fits. **C, D**, Plots of relative growth rate for DHFR mutants spanning a range of catalytic specificities (k_{cat}/K_m), and either a wild-type (WT, grey points) or catalytically dead (R166Q, red points) TYMS. Error bars correspond to standard error across triplicate measurements. For measurements in both 5 and 50 μg/ml thymidine we observe positive or "buffering" epistasis, in

which the cost of reducing the activity of one enzyme is partly or totally mitigated by reducing activity in the other.

Figure 3 A loss-of-function mutation in TYMS buffers metabolic changes from decreased DHFR activity. **A,** Liquid chromatography-mass spectrometry profiling of intracellular folate species in M9 media supplemented with 50 μg/ml thymidine. Rows reflect mutant DHFR/TYMS combinations, columns correspond to metabolites. Each folate species can be modified by the addition of 1-5 glutamates. Square intensity denotes the $\log_2$ abundance of each species relative to wild type. The data show that mutations reducing DHFR activity (G121V, F31Y.L54I, M42F.G121V, and F31Y.G121V) cause an accumulation of DHF and depletion of reduced folate species (THF) (bottom four rows). This effect is partly compensated by an inactivating mutation in TYMS (top four rows). **B,** The corresponding doubling time for each mutant, as measured in batch culture (conditions identical to panel A). See also Fig. S2.

Figure 4 Evolution of trimethoprim (TMP) resistance in *MG1655* cells using the morbidostat. **A,** Schematic of the continuous culture tube. Dilutions are made through the inlet tubes labeled "media A," "media B," and "media C." A constant volume of 15 ml is maintained by the outlet line, labeled "waste," which aspirates extra medium after mixing. The optical density at 600 nm (OD_{600}) of each culture is monitored by an LED-detector pair near the bottom of the tube. Drug concentration is dynamically varied to promote evolution of increased resistance. **B,** Control strategy for the addition of trimethoprim. Once the OD_{600} exceeds 0.06, dilutions of 3 ml were made every 20 minutes. For OD_{600}=0.06-0.15, "media A" is added, which contains no TMP. Above an OD_{600} of 0.15, drug was introduced through dilution with "media

B," which contains a lower amount of TMP. Once the TMP concentration in the culture tube exceeds 60% of the "media B" stock, then "media C," which contains 5X more TMP, is used. Following a decrease in growth rate in response to drug, dilutions resume with "media A". If "media C" was used in a particular day, then the TMP concentration in media "B & C" were incremented by a factor of 5X to enable further adaptation in the following day. **C,** A representative growth trajectory, color-coded by TMP concentration (day 7, 50 μg/ml thymidine). See Figure S3 for full growth trajectories over 13 days. **D**, The trajectory of estimated TMP resistance (as measured by the median TMP concentration on each day) versus number of generations for each experimental condition. Adaptation occurs more rapidly in thymidine-supplemented conditions.

Figure 5 Measurements of phenotype and genotype. Ten single colonies (strains) were selected at the endpoint of each forward evolution condition for phenotyping and genotyping (40 in total). **A,** Trimethoprim (TMP) IC50 measurements for each experimental strain. Error bars represent standard error over triplicate measurements. See also Table S2. **B,** Thymidine dependence for each experimental strain, as determined by total growth in 0 μg/ml thymidine over 10 hours. See Fig. S4 for growth rates across a range of thymidine concentrations. Experimental strains evolved in 5, 10, and 50 μg/ml thymidine are no longer viable in the absence of extracellular thymidine, indicating a loss-of-function mutation in TYMS. **C,** Mutations acquired by each strain during forward evolution. Genes that were mutated two or fewer times across all strains are excluded, as are synonymous mutations (see Table S3 for sequencing statistics, and Table S4 for a complete list of mutants). Gene names are labeled along the left edge of the map, with the corresponding residue or nucleotide change(s) denoted along

the right. If a strain acquires any mutation in a particular gene, the column section corresponding to that gene is shaded blue. Strains evolved in the 5, 10, and 50 μg/ml thymidine conditions acquired mutations in both *folA* and *thyA*, encoding DHFR and TYMS, with a few exceptions lacking a *folA* mutation. A small red arrow indicates one strain with mutations in only DHFR and TYMS. In contrast, the strains sampled from 0 μg/ml thymidine only contain a promoter region mutation in *folA*. See Fig. S5 for a comparison of the trimethoprim resistance of four strains engineered to include only *folA/thyA* mutations and the strains evolved in 5 and 50 μg/ml thymidine.

Figure 6. Genome-wide analysis of pairwise synteny in *E. coli*. A, Enrichment of physical and metabolic interactions as a function of synteny coupling. **B,** A scatter plot of synteny-based coupling for all analyzed gene pairs. Each point represents a pair of genes; coupling within the pair is shown on the x-axis, and the strongest coupling outside of the pair is shown on the y-axis. Color-coding reflects annotations from the STRING database (physical interactions) or KEGG database (metabolic pathways): green gene pairs bind, while pairs in dark blue or light blue do not interact but are found in the same metabolic pathway. Dark blue gene pairs share a metabolic intermediate. The DHFR/TYMS pair is highlighted in red. The orange lines indicate one possible working definition of evolutionary modules: pairs that satisfy the criteria $D_{ij}^{\mathrm{intra}} > 1.0$ and $D_{ij}^{\mathrm{exter}} < 0.5$. See Table S5 for an annotated list of gene pairs below the diagonal, and Table S6 for an analysis of the cutoff dependence of the modularity definition. Figure S6 shows a similar analysis using gene co-occurrence (rather than synteny). **C**, Pie charts showing the distribution of physical and metabolic interactions for: all gene pairs, coupled gene pairs ($D_{ij}^{\mathrm{intra}} > 1.0$) and evolutionary modules ($D_{ij}^{\mathrm{intra}} > 1.0$ and $D_{ij}^{\mathrm{exter}} < 0.5$).

Experimental Procedures

Statistical analysis of gene coevolution. Synteny analysis was conducted using a slightly modified version of the methods described in (Junier and Rivoire, 2013, 2016). See the Supplemental Experimental Procedures for a detailed description of the synteny and co-occurrence calculations.

Forward evolution of trimethoprim resistance in the morbidostat.

The morbidostat/turbidostat apparatus was constructed as described by Toprak and colleagues (Toprak et al., 2013). The founder strain for 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. The goal of this modification was to prevent and detect contamination with other strains. Throughout the forward evolution experiment, cells were grown at 30°C in M9 media supplemented with 0.4% glucose and 0.2% amicase (Sigma); 30 μg/ml of chloramphenicol (Cam) was added for positive selection.

To begin the experiment, the founder 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 0, 5, 10, or 50μg/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 founder strain. The remainder of the overnight cultures was used to inoculate four morbidostat tubes at containing M9 media with varying thymidine supplementation (0, 5, 10, and 50μg/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 $f = 3\ h^{-1}$, to give $r_{dil} = 0.55$. Above the $OD_{600} = 0.15$, these dilutions are used to introduce TMP into the culture (see also Fig. 3B). 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 300μl from the previous culture. Complete trajectories of OD_{600} and drug concentration are shown in Fig. S1. 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.

Whole genome 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 48 strains. Isolation of genomic DNA was performed using the QIAamp DNA Mini Kit (Qiagen). The Nextera XT DNA Library Prep Kit (Illumina) 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 S3. Genome assembly and mutation prediction was performed using *breseq* (Deatherage and Barrick, 2014). 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 founder strain. The

modified reference sequence and all complete genome sequences from the beginning and endpoint of forward evolution are available in the NCBI BioProject database (accession number: PRJNA378892, see also Table S4).

Measurements of thymidine dependence. All strains were grown overnight in LB + 5μg/ml thy, with the exception of the strains evolved in the 50μg/ml thy, which were supplemented with 50μg/ml thy to ensure viability. Cultures were then washed twice in M9 media without thymidine, and inoculated at an OD_{600}=0.005 in 96-well plates containing M9 media supplemented with 10-fold serial dilutions of thymidine, ranging from 0.005 μg/ml to 50 μg/ml (in singlicate). OD_{600} was monitored in a Victor X3 plate reader at 30°C over a period of 20 hours. 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., 2012). For each strain, we identified a start-time (t_0) at the end of lag-phase for the fully-rescued 50μg/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+10 hours using the trapezoid method.

Measurements of trimethoprim resistance (IC50). All strains were grown overnight in LB + 5μg/ml thy, with the exception of the strains evolved in the 50μg/ml thy, which were supplemented with 50μg/ml thy to ensure viability. Each strain was then washed into media conditions corresponding to the strain's forward evolution condition, and adapted for 4 hours at

30°C. The recovery cultures were used to inoculate 96-well plates containing M9 media sampling serial dilutions of TMP (in triplicate), with 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. The trimethoprim resistance of each strain was quantified by its absolute IC50, 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 was quantified using the positive integral of OD600 data over a 10h period of growth (see also the methods for measurement of thymidine dependence). For each strain, we identify a start-time (t_0) at the end of lag-phase for the uninhibited 0μg/ml TMP condition. Growth versus TMP concentration was fit to the above model using MATLAB. IC50 was calculated as the concentration X^* for which growth $Y(X^*) = a/2$.

Growth without trimethoprim selection in 50μg/ml thymidine using the turbidostat

The founder 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 (Sigma); 30 μg/ml of chloramphenicol (CAM) was added for positive selection. To begin the experiment, the founder 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 50μg/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 300µl from the previous culture.

Epistasis Measurements. 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 +30µg/ml Cam +50µg/ml thy cultures. Then, cultures were washed 2x in M9 media supplemented with 0.4% glucose and 0.2% amicase and 30µg/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 1ml 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). We obtained 14,348,937 reads. Data analysis was performed using a series of custom python scripts to count barcodes, and MATLAB to fit relative growth rates.

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. (Tas et al., 2015). In this method, a tetracycline 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 (Genbank accession number GU327533) into electrocompetent *E. coli* MG1655 with a genomic *egfp/cat* resistance cassette (the forward evolution founder strain). For the Δ25-26 TYMS mutation, we introduced the *tetA* landing pad between genome posisitons 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 49,684 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 (RDM, Teknova) 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 Spec, 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 ug/ml thy, and 30 ug/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 13000 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 tributylamine and 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 x 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).

Figure 1
Schober et al.

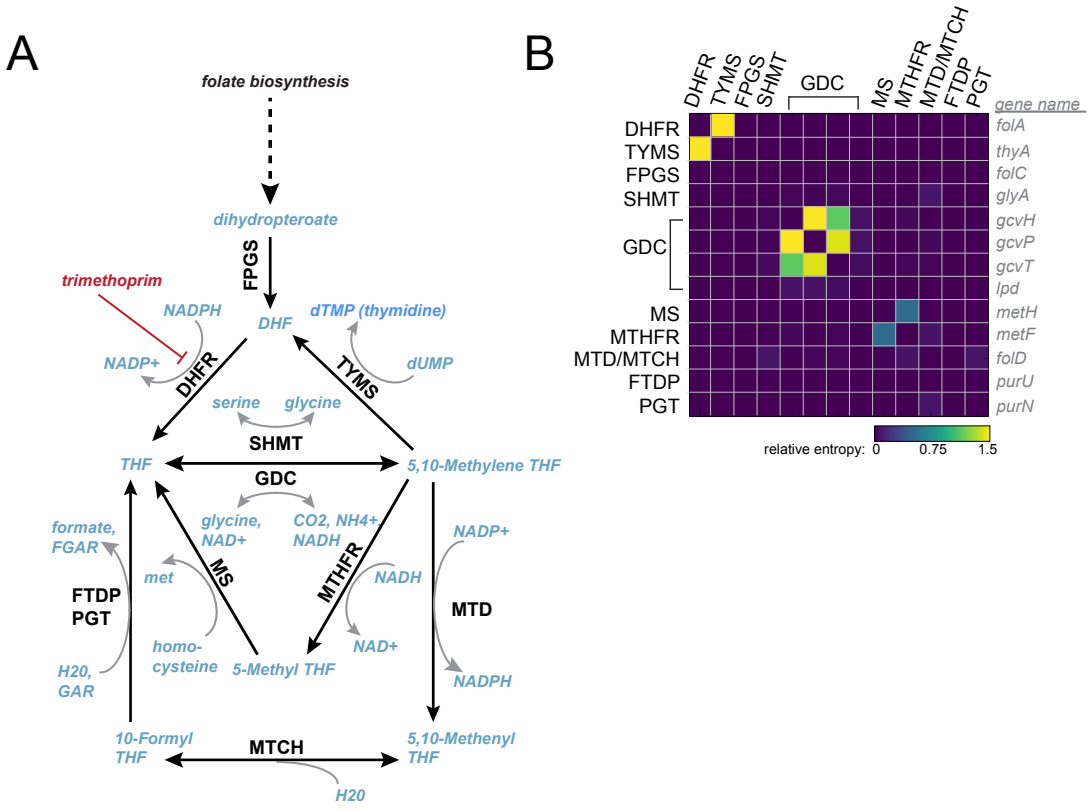


Figure2
Schober et al.

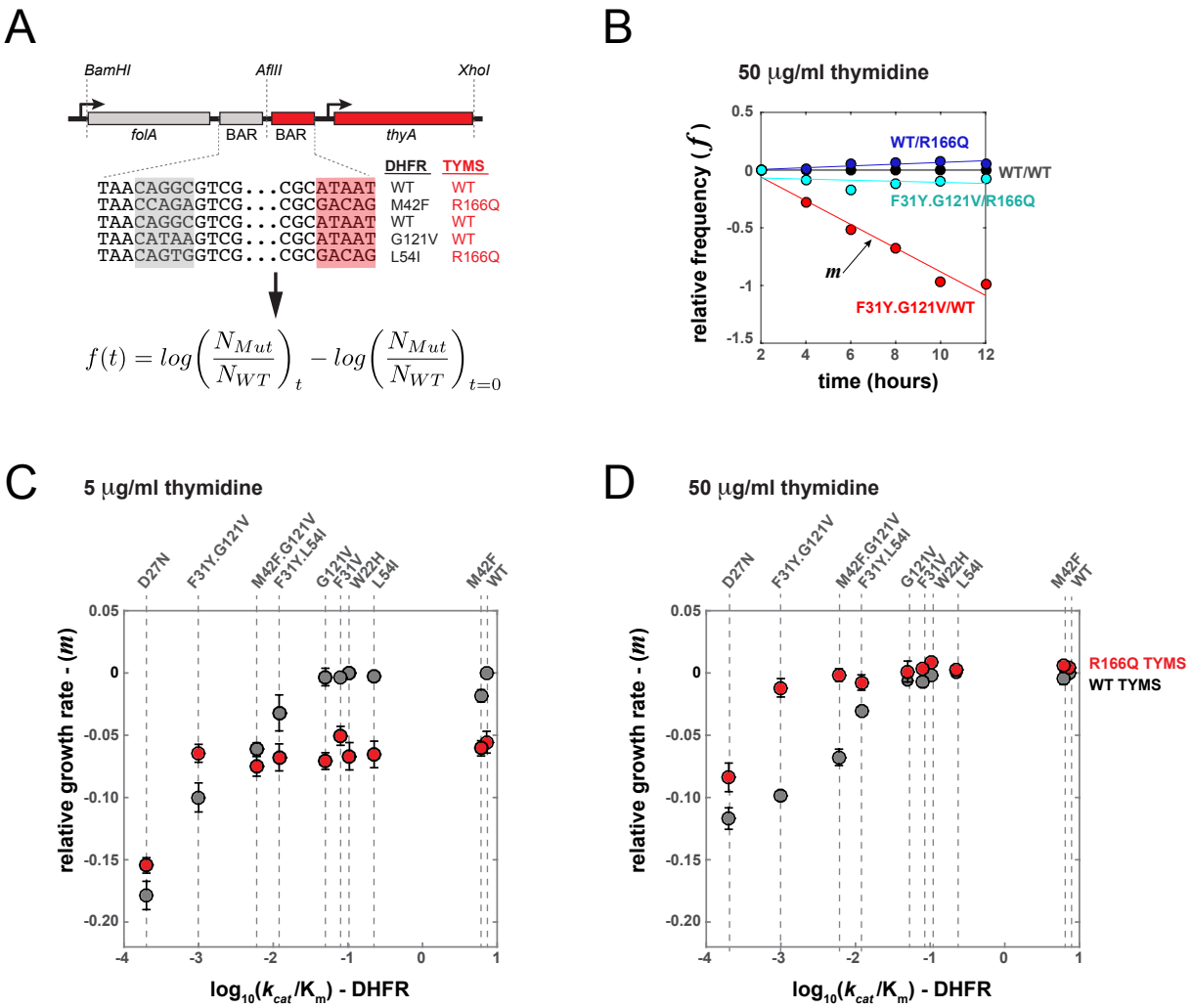


Figure 3
Schober et al.

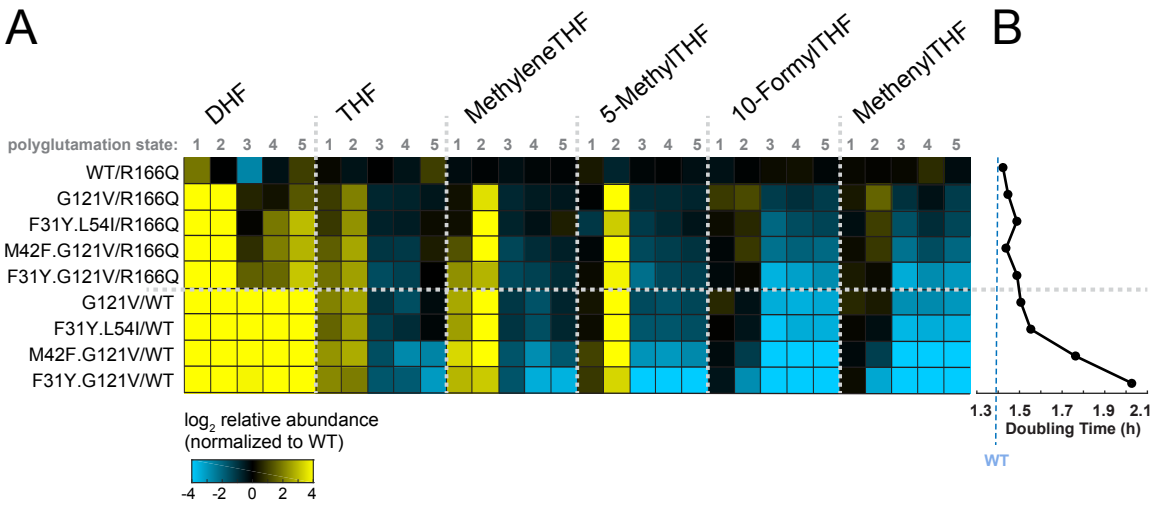


Figure 4
Schober et al.

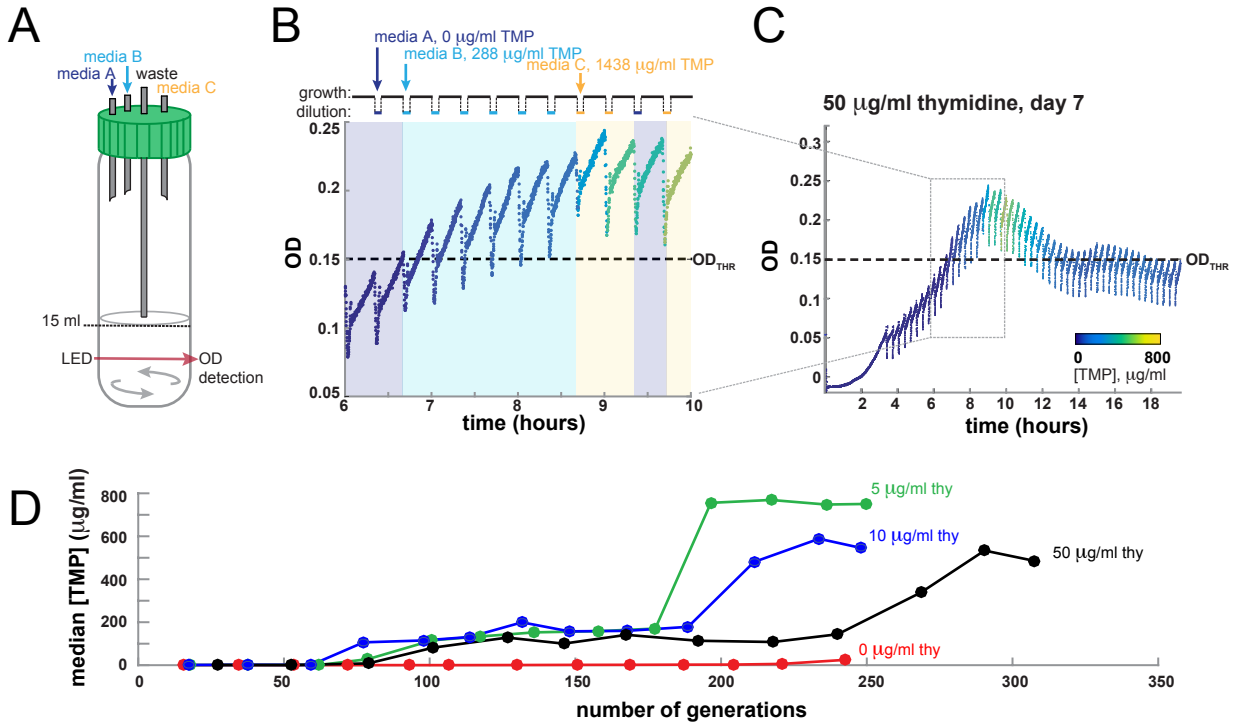


Figure 5
Schober et al.

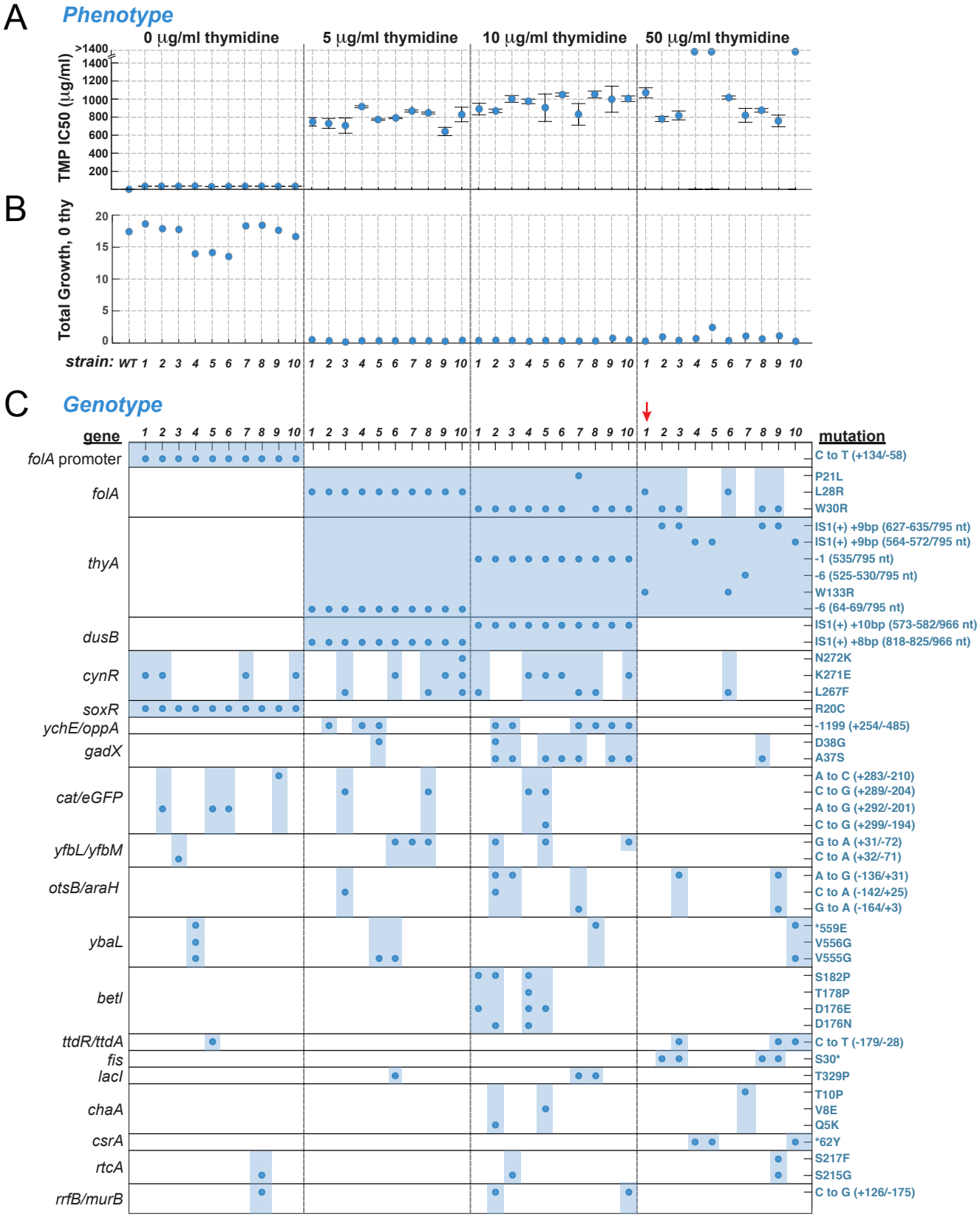
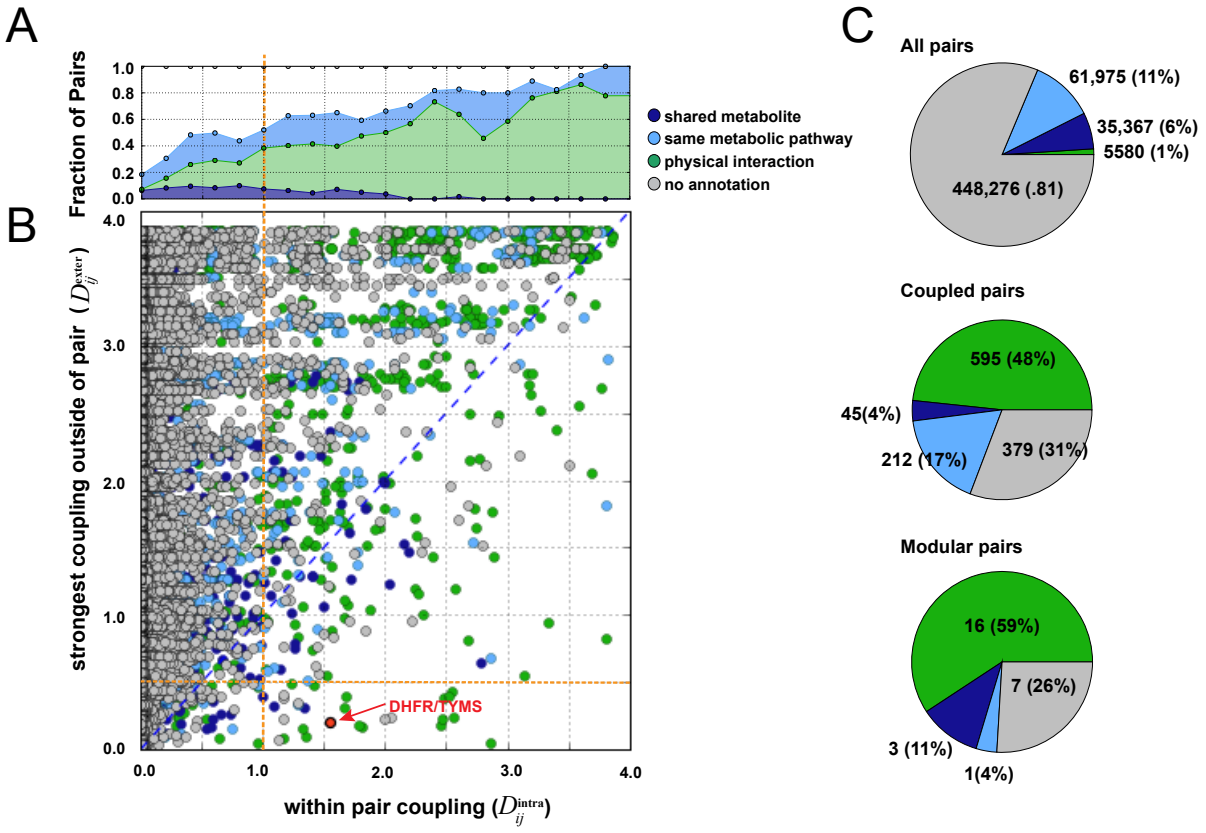


Figure 6
Schober et al.



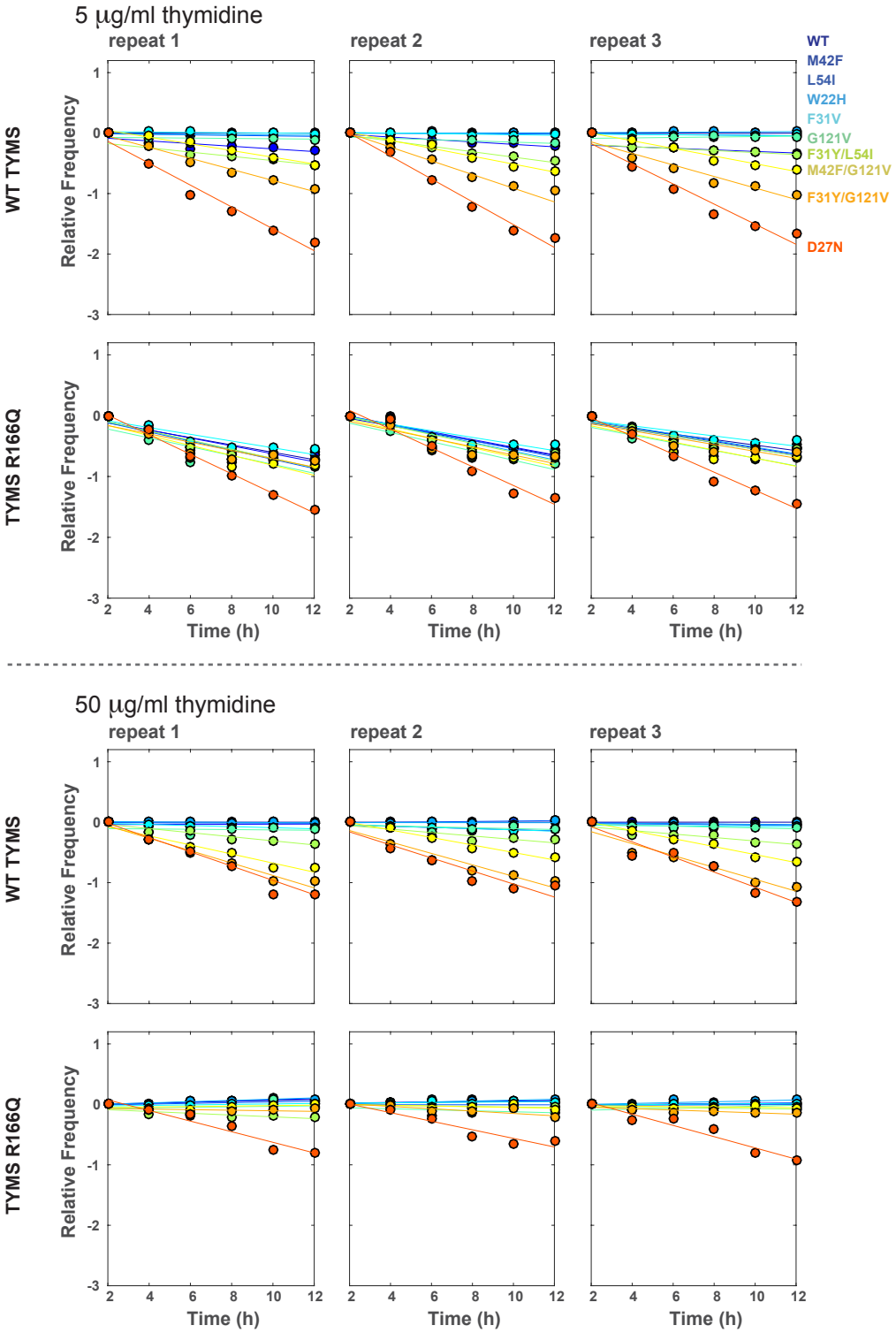


Figure S1. Relative growth rate measurements for DHFR/TYMS mutants. Related to Fig. 2. Points represent the normalized relative frequency (log scale) of DHFR mutants during turbidostat growth in either 5 or 50 µg/ml thymidine. The y-axis indicates the genetic background of TYMS: either WT or R166Q. Relative growth rate fits are shown by the solid lines.

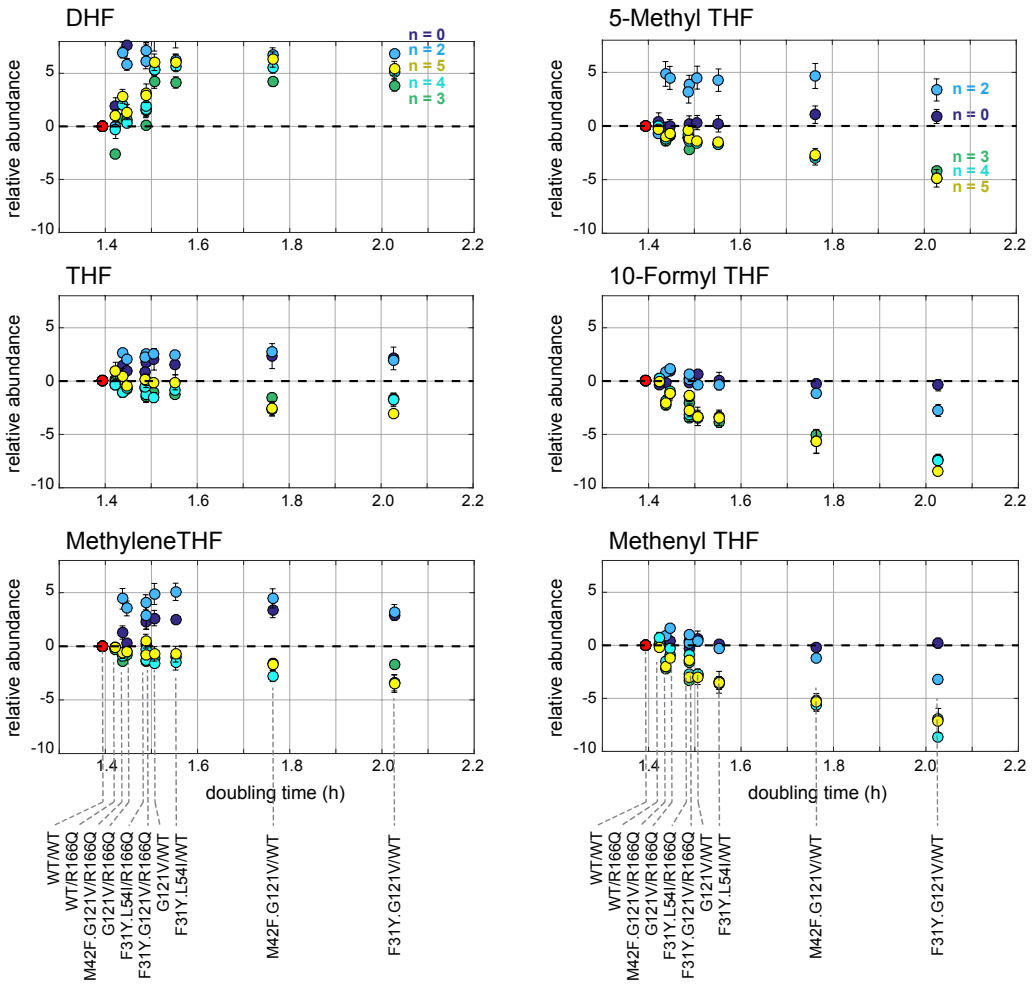


Figure S2. The relationship between intracellular folate species and doubling time. Related to Fig. 3. The y-axis depicts the $\log_2$ abundance of each metabolite normalized to WT. For each folate species, the five glutamylation states are shown in different colors. Doubling times were measured in M9 minimal media supplemented with 50μg/ml thymidine. Error bars denote standard error of the mean across triplicate measurements.

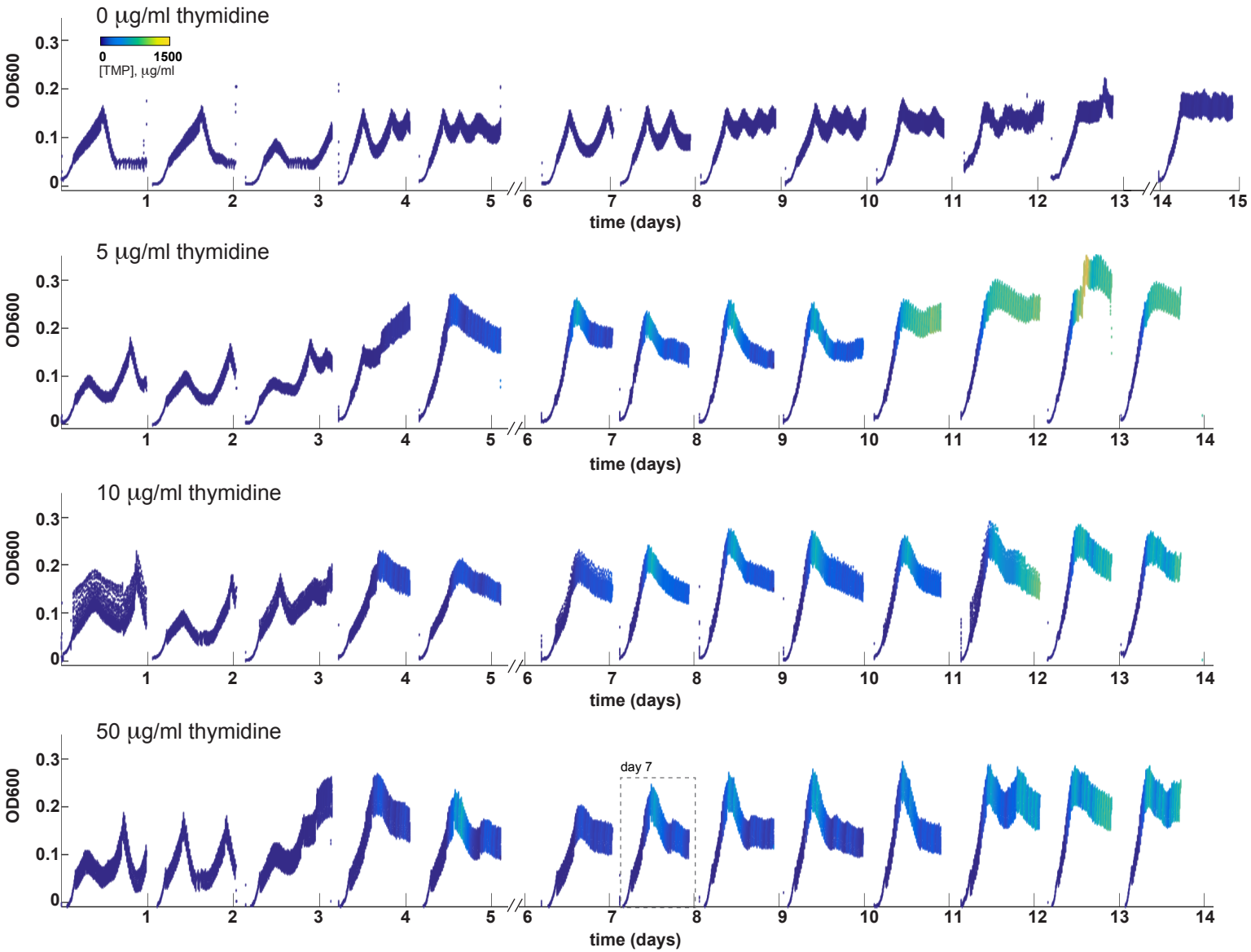


Figure S3. OD600 measurements and trimethoprim concentration over 13 days of forward evolution. Related to Fig. 4. Each plot corresponds to a different thymidine concentration, denoted in the upper left hand corner. The x-axis displays the number of days in real time. Discontinuities at day 5 for all four conditions and day 13 for the 0 μg/ml condition are the result of minor technical problems; cultures were restarted from the previous day's glycerol stock. An enhanced view of the 50 μg/ml trajectory on day 7 (dashed line) is shown in Fig 2B.

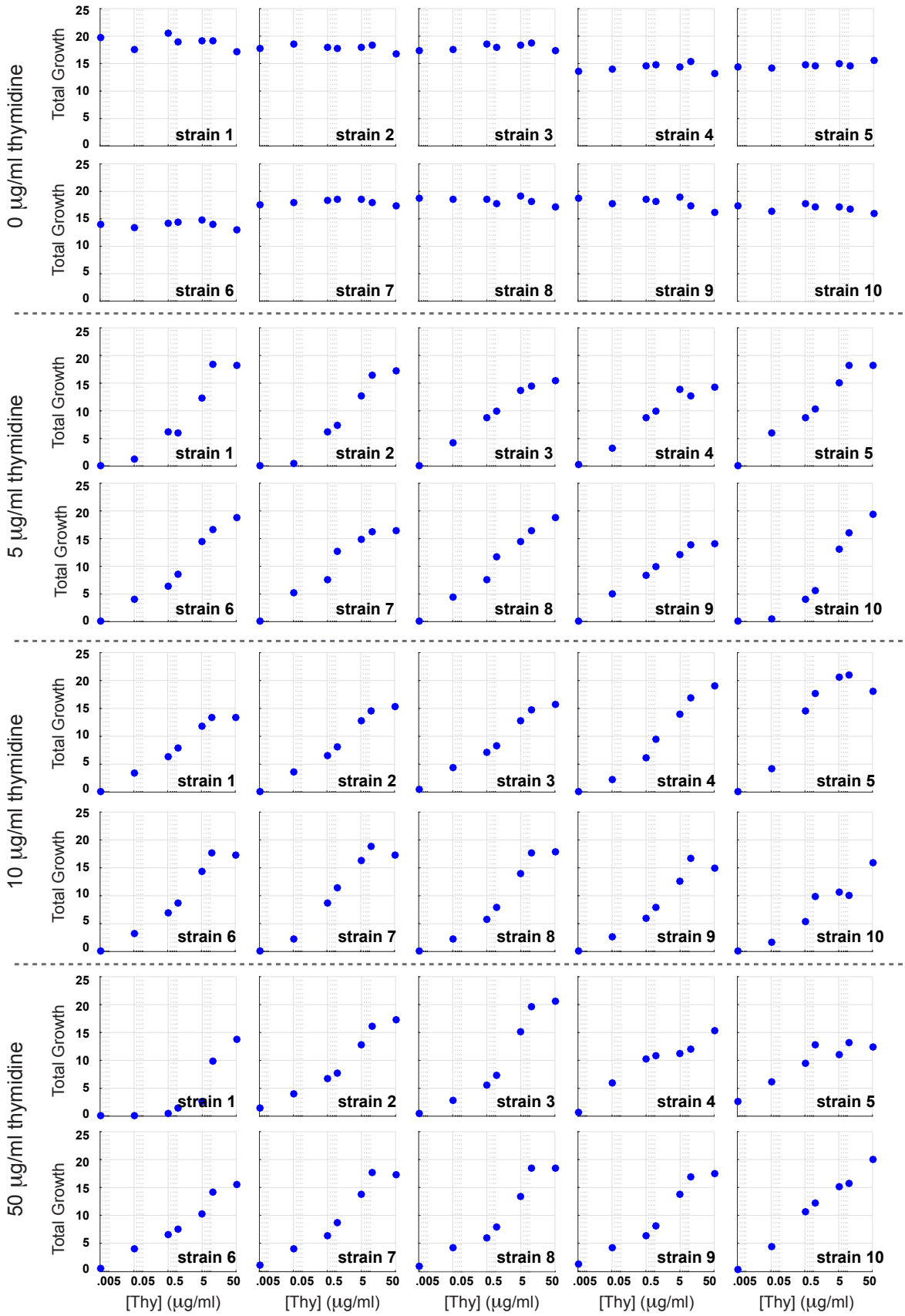


Figure S4. Thymidine dependence of the 40 evolved strains. Related to Fig. 5.
The y-axis denotes the positive integral of log(OD600) evaluated over 20 hours of growth (see Experimental Procedures). Strains evolved in the presence of thymidine become auxotrophs.

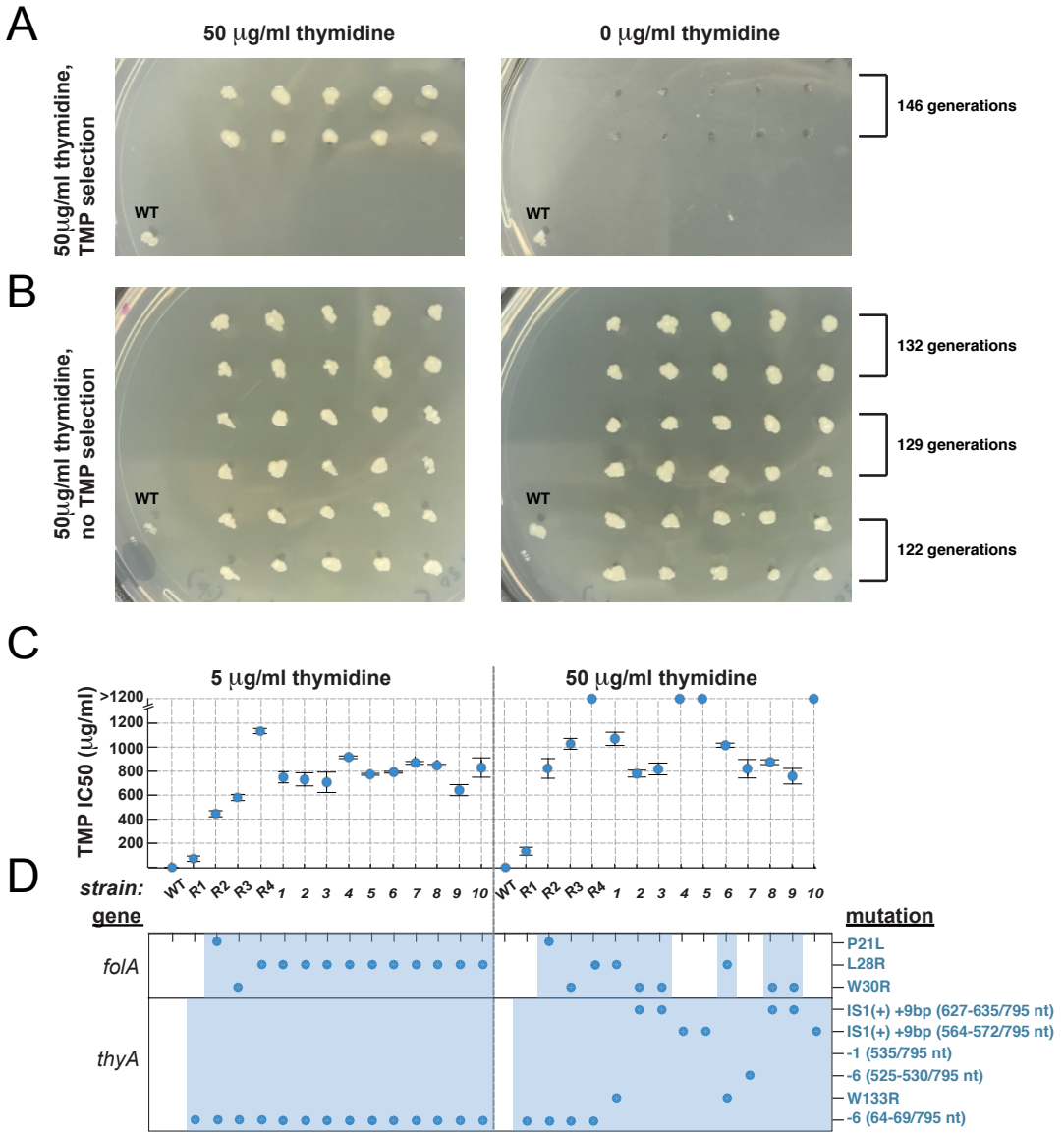


Figure S5. Mutations in folA/thyA are sufficient for TMP resistance. Related to Fig. 5. Loss of function in TYMS due to selection with TMP. **A**, Ten colonies from day 6 of the morbidostat TMP selection (50 μg/ml thymidine condition). Replica plating on 0 and 50 μg/ml thy indicates that all ten strains are thymidine auxotrophs. **B**, Ten colonies from three replicate growths in 50 μg/ml thymidine without TMP selection (turbidostat). These cultures were grown until biofilm formation became prohibitive. Replica plating on 0 and 50 μg/ml thymidine indicates that all strains retain TYMS activity. **C,** Strains R1,R2,R3 and R4 were engineered to contain *folA/thyA* mutations isolated from the morbidostat selection in a clean WT MG1655 *E. coli* background. IC50 measurements were made in both 5 and 50 μg/ml thymidine. The strains obtained from forward evolution in 5 or 50 μg/ml thymidine are shown for comparison. Error bars indicate standard error across triplicate measurements. **D,** Mutations for each strain are indicated with a blue circle. (mutations outside of the *folA/thyA* loci in the forward evolution strains are omitted for clarity).

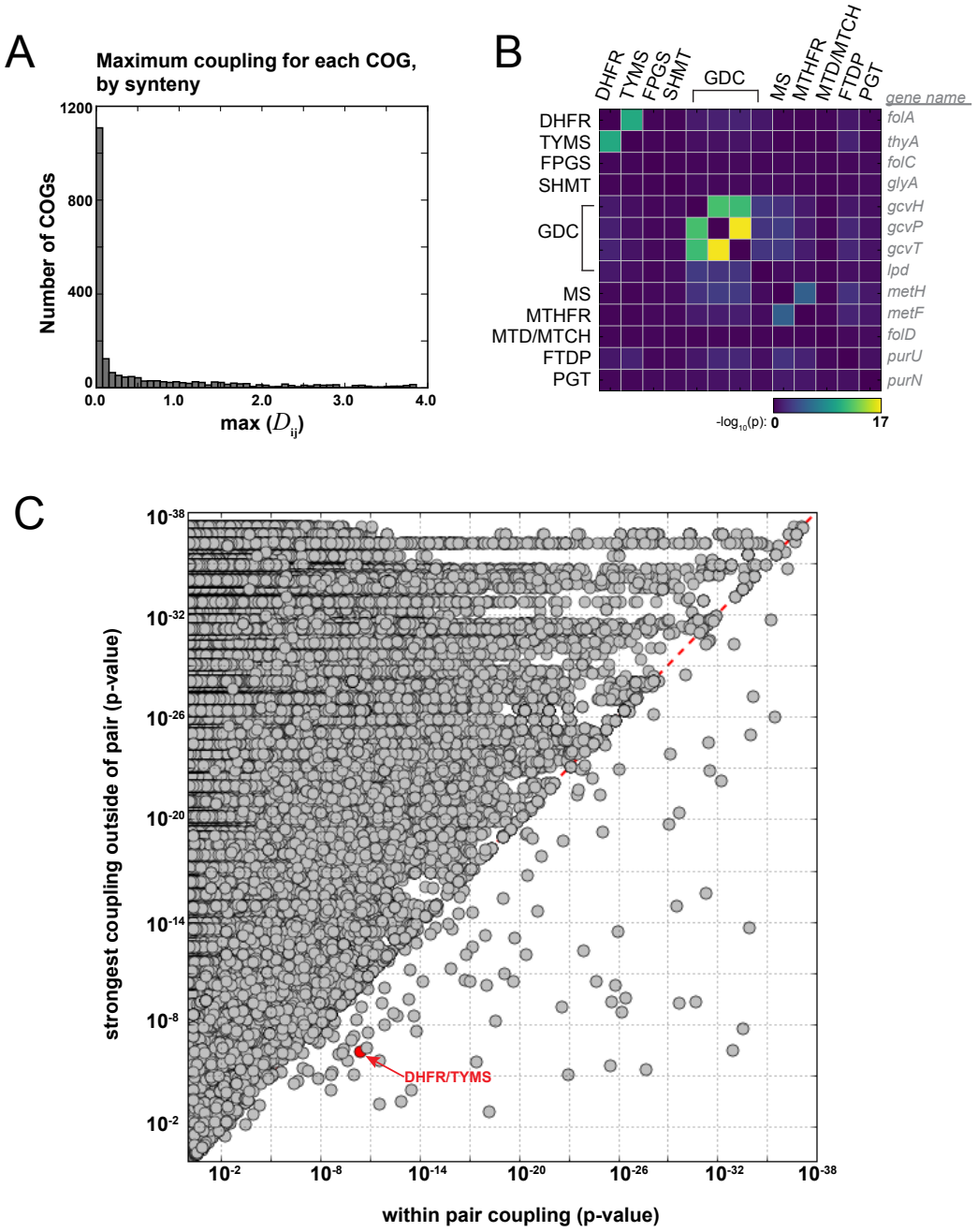


Figure S6. Additional analysis of evolutionary coupling. Related to Fig. 6.
A, The distribution of COGs as a function of maximum synteny coupling (to any other COG). The majority of COGs are not strongly coupled to any other gene. **B**, Heatmap of co-occurence couplings between gene pairs in folate metabolism. **C**, A scatter plot of co-occurrence based couplings across 3528 COGs. DHFR and TYMS are more strongly coupled to each other than any other COG in our dataset.

Supplemental Information For:

A conserved functional module in central metabolism

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

[1]*The Green Center for Systems Biology, and* [2]*Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390*

[3]*Lewis-Sigler Institute for Integrative Genomics,* [4]*Department of Chemical and Biological Engineering, and* [5]*Department of Chemistry, Princeton University, Princeton, NJ 08544*

[6]*Centre National de la Recherche Scientifique (CNRS), and* [7]*Université Grenoble Alpes, TIMC-IMAG, F-38000 Grenoble, France*

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

Contents

I. Supplementary Experimental Procedures

A. Analysis of synteny

II. Supplementary Tables

A. TableS1: Enzymes in central folate metabolism

B. TableS2: Trimethoprim resistance (IC50) for forward evolution strains

C. TableS3: Whole genome sequencing statistics

D. TableS4: Annotated list of genes mutated during the forward evolution experiment

E. TableS5: Modular enzyme pairs identified by synteny

F. TableS6: Evolutionary modules as a function of cutoff

Synteny calculations

A. Starting dataset

Calculating synteny requires 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.

B. Counting pairs in co-occurrence

Synteny is only relevant for the subset of genomes where both orthology classes are present. Thus, we begin by counting the number of genomes where orthology classes i and j co-occur. As previously published (Junier and Rivoire, 2016), we correct for the uneven phylogenetic distribution of sequenced genomes (strains) by introducing genome weights. To this end, we compute 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 the average sequence similarity). The weight w_s of strain s is then defined as $1/n_s$ where n_s is the number of strains within a given distance δ of s. Varying δ can provide information at different "phylogenetic depths" (Junier and Rivoire, 2013) but here we fix $\delta = 0.3$, 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

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

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, $\mathbb{1}[i \cap s \neq \emptyset] = 1$ if i is represented in strain s and 0 otherwise.

C. Defining gene proximity

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 $p^* = d^*/(L/2)$.

D. Counting pairs in synteny

The value p^* provides a measure of signifcance for finding two genes at a distance d^* in one genome. However, we are interested in the conservation of proximity across many species. To begin, we count the effective number of strains in which i and j are within a given distance d^*.

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

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 for the probability of observing a positive event under the null model to be common for all strains, we instead consider the normalized distance and compute:

$$X_{ij} = \sum_s w_s \mathbb{1}[2d(i, j)/L_s < p^*]. \tag{3}$$

For strains that contain multiple chromosomes, we take 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 take $p^* = 0.02$, corresponding to $d = 50$ kb in the context of a chromosome of length 5 Mb. This cutoff is chosen to represent a length scale longer than those typical for gene coexpression 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 ij, we correct Eq. (3) by averaging over all these pairs:

$$X_{ij} = \sum_s w_s \frac{1}{|i \cap s||j \cap s|} \sum_{g_i \in i \cap s, g_j \in j \cap s} \mathbb{1}[2d(g_i, g_j)/L_s < p^*], \tag{4}$$

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

E. Measuring significance

Now that we have counted the number of genomes in which i and j are proximal, we can assess the significance of this result. In a standard statistics "coin toss" problem, one computes the significance of obtaining X "tails" out of M "flips" (given a probability of tails $p^* = 0.5$) using the binomial distribution. Here, we compute the significance of finding a pair of genes in proximity X_{ij} times out of M_{ij} genomes (given a probability of $p^* = 0.02$) using the same approach:

$$\pi_{ij} = \sum_{K \geq X_{ij}} \binom{M_{ij}}{K} (p^*)^K (1 - p^*)^{M'_{ij} - K} = I(X_{ij}, M_{ij} - X_{ij} + 1, p^*) \tag{5}$$

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 apply the Bonferroni principle - we set here a threshold of significance to $\pi^* = 2/N(N-1) \simeq 10^{-7}$ where $N = 4764$ is the number of orthology classes defined by COGs. That is, we choose a cutoff such that we should not find 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).

F. Degree of synteny

The p-values π_{ij} depend on the number of genomes in the dataset. It is more meaningful to define a measure of conservation that depends only on rescaled variables, here the frequencies $f_{ij} = X_{ij}/M_{ij}$. For these frequencies to be meaningful, we need, however, to restrict to cases where the number M_{ij} of genomes where genes i and j co-occur is large. Here, we restrict to pairs of COGs with $M_{ij} \geq 100$. A degree of synteny is then given by the relative entropy:

$$D(f_{ij} \| p^*) = f_{ij} \ln \frac{f_{ij}}{p^*} + (1 - f_{ij}) \ln \frac{1 - f_{ij}}{1 - p^*}. \tag{6}$$

In the limit of large M_{ij}, $e^{-M_{ij} D_{ij}}$ approximates the first term of the sum in Eq. (5) 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.

G. Application to *E. coli*

To analyze synteny relationships relevant to *E. coli*, we keep only the COGs i that are represented in its genome, and analyze COG pairs for which $M_{ij} \geq 100$ (2095 COGs in total). In Fig. 6B, we plot for each pair ij of these COGs their degree of synteny D_{ij} (x-axis) against their maximal degree of synteny with any other COG $\max_{k \neq i,j}(D_{ik}, D_{jk})$ (y-axis). We define two-gene modules as all pairs where the within-pair coupling $D_{ij} > 1$, and the maximum coupling outside of the pair $D_{ij} < 0.5$. In this figure, we use the String database to annotate physical interactions, taking a threshold of 700 and the largest score when multiple paralogs are present.

Abbreviation	Name	E. coli gene	Uniprot ID	COG
FPGS	bifunctional dihydropteroate synthase/FPGS	*folC*	P08192	COG0285H
DHFR	dihydrofolate reductase	*folA*	P0ABQ4	COG0262H
TYMS	thymidylate synthase	*thyA*	P0A884	COG0207F
SHMT	serine hydroxymethyltransferase	*glyA*	P0A825	COG0112E
GDC	glycine cleavage system - protein H	*gcvH*	P0A6T9	COG0509E
	glycine cleavage system - protein L	*lpdA*	P0A9P0	COG1249C
	glycine cleavage system - protein P	*gcvP*	P33195	COG1003E
	glycine cleavage system - protein T	*gcvT*	P27248	COG0404E
MS	methionine synthase	*metH*	P13009	COG1410E
MTHFR	5,10-methylenetetrahydrofolate reductase	*metF*	P0AEZ1	COG0685E
MTD	methylenetetrahydrofolate dehydrogenase (bifunctional)	*folD*	P24186	COG0190H
MTCH	methenyltetrahydrofolate cyclohydrolase (bifunctional)	*folD*	P24186	COG0190H
FTDP	formyltetrahydrofolate deformylase	*purU*	P37051	COG0788F
PGT	phosphoribosylglycinamide formyltransferase	*purN*	P08179	COG0299F

Table S1 **Enzymes in central folate metabolism**. Related to Figure 1.

Evolved	Strain	IC50 (μg/ml)	Std Err
0 thy	1	34	0.6
	2	35	1.1
	3	36	0.83
	4	37	0.65
	5	31	0.22
	6	33	0.5
	7	35	1.1
	8	35	0.78
	9	33	1
	10	34	0.34
5 thy	1	750	21
	2	720	24
	3	710	85
	4	910	11
	5	770	7.7
	6	790	7.3
	7	870	12
	8	840	10
	9	640	46
	10	830	81
10 thy	1	890	63
	2	870	19
	3	1000	37
	4	970	25
	5	900	150
	6	1000	18
	7	830	120
	8	1100	37
	9	1000	140
	10	1000	31
50 thy	1	1100	55
	2	780	29
	3	820	49
	4	NA	NA
	5	NA	NA
	6	1000	18
	7	820	77
	8	870	20
	9	760	65
	10	NA	NA

Founder	Strain	IC50 (μg/ml)	Std Err
0 thy	1	0.93	0.11
	2	0.99	0.095
5 thy	1	0.91	0.036
	2	1.1	0.043
10 thy	1	0.86	0.0071
	2	0.95	0.11
50 thy	1	1.2	0.037
	2	1.3	0.091

Table S2 **Trimethoprim resistance (IC50) for forward evolution strains**. Related to Figure 3. Standard error is calculated across triplicate measurements. An estimate could not be obtained for the strains 4, 5, and 7 evolved in 50 μg/ml thymidine because these showed slow growth regardless of TMP concentration (trimethoprim insensitive).

Evolved	Strain	Coverage	Dispersion (σ^2/μ)	Reads	Avg read length (BP)
0 thy	1	40	3.0	1.46E+06	128
	2	47	3.2	1.72E+06	126
	3	47	3.0	1.73E+06	128
	4	47	3.0	1.57E+06	139
	5	43	2.7	1.57E+06	128
	6	58	4.0	2.15E+06	124
	7	60	3.8	2.22E+06	125
	8	55	3.5	2.01E+06	127
	9	67	3.6	2.52E+06	123
	10	57	3.3	2.05E+06	128
5 thy	1	55	4.4	2.07E+06	126
	2	40	3.4	1.52E+06	125
	3	40	3.4	1.43E+06	131
	4	35	3.5	1.46E+06	114
	5	29	2.3	9.85E+05	138
	6	38	3.0	1.45E+06	122
	7	46	3.0	1.53E+06	142
	8	68	4.1	2.30E+06	137
	9	52	3.2	1.70E+06	140
	10	52	4.2	1.80E+06	134
10 thy	1	32	3.4	1.13E+06	135
	2	34	3.3	1.16E+06	138
	3	34	3.4	1.19E+06	135
	4	48	3.9	1.67E+06	135
	5	42	3.6	1.46E+06	133
	6	31	2.9	1.12E+06	132
	7	29	2.9	1.03E+06	131
	8	25	2.6	8.43E+05	137
	9	31	2.9	1.07E+06	135
	10	41	3.4	1.44E+06	135
50 thy	1	42	3.5	1.46E+06	133
	2	35	3.4	1.32E+06	126
	3	33	3.0	1.16E+06	134
	4	21	3.1	8.79E+05	116
	5	18	3.7	7.93E+05	115
	6	25	3.5	9.90E+05	121
	7	24	2.9	8.87E+05	125
	8	31	3.2	1.13E+06	130
	9	29	2.9	1.00E+06	136
	10	24	3.2	9.36E+05	126

Founders	Strain	Coverage	Dispersion (σ^2/μ)	Reads	Avg read length (BP)
0 thy	1	61	3.2	2.18E+06	129
	2	49	3.3	1.86E+06	123
5 thy	1	44	3.1	1.53E+06	134
	2	50	3.2	1.92E+06	122
10 thy	1	36	2.2	1.32E+06	129
	2	40	2.5	1.45E+06	130
50 thy	1	35	2.3	1.18E+06	137
	2	38	2.5	1.26E+06	138

Table S3 **Whole genome sequencing statistics**. Related to Figure 3. Coverage is the average number of reads aligned to a particular position in the genome. Dispersion is the variance of read coverage normalized by the mean.

Name	Mutation		Strains	Annotation
kefC/folA	C to T	(+134/-58)	0thy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	potassium:proton antiporter/dihydrofolate reductase
folA	CCG to CTG	P21L	10thy: 7	dihydrofolate reductase
	CTC to CGC	L28R	5thy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; 50thy: 1, 6	
	TGG to AGG	W30R	10thy: 1, 2, 3, 4, 5, 6, 8, 9, 10; 50thy: 2, 3, 8, 9	
thyA	IS1(+) +9bp	(627-635/795 nt)	50thy: 2, 3, 8, 9	thymidylate synthetase
	IS1(+) +9bp	(564-572/795 nt)	50thy: 4, 5, 10	
	Δ1	(535/795 nt)	10thy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	
	Δ6	(525-530/795 nt)	50thy: 7	
	TGG to AGG	W133R	50thy: 1, 6	
	Δ6	(64-69/795 nt)	5thy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	
dusB	IS1(+) +10bp	(573-582/966 nt)	10thy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	tRNA-dihydrouridine synthase B
	IS1(+) +8bp	(818-825/966 nt)	5thy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	
cynR	AAT to AAA	N272K	5thy: 10	transcriptional activator of cyn operon; autorepressor
	AAA to GAA	K271E	0thy: 1, 2, 7, 10; 5thy: 6, 9, 10; 10thy: 4, 5, 6, 10	
soxR	CGC to TGC	R20C	0thy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	redox-sensitive transcriptional activator of soxS; autorepressor
gadX	GAT to GGT	D38G	5thy: 5; 10thy: 2	acid resistance regulon transcriptional activator; autoactivator
	GCG to TCG	A37S	10thy: 2, 3, 5, 6, 7, 9, 10; 50thy: 8	
cat/egfp	A to C	(+283/-210)	0thy: 9	chloramphenicol acetyltransferase/green fluorescent protein
	A to G	(+292/-201)	0thy: 2, 5, 6	
	C to G	(+299/-194)	10thy: 5	
yfbL/yfbM	G to A	(+31/-72)	5thy: 6, 7, 8; 10thy: 2, 5, 10	putative M28A family peptidase/DUF1877 family protein
	C to A	(+32/-71)	0thy: 3	
otsB/araH	A to G	(-136/+31)	10thy: 2, 3; 50thy: 3, 9	trehalose-6-phosphate phosphatase, biosynthetic/L-arabinose ABC transporter permease
	C to A	(-142/+25)	5thy: 3; 10thy: 2	
	G to A	(-164/+3)	10thy: 7; 50thy: 9	
ybaL	TAA to GAA	*559E	0thy: 4; 10thy: 8; 50thy: 10	inner membrane putative NAD(P)-binding transporter
	GTG to GGG	V556G	0thy: 4	
	GTG to GGG	V555G	0thy: 4; 5thy: 5, 6	
betI	TCC to CCC	S182P	10thy: 1, 2, 4	choline-inducible betIBA-betT divergent operon transcriptional repressor
	ACC to CCC	T178P	10thy: 4	
	GAT to GAA	D176E	10thy: 1, 4, 5	
	GAT to AAT	D176N	10thy: 2, 4	
ttdR/ttdA	C to T	(-179/-28)	0thy: 5	transcriptional activator of ttdABT/L-tartrate dehydratase, alpha subunit
fis	TCG to TAG	S30*	50thy: 2, 3, 8, 9	global DNA-binding transcriptional dual regulator

lacI	CCC to CCA	P332P	10thy: 8	lactose-inducible lac operon transcriptional repressor
	ACC to CCC	T329P	5thy: 6; 10thy: 7, 8	
chaA	ACC to CCC	T10P	50thy: 7	calcium/sodium:proton antiporter
	GTA to GAA	V8E	10thy: 5	
	CAA to AAA	Q5K	10thy: 2	
csrA	TAA to TAC	*62Y	50thy: 4, 5, 10	pleiotropic regulatory protein for carbon source metabolism
rtcA	AGT to GGT	S215G	0thy: 8; 10thy: 3; 50thy: 9	RNA 3'-terminal phosphate cyclase
rrfB/murB	C to G	(+126/-175)	0thy: 8; 10thy: 2, 10	5S ribosomal RNA of rrnB operon/UDP-N-acetylenolpyruvoylglucosamine reductase, FAD-binding
citG	ACC to CCC	T255P	5thy: 2; 10thy: 3	2-(5"-triphosphoribosyl)-3'-dephosphocoenzyme-A synthase
ompF/asnS	G to T	(-529/+74)	10thy: 2	outer membrane porin 1a (Ia;b;F)/asparaginyl tRNA synthetase
	C to A	(-540/+63)	10thy: 8	
lpoB	CAA to CAC	Q38H	5thy: 6; 50thy: 9	OM lipoprotein stimulator of MrcB transpeptidase
sapA	ACC to CCC	T304P	5thy: 2; 50thy: 4	antimicrobial peptide transport ABC transporter periplasmic binding protein
yddE	CAA to CAC	Q14H	50thy: 4, 5	PhzC-PhzF family protein
	ACC to CCC	T12P	50thy: 4	
yghQ	GTG to GGG	V332G	5thy: 7; 10thy: 10	putative inner membrane polysaccharide flippase
	GGA to GGG	G323G	10thy: 10	
agaD	GGA to GGG	G120G	10thy: 1	N-acetylgalactosamine-specific enzyme IID component of PTS
	GCC to TCC	A126S	10thy: 1, 3	
gntR	GAA to GGA	E147G	10thy: 4, 5	d-gluconate inducible gluconate regulon transcriptional repressor
	GTG to GGG	V146G	10thy: 4	
yiaK	ACC to CCC	T309P	10thy: 7; 50thy: 6	2,3-diketo-L-gulonate reductase, NADH-dependent
	GAA to AAA	E313K	10thy: 7	
ampC	GTA to GGA	V48G	10thy: 1, 8	penicillin-binding protein; beta-lactamase, intrinsically weak
thrC	CTC to ATC	L3I	50thy: 7	L-threonine synthase
dapB/carA	T to A	(+301/-155)	50thy: 6	dihydrodipicolinate reductase/carbamoyl phosphate synthetase small subunit, glutamine amidotransferase
paoC	CAA to AAA	Q72K	50thy: 9	PaoABC aldehyde oxidoreductase, Moco-containing subunit
acrR	IS1(+) +9bp	(320-328/648 nt)	50thy: 7	transcriptional repressor
ybdK	TGG to CGG	W263R	5thy: 5	weak gamma-glutamyl:cysteine ligase
dtpD/ybgI	T to A	(-84/-187)	10thy: 3	dipeptide and tripeptide permease D/NIF3 family metal-binding protein
ssuB	GGC to GGG	G44G	10thy: 3, 8	aliphatic sulfonate ABC transporter ATPase
putP	GAT to GGT	D55G	50thy: 3	proline:sodium symporter
serX	A to G	(72/88 nt)	10thy: 8	tRNA-Ser
flgF	CAG to CGG	Q19R	50thy: 10	flagellar component of cell-proximal portion of basal-body rod
pabC	TAC to GAC	Y92D	10thy: 3	4-amino-4-deoxychorismate lyase component of para-aminobenzoate synthase multienzyme complex

dadX	ACC to CCC	T284P	5thy: 2	alanine racemase, catabolic, PLP-binding
oppF	CCG to CAG	P273Q	10thy: 3	oligopeptide ABC transporter ATPase
uspF/ompN	G to A	(-108/+33)	5thy: 7	stress-induced protein, ATP-binding protein/outer membrane pore protein N, non-specific
yneM/dgcZ	G to A	(+75/+144)	10thy: 3	inner membrane-associated protein/diguanylate cyclase, zinc-sensing
yebV/yebW	G to T	(+26/-79)	10thy: 7	uncharacterized protein/uncharacterized protein
araH	CAA to AAA	Q322K	10thy: 2	L-arabinose ABC transporter permease
mntH	GTG to GGG	V313G	10thy: 1	manganese/divalent cation transporter
xapR	ATG to ATA	M176I	5thy: 6	transcriptional activator of xapAB
uraA	ATT to GTT	I311V	50thy: 7	uracil permease
relA	CAT to CAA	H518Q	5thy: 3	(p)ppGpp synthetase I/GTP pyrophosphokinase
ptrA	GAT to GAA	D38E	10thy: 1	protease III
	GAT to AAT	D38N	10thy: 1	
	CGT to CGA	R35R	10thy: 1	
rsmI	CAT to CAA	H235Q	10thy: 6	16S rRNA C1402 2'-O-ribose methyltransferase, SAM-dependent
gltF/yhcA	Δ4	(+90/-79)	50thy: 7	periplasmic protein/putative periplasmic chaperone protein
fis-yhdX	Δ9555		50thy: 7	fis, yhdJ, yhdU, acrS, acrE, acrF, yhdV, yhdW, yhdX
acrS	TAT to TTT	Y187F	10thy: 3	acrAB operon transcriptional repressor
secY	GTA to GGA	V274G	10thy: 2	preprotein translocase membrane subunit
xylF	GAA to AAA	E195K	50thy: 6	D-xylose transporter subunit
uhpT	GAA to GGA	E447G	50thy: 8	hexose phosphate transporter
pstA	GGT to GGG	G112G	50thy: 5	phosphate ABC transporter permease
	ATT to GTT	I106V	50thy: 5	
yiiQ	AAC to AAA	N156K	0thy: 5	DUF1454 family putative periplasmic protein
pyrB	ACC to CCC	T54P	50thy: 4	aspartate carbamoyltransferase, catalytic subunit

Table S4 **Annotated list of genes mutated during the forward evolution experiment**. Related to Figure 3. Mutations identified in any of the founder strains are omitted. The first column indicates the affected gene; two names with a slash indicate neighboring genes to the affected intergenic region (ordered 5' to 3' along the sense strand). For proteins, both the codon change and amino acid change are included, synonymous mutations are omitted (an asterisk * indicates a stop codon). For intergenic mutations, the base change(s) and position relative to each neighboring gene are displayed. Insertion-sequence mediated changes are preceded with "IS#."

Table S5 **Modular enzyme pairs identified by synteny**

Sorted by distance from the diagonal. Pairs identified as modular according to the thresholds in Fig. 6 are highlighted in orange.

COG 1	gene 1	COG 2	gene 2	physical interaction	shared metabolite	in-pair	out-pair
COG0103	*rpsI*	COG0102	*rplM*	in interaction	no shared intermediate	3.81	0.82
COG1271	*appC*	COG1294	*appB*	in interaction	no shared intermediate	2.87	0.04
COG1220	*hslU*	COG5405	*hslV*	in interaction	no shared intermediate	3.14	0.55
COG0719	*sufB*	COG0396	*sufC*	in interaction	no shared intermediate	3.33	0.95
COG0459	*groL*	COG0234	*groS*	in interaction	no shared intermediate	2.56	0.23
COG1108	*znuB*	COG1121	*znuC*	in interaction	no shared intermediate	2.46	0.18
COG1838	*ttdB*	COG1951	*fumB*	in interaction	shared int.	2.47	0.23
COG2884	*ftsE*	COG2177	*ftsX*	no interaction	no shared intermediate	2.86	0.68
COG0041	*purE*	COG0026	*purK*	no interaction	shared int.	2.79	0.64
COG3261	*hycE*	COG3260	*hycG*	in interaction	no shared intermediate	2.52	0.40
COG0074	*sucD*	COG0045	*sucC*	in interaction	shared int.	2.55	0.43
COG2025	*ygcQ*	COG2086	*ygcR*	in interaction	no shared intermediate	2.48	0.38
COG0261	*rplU*	COG0211	*rpmA*	no interaction	no shared intermediate	3.81	1.82
COG0752	*glyQ*	COG0751	*glyS*	in interaction	shared int.	2.67	0.82
COG1918	*feoA*	COG0370	*feoB*	no interaction	no shared intermediate	2.05	0.24
COG1203	*ygcB*	COG1518	*ygbT*	no interaction	no shared intermediate	2.00	0.23
COG0048	*rpsL*	COG0049	*rpsG*	in interaction	no shared intermediate	3.69	2.05
COG0420	*sbcD*	COG0419	*sbcC*	in interaction	no shared intermediate	1.80	0.17
COG0208	*nrdF*	COG0209	*nrdE*	in interaction	shared int.	1.79	0.18
COG0184	*rpsO*	COG1185	*pnp*	in interaction	no shared intermediate	2.89	1.43
COG0052	*rpsB*	COG0264	*tsf*	no interaction	no shared intermediate	3.51	2.12
COG0725	*modA*	COG4149	*modB*	in interaction	no shared intermediate	2.57	1.20
COG0732	*hsdS*	COG0286	*hsdM*	in interaction	no shared intermediate	1.67	0.32
COG0262	*folA*	COG0207	*thyA*	no interaction	shared int.	1.55	0.20
COG1291	*motA*	COG1360	*motB*	in interaction	no shared intermediate	2.31	1.01
COG0505	*carA*	COG0458	*carB*	in interaction	shared int.	1.68	0.39
COG1702	*phoH*	COG0319	*ybeY*	no interaction	no shared intermediate	2.80	1.51
COG2104	*thiS*	COG2022	*thiG*	no interaction	no shared intermediate	2.21	0.95
COG1826	*tatB*	COG0805	*tatC*	in interaction	no shared intermediate	1.36	0.10
COG0004	*amtB*	COG0347	*glnB*	in interaction	no shared intermediate	1.33	0.16
COG1516	*fliS*	COG1345	*fliD*	no interaction	no shared intermediate	2.71	1.54
COG0066	*leuD*	COG0065	*leuC*	in interaction	shared int.	2.28	1.16
COG1843	*flgD*	COG1749	*flgE*	in interaction	no shared intermediate	2.88	1.77
COG1740	*hybO*	COG0374	*hybC*	in interaction	shared int.	2.74	1.65
COG0851	*minE*	COG2894	*minD*	in interaction	no shared intermediate	3.77	2.68
COG0168	*trkH*	COG0569	*trkA*	in interaction	no shared intermediate	1.17	0.08
COG1585	*ybbJ*	COG0330	*hflC*	no interaction	no shared intermediate	1.12	0.07
COG1077	*mreB*	COG1792	*mreC*	in interaction	no shared intermediate	2.25	1.20
COG0086	*rpoC*	COG0085	*rpoB*	in interaction	shared int.	3.59	2.55
COG0321	*lipB*	COG0320	*lipA*	no interaction	shared int.	1.34	0.31
COG0238	*rpsR*	COG0360	*rpsF*	in interaction	no shared intermediate	3.41	2.37
COG1925	*npr*	COG1080	*ptsA*	in interaction	no shared intermediate	1.53	0.55
COG0072	*pheT*	COG0016	*pheS*	in interaction	shared int.	2.39	1.43
COG1057	*nadD*	COG0799	*rsfS*	no interaction	no shared intermediate	1.84	0.88
COG0280	*eutD*	COG0282	*tdcD*	no interaction	no shared intermediate	1.28	0.33
COG0159	*trpA*	COG0133	*trpB*	in interaction	shared int.	2.73	1.79
COG1138	*nrfE*	COG0755	*ccmC*	no interaction	no shared intermediate	1.17	0.23
COG1921	*selA*	COG3276	*selB*	no interaction	no shared intermediate	1.29	0.35
COG0848	*exbD*	COG0811	*exbB*	in interaction	no shared intermediate	1.41	0.48
COG0292	*rplT*	COG0291	*rpmI*	no interaction	no shared intermediate	3.82	2.90
COG0333	*rpmF*	COG1399	*yceD*	no interaction	no shared intermediate	3.10	2.19
COG1703	*argK*	COG1884	*scpA*	in interaction	no shared intermediate	0.96	0.05
COG0718	*ybaB*	COG0353	*recR*	no interaction	no shared intermediate	2.60	1.71
COG0742	*rsmD*	COG0669	*coaD*	no interaction	no shared intermediate	1.45	0.57
COG0194	*gmk*	COG1561	*yicC*	no interaction	no shared intermediate	2.67	1.82
COG0245	*ispF*	COG1211	*ispD*	no interaction	no shared intermediate	1.36	0.52
COG0263	*proB*	COG0014	*proA*	in interaction	shared int.	1.88	1.03
COG1825	*rplY*	COG0193	*pth*	no interaction	no shared intermediate	2.29	1.47
COG0689	*rph*	COG0127	*rdgB*	no interaction	no shared intermediate	1.43	0.62
COG2145	*thiM*	COG0352	*thiE*	no interaction	shared int.	2.03	1.23
COG0437	*nrfC*	COG5557	*hybB*	no interaction	no shared intermediate	1.17	0.39
COG0391	*ybhK*	COG1660	*yhbJ*	no interaction	no shared intermediate	2.32	1.54
COG0060	*ileS*	COG0597	*lspA*	no interaction	no shared intermediate	1.17	0.40
COG0248	*gpp*	COG0855	*ppk*	no interaction	shared int.	1.09	0.33
COG1923	*hfq*	COG0324	*miaA*	no interaction	no shared intermediate	1.70	0.95
COG0341	*secF*	COG0342	*secD*	in interaction	no shared intermediate	3.27	2.52
COG0439	*accC*	COG0511	*accB*	in interaction	no shared intermediate	1.23	0.49
COG0149	*tpiA*	COG1314	*secG*	no interaction	no shared intermediate	1.57	0.84
COG2087	*cobU*	COG0368	*cobS*	no interaction	shared int.	2.19	1.47
COG0528	*pyrH*	COG0233	*frr*	no interaction	no shared intermediate	2.83	2.12
COG0443	*hscA*	COG0576	*grpE*	in interaction	no shared intermediate	1.67	0.98
COG0539	*rpsA*	COG0283	*cmk*	no interaction	no shared intermediate	1.44	0.76
COG0216	*prfA*	COG2890	*prmB*	in interaction	no shared intermediate	1.56	0.88
COG1137	*lptB*	COG1934	*lptA*	in interaction	no shared intermediate	2.26	1.58
COG2137	*recX*	COG0468	*recA*	in interaction	no shared intermediate	1.41	0.74
COG1975	*yqeB*	COG2068	*mocA*	no interaction	no shared intermediate	1.49	0.86
COG0059	*ilvC*	COG0440	*ilvN*	no interaction	shared int.	2.15	1.53
COG2804	*gspE*	COG1459	*gspF*	in interaction	no shared intermediate	1.69	1.08
COG1177	*ydcV*	COG1176	*ydcU*	in interaction	no shared intermediate	2.76	2.16
COG0173	*yfeK*	COG0124	*hisS*	no interaction	shared int.	1.00	0.39
COG0407	*hemE*	COG0276	*hemH*	no interaction	no shared intermediate	0.74	0.15
COG1219	*clpX*	COG0544	*tig*	no interaction	no shared intermediate	2.53	1.96
COG1126	*yhdZ*	COG0765	*yhdY*	in interaction	no shared intermediate	1.67	1.11
COG0636	*atpE*	COG0356	*atpB*	in interaction	no shared intermediate	3.19	2.64
COG1136	*lolD*	COG4591	*lolE*	in interaction	no shared intermediate	0.64	0.09
COG0150	*purM*	COG0299	*purN*	no interaction	no shared intermediate	1.44	0.89
COG0297	*glgA*	COG0448	*glgC*	no interaction	shared int.	1.14	0.59
COG2009	*sdhC*	COG0479	*frdB*	in interaction	no shared intermediate	2.69	2.15

COG 1	gene 1	COG 2	gene 2	physical interaction	shared metabolite	in-pair	out-pair
COG0409	*hypD*	COG0298	*hybG*	in interaction	no shared intermediate	2.25	1.72
COG0165	*argH*	COG0137	*argG*	no interaction	shared int.	1.26	0.76
COG1977	*moaD*	COG0314	*moaE*	in interaction	no shared intermediate	0.95	0.46
COG0470	*holB*	COG0125	*yghT*	no interaction	no shared intermediate	0.96	0.47
COG0594	*rnpA*	COG0706	*yidC*	no interaction	no shared intermediate	2.91	2.43
COG0044	*iadA*	COG0540	*pyrB*	no interaction	no shared intermediate	1.09	0.61
COG0548	*argB*	COG0002	*argC*	no interaction	shared int.	1.26	0.79
COG0083	*thrB*	COG0498	*thrC*	no interaction	shared int.	0.73	0.26
COG0581	*pstA*	COG0573	*pstC*	in interaction	no shared intermediate	3.18	2.71
COG1674	*ftsK*	COG2834	*lolA*	no interaction	no shared intermediate	0.76	0.30
COG1043	*lpxA*	COG0764	*fabA*	no interaction	shared int.	1.71	1.26
COG0363	*yieK*	COG0364	*zwf*	no interaction	no shared intermediate	1.03	0.58
COG0329	*yjhH*	COG0289	*dapB*	no interaction	shared int.	0.97	0.53
COG0057	*epd*	COG0126	*pgk*	in interaction	shared int.	1.28	0.84
COG2011	*metI*	COG1135	*metN*	in interaction	no shared intermediate	2.92	2.49
COG1385	*rsmE*	COG2264	*prmA*	no interaction	no shared intermediate	0.92	0.49
COG0802	*tsaE*	COG1214	*tsaB*	in interaction	no shared intermediate	1.02	0.60
COG0414	*panC*	COG0853	*panD*	no interaction	shared int.	1.49	1.07
COG1558	*flgC*	COG1815	*flgB*	no interaction	no shared intermediate	3.63	3.21
COG0055	*atpD*	COG0355	*atpC*	in interaction	no shared intermediate	3.20	2.79
COG0516	*guaB*	COG0519	*guaA*	no interaction	shared int.	0.57	0.17
COG0244	*rplJ*	COG0222	*rplL*	in interaction	no shared intermediate	3.71	3.31
COG1806	*ppsR*	COG0574	*ppsA*	no interaction	no shared intermediate	1.16	0.77
COG1127	*mlaF*	COG0767	*mlaE*	no interaction	no shared intermediate	1.77	1.38
COG0421	*speE*	COG1586	*speD*	no interaction	shared int.	0.65	0.26
COG1410	*metH*	COG0685	*metF*	no interaction	shared int.	0.53	0.16
COG0713	*nuoK*	COG0839	*nuoJ*	in interaction	shared int.	3.16	2.79
COG0155	*cysI*	COG0175	*cysH*	no interaction	shared int.	0.89	0.53
COG1932	*serC*	COG0111	*serA*	no interaction	shared int.	0.58	0.23
COG1116	*ssuB*	COG0600	*ssuC*	in interaction	no shared intermediate	1.85	1.51
COG1120	*fecE*	COG0609	*fecC*	in interaction	no shared intermediate	1.55	1.21
COG1173	*dppC*	COG0601	*dppB*	in interaction	no shared intermediate	1.26	0.93
COG2127	*clpS*	COG2360	*aat*	no interaction	no shared intermediate	0.83	0.50
COG1438	*argR*	COG0497	*recN*	no interaction	no shared intermediate	1.53	1.22
COG0823	*yidR*	COG1729	*ybgF*	no interaction	no shared intermediate	0.77	0.46
COG1905	*nuoE*	COG1894	*nuoF*	in interaction	shared int.	1.96	1.65
COG0106	*hisA*	COG0131	*hisB*	no interaction	no shared intermediate	2.07	1.77
COG2001	*mraZ*	COG0275	*rsmH*	no interaction	no shared intermediate	3.45	3.15
COG0069	*gltB*	COG0493	*gltD*	in interaction	shared int.	0.53	0.23
COG1195	*recF*	COG0592	*dnaN*	no interaction	no shared intermediate	1.77	1.47
COG0247	*glcF*	COG1139	*ykgF*	no interaction	no shared intermediate	0.45	0.16
COG0375	*hybF*	COG0378	*hypB*	no interaction	no shared intermediate	1.74	1.46
COG0260	*pepA*	COG0795	*lptG*	no interaction	no shared intermediate	0.37	0.10
COG1319	*ygfM*	COG2080	*xdhC*	in interaction	no shared intermediate	1.91	1.64
COG0317	*spoT*	COG1490	*dtd*	no interaction	no shared intermediate	0.66	0.39
COG1256	*flgK*	COG1551	*csrA*	no interaction	no shared intermediate	1.96	1.70

COG 1	gene 1	COG 2	gene 2	physical interaction	shared metabolite	in-pair	out-pair
COG0304	*fabB*	COG0236	*acpP*	in interaction	no shared intermediate	1.32	1.06
COG0543	*fre*	COG0167	*preA*	in interaction	shared int.	0.74	0.49
COG0550	*topB*	COG0758	*smf*	no interaction	no shared intermediate	0.80	0.55
COG0836	*cpsB*	COG1089	*gmd*	no interaction	shared int.	0.59	0.35
COG0701	*yraQ*	COG0640	*arsR*	no interaction	no shared intermediate	0.45	0.22
COG1338	*fliP*	COG1987	*fliQ*	no interaction	no shared intermediate	3.10	2.87
COG0664	*crp*	COG1151	*hcp*	no interaction	no shared intermediate	0.30	0.06
COG0161	*bioA*	COG0132	*ynfK*	no interaction	shared int.	1.24	1.01
COG2513	*prpB*	COG0372	*gltA*	no interaction	no shared intermediate	0.33	0.12
COG0801	*folK*	COG1539	*folB*	no interaction	shared int.	0.99	0.78
COG0228	*rpsP*	COG0806	*rimM*	no interaction	no shared intermediate	3.05	2.85
COG0128	*aroA*	COG0287	*tyrA*	no interaction	no shared intermediate	0.97	0.76
COG4775	*bamA*	COG2825	*skp*	in interaction	no shared intermediate	1.47	1.26
COG2255	*ruvB*	COG0632	*ruvA*	in interaction	shared int.	1.99	1.79
COG0710	*aroD*	COG0169	*aroE*	no interaction	shared int.	0.81	0.61
COG0838	*nuoA*	COG0377	*nuoB*	in interaction	shared int.	2.70	2.50
COG0643	*cheA*	COG2201	*cheB*	in interaction	no shared intermediate	1.56	1.36
COG1074	*recB*	COG0507	*recD*	in interaction	no shared intermediate	0.30	0.11
COG1937	*rcnR*	COG2217	*zntA*	no interaction	no shared intermediate	0.29	0.11
COG0691	*smpB*	COG0557	*rnr*	no interaction	no shared intermediate	0.57	0.39
COG0274	*deoC*	COG0213	*deoA*	no interaction	no shared intermediate	0.59	0.41
COG0112	*glyA*	COG0698	*rpiB*	no interaction	no shared intermediate	0.51	0.34
COG0237	*coaE*	COG0749	*polA*	no interaction	no shared intermediate	0.55	0.38
COG0568	*rpoH*	COG0358	*dnaG*	no interaction	no shared intermediate	0.92	0.75
COG0195	*nusA*	COG0779	*rimP*	no interaction	no shared intermediate	3.51	3.35
COG2919	*ftsB*	COG0148	*eno*	no interaction	no shared intermediate	0.86	0.70
COG0736	*acpS*	COG0063	*nnr*	no interaction	no shared intermediate	0.74	0.60
COG0279	*diaA*	COG0241	*gmhB*	no interaction	no shared intermediate	0.83	0.68
COG0080	*rplK*	COG0081	*rplA*	in interaction	no shared intermediate	3.52	3.38
COG0524	*yihV*	COG0800	*dgoA*	no interaction	shared int.	0.33	0.20
COG0825	*accA*	COG0777	*accD*	in interaction	shared int.	0.76	0.63
COG1587	*hemD*	COG0181	*hemC*	no interaction	shared int.	1.41	1.29
COG3383	*fdhF*	COG1526	*fdhD*	no interaction	no shared intermediate	0.51	0.39
COG0084	*yjjV*	COG0143	*metG*	no interaction	no shared intermediate	0.50	0.38
COG0554	*glpK*	COG0578	*glpD*	no interaction	shared int.	1.41	1.29
COG2878	*rsxB*	COG0177	*nth*	no interaction	no shared intermediate	0.23	0.12
COG0445	*mnmG*	COG0357	*rsmG*	no interaction	no shared intermediate	1.22	1.11
COG2877	*kdsA*	COG0504	*pyrG*	no interaction	shared int.	0.83	0.72
COG1596	*wza*	COG2148	*wcaJ*	no interaction	no shared intermediate	0.52	0.41
COG0583	*yjiE*	COG2855	*yeiH*	no interaction	no shared intermediate	0.14	0.03
COG0508	*sucB*	COG0567	*sucA*	in interaction	shared int.	0.58	0.47
COG0325	*yggS*	COG1496	*yfiH*	no interaction	no shared intermediate	0.50	0.40
COG0337	*aroB*	COG0703	*aroK*	no interaction	no shared intermediate	1.01	0.91
COG1995	*pdxA*	COG0030	*rsmA*	no interaction	no shared intermediate	0.54	0.44
COG0043	*ubiD*	COG0163	*ubiX*	in interaction	no shared intermediate	0.28	0.19
COG0381	*wecB*	COG0677	*wecC*	no interaction	shared int.	0.50	0.41

COG 1	gene 1	COG 2	gene 2	physical interaction	shared metabolite	in-pair	out-pair
COG0272	*ligB*	COG0210	*uvrD*	no interaction	no shared intermediate	0.19	0.09
COG0054	*ribE*	COG0307	*ribC*	no interaction	shared int.	1.70	1.61
COG0242	*def*	COG0223	*fmt*	no interaction	no shared intermediate	1.04	0.94
COG1129	*ytfR*	COG1172	*yjfF*	in interaction	no shared intermediate	1.51	1.42
COG1488	*pncB*	COG1335	*yecD*	no interaction	shared int.	0.21	0.12
COG0746	*mobA*	COG1763	*mobB*	in interaction	no shared intermediate	0.42	0.34
COG0328	*rnhA*	COG2334	*rdoA*	no interaction	no shared intermediate	0.18	0.10
COG1003	*gcvP*	COG0509	*gcvH*	no interaction	no shared intermediate	1.52	1.43
COG0845	*mdtE*	COG0841	*mdtF*	in interaction	no shared intermediate	0.35	0.27
COG0227	*rpmB*	COG1200	*recG*	no interaction	no shared intermediate	0.65	0.57
COG1886	*fliN*	COG1868	*fliM*	in interaction	no shared intermediate	2.26	2.18
COG0395	*ugpE*	COG1175	*malF*	in interaction	no shared intermediate	1.03	0.95
COG1045	*cysE*	COG0031	*cysM*	in interaction	shared int.	0.32	0.24
COG0178	*uvrA*	COG0556	*uvrB*	in interaction	no shared intermediate	0.49	0.42
COG0203	*rplQ*	COG0100	*rpsK*	in interaction	no shared intermediate	3.85	3.77
COG0046	*purL*	COG0152	*purC*	no interaction	shared int.	0.82	0.75
COG0164	*rnhB*	COG0792	*yraN*	no interaction	no shared intermediate	1.18	1.11
COG3956	*mazG*	COG1188	*hslR*	no interaction	no shared intermediate	0.78	0.71
COG1162	*rsgA*	COG0036	*sgcE*	no interaction	no shared intermediate	0.69	0.63
COG1250	*fadB*	COG0183	*fadA*	in interaction	shared int.	0.30	0.24
COG0782	*greB*	COG1190	*lysU*	no interaction	no shared intermediate	0.42	0.36
COG0602	*nrdG*	COG0603	*queC*	no interaction	shared int.	0.98	0.92
COG1841	*rpmD*	COG0200	*rplO*	in interaction	no shared intermediate	3.73	3.67
COG0712	*atpH*	COG0056	*atpA*	in interaction	no shared intermediate	3.27	3.22
COG0527	*lysC*	COG0136	*asd*	no interaction	shared int.	0.43	0.38
COG1570	*xseA*	COG1722	*xseB*	in interaction	no shared intermediate	1.85	1.80
COG0465	*ftsH*	COG0037	*ttcA*	no interaction	no shared intermediate	0.77	0.72
COG0088	*rplD*	COG0089	*rplW*	in interaction	no shared intermediate	3.86	3.81
COG1622	*cyoA*	COG0843	*cyoB*	in interaction	shared int.	2.04	1.99
COG0229	*msrB*	COG0225	*msrA*	no interaction	shared int.	0.11	0.06
COG0040	*hisG*	COG0141	*hisD*	no interaction	no shared intermediate	1.15	1.10
COG0092	*rpsC*	COG0255	*rpmC*	in interaction	no shared intermediate	3.86	3.82
COG1597	*yegS*	COG1409	*cpdA*	no interaction	no shared intermediate	0.07	0.03
COG1070	*rhaB*	COG0235	*ulaF*	no interaction	shared int.	0.15	0.11
COG0547	*trpD*	COG0134	*trpC*	no interaction	shared int.	1.83	1.79
COG0256	*rplR*	COG0096	*rpsH*	in interaction	no shared intermediate	3.72	3.67
COG1131	*rbbA*	COG0842	*yhhJ*	in interaction	no shared intermediate	0.32	0.28
COG4177	*livM*	COG0559	*livH*	in interaction	no shared intermediate	1.70	1.66
COG0787	*alr*	COG2337	*chpB*	no interaction	no shared intermediate	0.40	0.37

COG 1	gene 1	COG 2	gene 2	physical interaction	shared metabolite	in-pair	out-pair
COG0481	*lepA*	COG0635	*hemN*	no interaction	no shared intermediate	0.49	0.45
COG1555	*ybaV*	COG2333	*ycaI*	no interaction	no shared intermediate	0.79	0.76
COG1317	*fliH*	COG1766	*fliF*	no interaction	no shared intermediate	3.18	3.15
COG0743	*dxr*	COG0575	*cdsA*	no interaction	no shared intermediate	2.37	2.34
COG0187	*gyrB*	COG0188	*parC*	in interaction	no shared intermediate	1.01	0.98
COG0659	*dauA*	COG0288	*cynT*	no interaction	no shared intermediate	0.06	0.03
COG0461	*pyrE*	COG0284	*pyrF*	in interaction	shared int.	0.63	0.61
COG0469	*pykA*	COG0205	*pfkA*	no interaction	shared int.	0.31	0.29
COG1559	*yceG*	COG0816	*yqgF*	no interaction	no shared intermediate	0.92	0.90
COG1576	*rlmH*	COG1235	*phnP*	no interaction	no shared intermediate	0.40	0.38
COG1575	*menA*	COG0318	*aas*	no interaction	shared int.	0.11	0.09
COG1252	*ndh*	COG1477	*apbE*	no interaction	no shared intermediate	0.09	0.07
COG1528	*ftnA*	COG0450	*ahpC*	no interaction	no shared intermediate	0.09	0.08
COG1251	*nirB*	COG2223	*yhjX*	no interaction	no shared intermediate	0.15	0.13
COG1088	*rffG*	COG1898	*rfbC*	no interaction	shared int.	2.00	1.98
COG1328	*yjjI*	COG1180	*yjjW*	no interaction	no shared intermediate	0.41	0.39
COG0709	*selD*	COG0425	*tusA*	no interaction	no shared intermediate	0.37	0.35
COG1519	*waaA*	COG1663	*lpxK*	no interaction	shared int.	0.45	0.44
COG0157	*nadC*	COG0379	*nadA*	no interaction	shared int.	1.30	1.29
COG0863	*yhdJ*	COG1194	*mutY*	no interaction	no shared intermediate	0.12	0.10
COG1092	*rlmI*	COG2606	*yeaK*	no interaction	no shared intermediate	0.05	0.03
COG1902	*fadH*	COG2141	*yhbW*	no interaction	no shared intermediate	0.04	0.02
COG0017	*asnS*	COG0116	*rlmL*	no interaction	no shared intermediate	0.10	0.09
COG0090	*rplB*	COG0185	*rpsS*	in interaction	no shared intermediate	3.58	3.57
COG0655	*wrbA*	COG1733	*ytfH*	no interaction	no shared intermediate	0.04	0.03
COG0122	*alkA*	COG0350	*ogt*	no interaction	no shared intermediate	0.10	0.09
COG1752	*rssA*	COG3264	*mscM*	no interaction	no shared intermediate	0.04	0.03
COG1607	*yciA*	COG0604	*qorA*	no interaction	no shared intermediate	0.04	0.03
COG1381	*recO*	COG1159	*era*	no interaction	no shared intermediate	1.21	1.20
COG1171	*ilvA*	COG0120	*rpiA*	no interaction	no shared intermediate	0.09	0.09
COG0038	*yfeO*	COG0025	*yjcE*	no interaction	no shared intermediate	0.03	0.02
COG4108	*prfC*	COG0791	*spr*	no interaction	no shared intermediate	0.05	0.04
COG0657	*aes*	COG4221	*ydfG*	no interaction	no shared intermediate	0.03	0.02
COG1004	*ugd*	COG1210	*wcaN*	no interaction	shared int.	0.31	0.30
COG1015	*deoB*	COG0180	*trpS*	no interaction	no shared intermediate	0.16	0.16
COG1055	*arsB*	COG0598	*corA*	no interaction	no shared intermediate	0.03	0.03
COG2110	*ymdB*	COG0705	*glpG*	no interaction	no shared intermediate	0.04	0.04
COG2896	*moaA*	COG0315	*moaC*	in interaction	shared int.	0.66	0.66
COG1012	*aldB*	COG0160	*gabT*	in interaction	shared int.	0.03	0.03

Table S6. **Evolutionary modules as a function of cutoff**

Dij inter	Dij exter	Number of Modules	Number Bind (Fraction)	Number Shared Intermediate (Fraction)	Number Shared Path (Fraction)	ln(Probability) (full distribution)	ln(Probability) (coupled genes)
0.75	0.25	15	10 (.67)	1 (0.07)	0 (0.00)	-37	-3.7
0.75	0.5	35	18 (.51)	4 (0.11)	2 (0.06)	-61	-3.5
1	0.25	14	9 (.64)	1 (0.07)	0 (0.00)	-33	-2.9
1	**0.5**	**27**	**16 (0.59)**	**3 (0.11)**	**1 (0.04)**	**-58**	**-3.9**
1	0.75	39	20 (0.51)	5 (0.13)	5 (0.13)	-71	-3.4
1.25	0.25	11	8 (0.73)	1 (0.09)	0 (0.00)	-32	-3.1
1.25	0.5	20	14 (0.70)	2 (0.10)	1 (0.05)	-54	-3.7
1.25	0.75	28	17 (0.61)	3 (0.11)	3 (0.11)	-63	-3.1