Technical University of Denmark

DTU

Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations

Aziz, Ramy K.; Monk, Jonathan M.; Lewis, R. M.; Loh, Suh I.; Mishra, A.; Nagle, A. A.; Satyanarayana, C.; Dhakshinamoorthy, S.; Luche, M.; Kitchen, D. B.; Andrews, K. A.; Fong, N. L.; Li, H. J.; Palsson, B. O.; Charusanti, Pep

DTU Library
Technical Information Center of Denmark

SCIENTIFIC REP⚙RTS

Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations

Ramy K. Aziz[1,2], Jonathan M. Monk[2], Robert M. Lewis[3], Suh In Loh[4], Arti Mishra[4], Amrita Abhay Nagle[4], Chitkala Satyanarayana[4], Saravanakumar Dhakshinamoorthy[4], Michele Luche[3], Douglas B. Kitchen[3], Kathleen A. Andrews[2], Nicole L. Fong[2], Howard J. Li[2], Bernhard O. Palsson[2] & Pep Charusanti[2,5]

Mathematical models of metabolism from bacterial systems biology have proven their utility across multiple fields, for example metabolic engineering, growth phenotype simulation, and biological discovery. The usefulness of the models stems from their ability to compute a link between genotype and phenotype, but their ability to accurately simulate gene-gene interactions has not been investigated extensively. Here we assess how accurately a metabolic model for *Escherichia coli* computes one particular type of gene-gene interaction, synthetic lethality, and find that the accuracy rate is between 25% and 43%. The most common failure modes were incorrect computation of single gene essentiality and biological information that was missing from the model. Moreover, we performed virtual and biological screening against several synthetic lethal pairs to explore whether two-compound formulations could be found that inhibit the growth of Gram-negative bacteria. One set of molecules was identified that, depending on the concentrations, inhibits *E. coli* and *S. enterica* serovar Typhimurium in an additive or antagonistic manner. These findings pinpoint specific ways in which to improve the predictive ability of metabolic models, and highlight one potential application of systems biology to drug discovery and translational medicine.

Genome-scale metabolic network reconstructions form a cornerstone of microbial systems biology as they capture in one framework all known genes, proteins, and reactions within the metabolic network of an organism[1,2]. The subsequent conversion of a reconstruction into a mathematical model suitable for numerical computation allows one to compute cellular phenotypes for different growth conditions. In this way, metabolic reconstructions and their associated models constitute a computational platform that can be used to investigate the link between genotype and phenotype. This property distinguishes a metabolic model from static maps of biochemical pathways. The latter is a catalog of all known pathways in a network, whereas the former provides additional information such as growth rates in different environments; substrate uptake and product secretion rates; and the identity of pathways that are likely to be active for a given growth condition versus pathways that are present but not utilized. Such models have found use in metabolic engineering[3,4], network analysis[5,6], and have driven biological discovery[7,8].

[1]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, 11562, Egypt. [2]Department of Bioengineering, University of California, San Diego, La Jolla, California, 92093, USA. [3]Computer-Aided Drug Discovery, Albany Molecular Research, Inc., Albany, New York, 12203, USA. [4]Albany Molecular Research Singapore Research Centre, Pte Ltd, Singapore, 117525. [5]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2970 Hørsholm, Denmark. Correspondence and requests for materials should be addressed to P.C. (email: pcharusanti@ucsd.edu)

For well-characterized organisms such as *Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae*, metabolic models simulate single gene essentiality with overall accuracy of approximately 90%[9–12]. This rate drops to 49%, however, when the *S. cerevisiae* iLL672 metabolic model is used to compute synthetic lethality instead[13]. Although this rate is two orders of magnitude better than if two genes are selected by random chance[13], it still represents a drop-off when compared to the ~90% accuracy rate for single gene essentiality. The simulation accuracy of synthetically lethal (SL) gene pairs for the *E. coli* metabolic model[10] is not known since it is much more difficult to create large numbers of double deletion mutants in *E. coli* than it is in *S. cerevisiae*. Correct simulation of synthetic lethality, and more generally gene-gene interactions, therefore constitutes one area in which metabolic models can be greatly improved.

With greater biological accuracy, the *in silico* computation of synthetic lethal gene pairs could become one strategy to rationally discover combination therapeutics. Currently, some common strategies to discover drug combinations include: the inhibition of multiple key steps in pathways known to be essential, such as folate biosynthesis[14,15]; the inhibition of an essential enzyme and another protein known to confer resistance, as is the case with Augmentin[16]; and high-throughput screening of a compound library in the presence of a second compound[17,18]. The inhibition of synthetic lethal proteins has been applied to the field of cancer therapeutics[19,20], most notably the inhibition of poly(adenosine diphosphate [ADP]–ribose) polymerases (PARPs) in different cancers[21–23]. Specifically, PARPs and *BRCA1*/*BRCA2* form a synthetic lethal pair in cancer cells, but in some cases the latter has developed a loss-of-function mutation such that only one compound targeting the PARPs needed to be developed. The situation would differ for antibiotics: two compounds targeting both members of the SL pair would likely be needed. This strategy becomes more attractive if the two molecules act synergistically, since synergy can result in lower individual dosages of the two compounds and possibly more target specificity[24]. Metabolic models can significantly accelerate the search for SL gene pairs. To take *Escherichia coli* K12 MG1655 as an example, a strategy based on the inhibition of two different enzymes would yield a total search space of approximately 8 million protein pairs since this strain has 4140 annotated coding DNA sequences (CDSs). In practice, however, the search for potential targets within this space is more focused, and metabolic models can narrow the search space even further due to their ability to simulate the effect of gene knockouts on growth.

We analyze here how accurately the most current version of the *E. coli* metabolic model[10] computes synthetic lethality, and identify the most common failure modes to pinpoint specific areas in which the model can be improved. Furthermore, we perform virtual screening against a subset of validated SL pairs as a first step to find potential inhibitors, and then screen the most promising compounds both as single agents and in combination to assess the feasibility with which combinations of inhibitors can be found against the SL proteins. In the combination studies, we also assess whether the inhibitors exhibit synergy, additivity, or antagonism, and examine whether these effects match the expected outcome based on their protein targets and simulation data.

Results

Metabolic network models can guide the identification of synthetic lethal gene pairs. We used metabolic network models for *E. coli* EDL933[25], *K. pneumoniae* MGH78578[26], *Y. pestis* CO92[27], and *Salmonella enterica* serovar Typhimurium LT2[28] (hereafter referred to as *S.* Typhimurium) to guide our search for possible synthetic lethal pairs in these organisms. Other model-driven studies of synthetic lethality in *E. coli* have used metabolic models of the non-pathogenic K12 MG1655 strain[29,30]; however, we used a model for strain EDL933 to allow for potential application to infectious disease and drug discovery (Fig. 1). This strain is a prototypical one for the enterohemorrhagic O157:H7 serotype that has caused multiple outbreaks of food poisoning in the United States[31]. The strains chosen to represent the three other bacteria are likewise pathogenic variants for those organisms.

In LB media, simulations of the four models resulted in 65 (*E. coli*), 76 (*K. pneumoniae*), 78 (*Y. pestis*), and 51 (*S.* Typhimurium) computed SL pairs. Of these, three pairs are shared among all four models: *asnA*/*asnB* (amino acid metabolism), *cmk*/*pyrH* (nucleotide salvage pathway), and *tdk*/*thyA* (nucleotide salvage pathway). After these three pairs are removed, eight, five, three, and zero pairs are shared among three of the four models, respectively. These pairs fall predominantly into the following classifications: folate metabolism, amino acid metabolism (asparagine, methionine, and aromatic amino acids via the shikimate pathway), cofactor biosynthesis (pantothenate), and TCA cycle. These 19 SL gene pairs correspond to 13 unique reaction pairs, while there were six cases in which two or more SL gene pairs catalyzed the same reaction in the models. For example, *lpdA*/*sucC* and *lpdA*/*sucD* are two distinct SL gene pairs, but they correspond to only one SL reaction pair since SucC and SucD form a heterodimer. These data are summarized in Fig. 2.

In defined media, the number of computed SL pairs was 69 (*E. coli*), 70 (*K. pneumoniae*), 73 (*Y. pestis*), and 87 (*S.* Typhimurium). The defined medium for *Y. pestis* was BCS media[32], while the recipe for glucose M9 was used as the defined medium for the other three organisms. Five pairs were shared among all four models, *asnA*/*asnB* and *cmk*/*pyrH* as in LB medium plus *glyA*/*serA*, *glyA*/*serB*, and *purN*/*purT*. Unlike in LB, simulations did not identify *tdk*/*thyA* as a putative SL pair in defined media because *thyA* is singly essential under these conditions. Besides these five pairs, 21 other SL gene pairs were present in at least three of the four models. They again belong predominantly to folate metabolism, amino acid
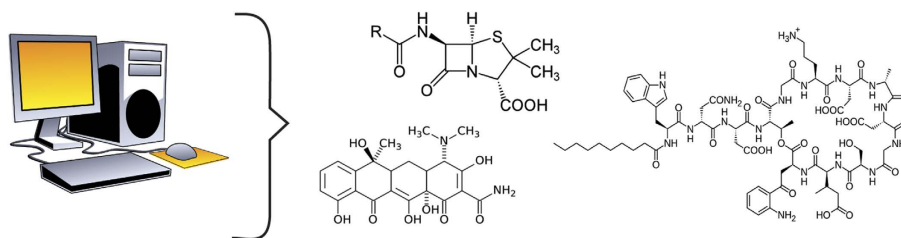
1. Build metabolic models for multiple pathogens. Compute SL pairs for each model and overlapping SL pairs shared among all models
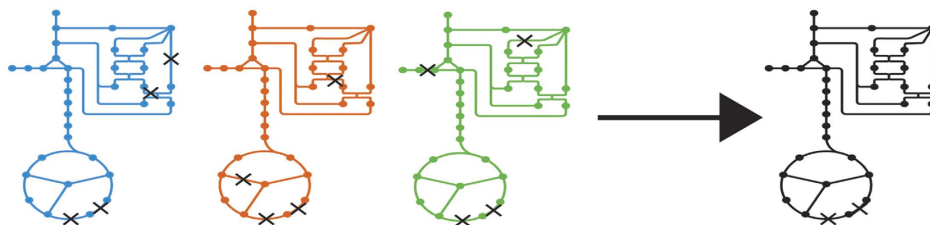
2. Construct and test double mutants to validate simulation data
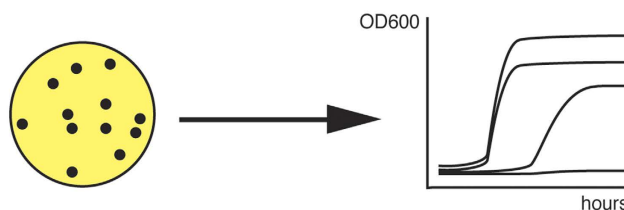
3. Perform virtual screening against validated pairs

4. Test compounds from Step 3 as single agents in whole cell assays

MIC (µg/mL)	8	4	2	1	1/2	1/4
Compound 1	○	●	●	●	●	●
Compound 2	○	○	○	○	●	●
Compound 3	○	○	○	●	●	●

5. Test compounds for synergy
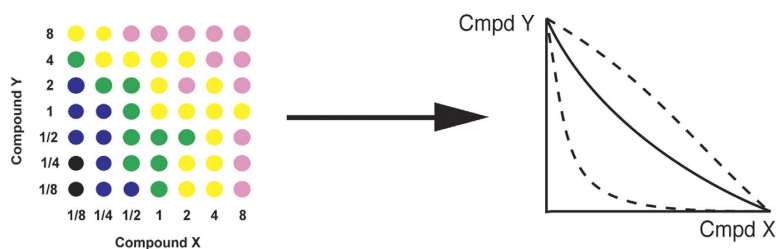
Figure 1. Five-step workflow for the model-guided identification of compounds that inhibit synthetic lethal protein pairs. The use of mathematical models in step 1 narrows the search space for SL pairs, after which the pairs are validated experimentally and then subjected to virtual and biological screening.

metabolism, and TCA cycle, with several additional pairs from ion transport and the pentose phosphate pathway. Overall, the 26 SL gene pairs correspond to 16 unique SL reaction pairs. These data are likewise summarized in Fig. 2. The full list of SL gene pairs for both LB and defined media can be found in Supplementary Dataset 1.
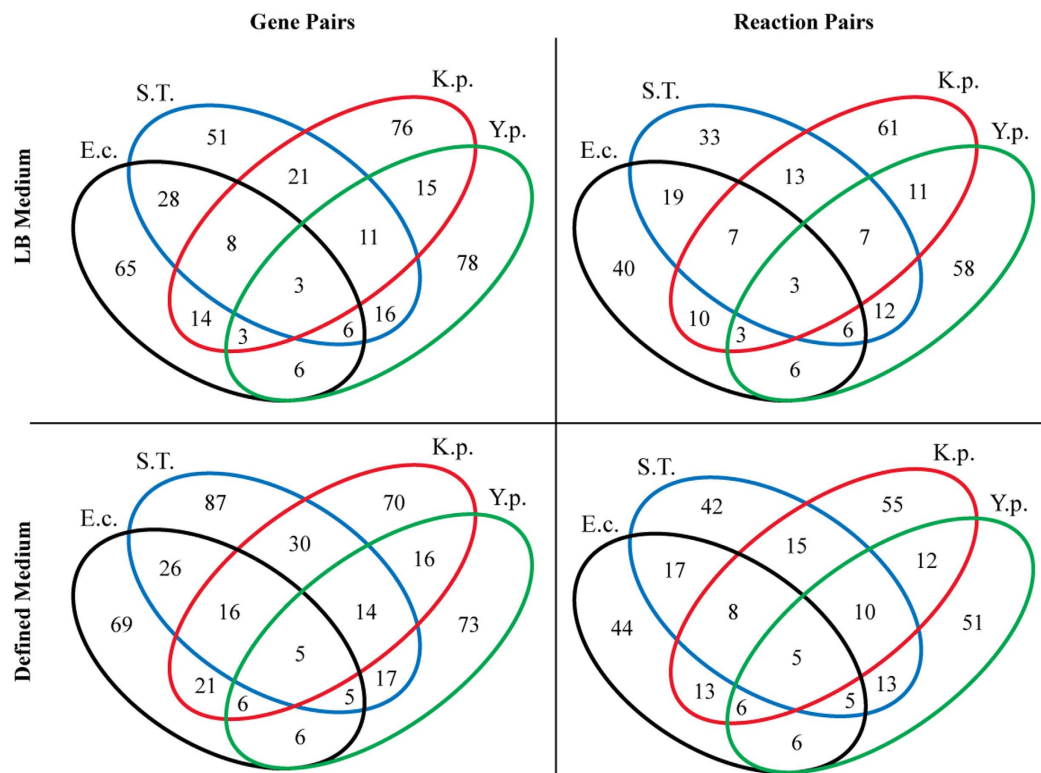
Figure 2. The number of SL genes and their corresponding reactions in each of the four individual models, and the number that were shared among multiple models. BCS media was the defined medium for the *Y. pestis* model whereas glucose M9 was the defined medium for the other three. Abbreviations: E.c.: *E. coli* EDL933; K.p.: *K. pneumoniae* MGH78578; Y.p.:*Y. pestis* CO92; S.T.: *S.* Typhimurium.

The overall accuracy of model-guided prediction of synthetic lethality was between 25% and 43%. We tested how accurately the models could predict synthetic lethality *a priori* by constructing eight double knockout mutants in *E. coli* EDL933 and four in *Salmonella enterica* strain 14028s. In *E. coli*, one of the eight double mutants was experimentally synthetically lethal (*hemF/hemN*), whereas in *Salmonella* it was two out of four (*gltB/gdhA* and *lpdA/sucC*) (Table 1). Therefore, the overall accuracy was three out of twelve (25%). Five of the false positives resulted from cases in which one of the two genes was singly essential, and therefore synthetic lethality is not possible. If these are removed from the calculation, the accuracy rate is three out of seven (43%). No proper benchmark of these results exists since a systematic, large-scale study that examines the number of SL pairs in *E. coli* has not been carried out yet; however, of 692,865 randomly chosen gene pairs in yeast that were experimentally tested, only 0.56% displayed a synthetically lethal or sick phenotype[13]. If a similar ratio holds for *E. coli*, then the models will have outperformed random selection by a statistically significant amount.

Different mechanisms lead to the lethal phenotype in the three identified SL pairs. HemF and HemN both catalyze the same essential oxidative decarboxylation step during porphyrin biosynthesis, but HemF is active under aerobic conditions while HemN is active under anaerobic conditions. Their removal consequently eliminates an essential reaction. GltB and GdhA catalyze parallel reactions within glutamate biosynthesis: the deletion of *gltB* and *gdhA* removes all possible routes from 2-oxoglutarate to L-glutamate, leading to glutamate auxotrophy. LpdA is a component of two separate complexes that catalyze different steps within the TCA cycle, and SucC is the β-subunit of succinyl-CoA synthetase, also within the TCA cycle. The deletion of these two genes severely reduces flux through this critical pathway, thereby accounting for the SL phenotype.

Incorrect computation of single gene essentiality and missing biological information were the two most common causes of inaccurate SL predictions. We analyzed the nine cases in which the predicted SL phenotype proved to be inaccurate to determine the underlying cause, since such information reveals areas for further improvement of the models. The two most common causes were situations in which one of the two genes in the pair turned out to be singly essential, and missing biological information that consequently was not represented in the models.

There were five cases in which at least one of the two genes is singly essential (Table 1), but the reasons for the discrepancies with simulation data varied. In one case, the two genes (*aroE/ydiB*) catalyze the same essential reaction in the model, but *aroE* is singly essential in LB. RNA-seq data indicate that

Model-predicted SL pair	Organism	Growth medium on which synthetic lethality was predicted to occur	Experimentally valid SL pair?	Reason for incorrect prediction
hemF/hemN	E. coli	LB	Yes	
gltB/gdhA	S. Typhimurium	M9	Yes	
lpdA/sucC	S. Typhimurium	M9	Yes	
metE/metH	E. coli	LB	No	1
ntpA/mutT	E. coli	M9	No	4
fabA*/tesB	E. coli	LB	No	3
purT/purN	E. coli	LB	No	4
ydiB/aroE*	E. coli	LB	No	3
metL*/thrA	E. coli	M9	No	3
cstC/argD	E. coli	M9	No	4
glyA*/serA*	S. Typhimurium	M9	No	3
mdh/ppc*	S. Typhimurium	M9	No	2,3
ntpA/hemF	EDL933	Negative control; double mutant was viable on LB as expected		

Table 1. Putative SL gene pairs that were experimentally constructed. Eight and four double mutants were constructed in *E. coli* EDL933 and *S.* Typhimurium 14028s, respectively, and tested for synthetic lethality in the indicated growth medium. The nine pairs that did not show synthetic lethality as predicted were classified into one of four classes as follows: *1.* Important biological information was not incorporated into the metabolic model due to inherent limitations of the model; *2.* Missing regulatory information not captured by the model; *3.* At least one of the two genes appears to be singly essential experimentally, but was not singly essential in the model. The singly essential gene is denoted with an asterisk. For the *glyA/serA* pair, both genes are singly essential; *4.* Missing biological information, e.g. the possible presence of another unidentified homolog.

Target Protein	Number of Compounds Tested	Number of Active Compounds			
		E. coli		S. Typhimurium	
		IC$_{50}$ of 0.1–200 µM	>30% inhibition at 200 µM	IC$_{50}$ of 0.1–200 µM	>30% inhibition at 200 µM
HemN	89	0	1	0	0
SucC	279	0	5	1	3
LpdA	246	0	0	0	0
GlyA	195	0	0	0	0
SerA	154	2	0	2	1
HemF	84	0	0	0	0
Mdh1	203	0	0	0	0
Ppc1	78	0	4	0	1

Table 2. Results of antimicrobial screening of 1,328 compounds against *E. coli* and *S.* Typhimurium. The table shows the number of compounds tested for each of the target proteins and their respective confirmed activity profile against the two bacteria using repurchased and reweighed samples.

ydiB is not expressed during exponential growth in LB[33,34], implying that missing regulatory information that is beyond the scope of metabolic models is responsible for this error. Likewise, missing regulatory information centered on the glyoxylate shunt underlies the error with *mdh/ppc*. In contrast, in two cases (*metL/thrA* and *fabA/tesB*) the non-essential genes (*thrA* and *tesB*) are expressed under the growth conditions expected to result in synthetic lethality. This scenario points to incorrect assignment of gene function as a likely cause. In the last case (*glyA/serA*), both genes were singly essential.

There were three cases in which missing biological information led to erroneous simulation results. Simulation of a *purT/purN* double knockout resulted in a model for *E. coli* that could not synthesize

purine, an essential metabolite; however, the double knockout was viable in LB medium (but not glucose M9), an experimental outcome that has been seen previously in other strains of *E. coli*[35]. These data suggest the presence of an unidentified purine transporter. In the model, the two proteins MutT and NtpA catalyze the same essential reaction within folate biosynthesis, forming the basis for their computed synthetic lethality. The experimental viability of this double mutant strongly suggests that at least one other unidentified protein can catalyze the same reaction. Similarly, ArgD and CstC are putative paralogs that also catalyze the same essential reaction in the model.

Validated synthetic lethal pairs can guide the search for possible combination therapeutics. We performed virtual and biological screening to identify potential inhibitors against four sets of proteins: *hemF/hemN*, *lpdA/sucC*, *glyA/serA*, and *mdh/ppc*. The first two are synthetic lethal pairs (Table 1), and compounds that inhibit them were hypothesized to act synergistically. The latter two were screened for use as controls. The *glyA/serA* pair is a case in which both genes are singly essential in *S.* Typhimurium, so they were anticipated to exhibit additivity instead of synergy. The *mdh/ppc* pair is a case in which one gene (*ppc*) is singly essential. Consequently, neither synergy nor additivity was anticipated for this pair. We followed the following general strategy to identify potential combinations of inhibitors: 1) perform virtual screening against the enzyme targets; 2) screen the compounds against *E. coli* and *S.* Typhimurium to determine their activity as single agents; and 3) screen a subset of the compounds in combination according to the SL pairing to evaluate potential synergy, additivity, and antagonism. It should be noted that, for true SL pairs, compounds that inhibit either of the two proteins are not expected to show strong activity as single agents. Instead, inhibition is expected to be strongest when both compounds are present.

We first used molecular docking to identify compounds likely to inhibit the eight proteins in either *E. coli* or *S.* Typhimurium as single agents. A diverse collection of 300,000 to 600,000 compounds was docked to binding sites within each of the eight proteins using a high throughput docking procedure (HTVS mode, see Supplementary File 1). High scoring compounds from these searches were used as similarity probes against a 70 million compound library and the hits were docked in a higher precision mode (SP mode). This step resulted in a set of 1,039 compounds.
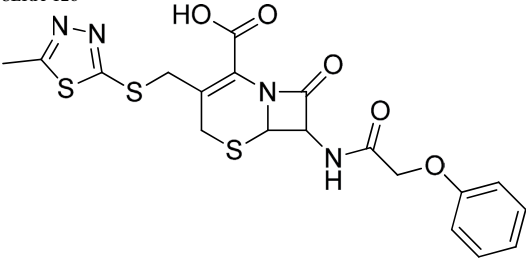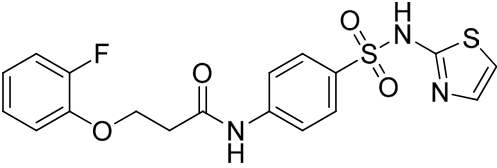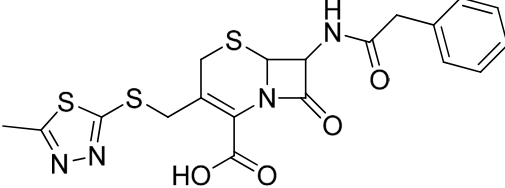
The 1,039 compounds were purchased and tested in a 10-point dose response format (top test concentration of $200\,\mu M$ and 2-fold dilution) to determine their IC_{50} and % inhibition against the two bacteria. Although virtual screening was carried out against protein targets, this step was performed as a bacterial growth inhibition assay in which the readout was cellular growth. Standard compounds were also tested along with every batch of compounds screened in the assay for quality control purposes. This step identified 67 compounds that exhibited over 20% inhibition at any dose.

These 67 weak hits were then used as probes to search for chemically similar structures that might also be active, and these compounds were likewise docked against the eight target proteins. This step resulted in a group of 289 compounds that were also purchased and screened in the cellular assays. Therefore, a total of 1328 compounds were identified by virtual screening and tested in the first round of biological screening as potential inhibitors against *E. coli* and *S.* Typhimurium (Table 2). After a hit was identified against one protein (% inhibition > 20% against either bacterium), it was subsequently docked against all eight proteins to assess whether it might bind to the other targets as well. If the docking score was at least one standard deviation above the average score for that site, then it was presumed to bind that additional site. A third qualitative assignment was made if the scores from two methods were deemed significant (see Materials and Methods). These multiple assignments are listed in Table 3 for the most active compounds.

Three of the 1328 compounds showed reproducible activity against *E. coli*, *S.* Typhimurium, or both during individual compound screening with $IC_{50} < 200\,\mu M$ and/or percent inhibition over 30% (Table 3), and we selected them for combination studies. SERA-126 inhibited the growth of both bacteria. Its IC_{50} value was lower for *S.* Typhimurium than it was for *E. coli*. SIM1-074 inhibited the growth of *S.* Typhimurium but did not have significant effect on the growth of *E. coli*. SIM4-003 inhibited the growth of both *E. coli* and *S.* Typhimurium with IC_{50} values of $<50\,\mu M$.

From the docking calculations against all eight proteins, combinations of these three compounds were presumed to target three of the four protein pairs (Table 3); however, the calculations suggested that SerA/GlyA would be the primary target. SERA-126 and SIM4-003 most likely target SerA and SIM1-074 is predicted to inhibit GlyA. This pair was not an experimentally valid SL pair because the two genes are singly essential (Table 1). Although SIM1-074 targets SucC in the calculations, combinations of these three compounds were not predicted to inhibit SucC/LpdA because SERA-126 and SIM4-003 are predicted to inhibit LpdA only weakly. The compounds exhibited weak inhibition for the HemN/HemF pair in the calculations, and it is equally likely that each compound would bind to HemN only. None of the compounds bound to Ppc in the calculations; therefore, the compounds were not predicted to have an effect on the Mdh/Ppc pair.

Two concentrations of SIM1-074 were tested against two concentrations each of SERA-126 and SIM4-003 against both *E. coli* and *S.* Typhimurium. The chemical similarity between SERA-126 and SIM4-003 suggested that these compounds would share a similar mechanism of action, and therefore they were not tested in combination with each other. The tested combinations were assigned to have either a

Compound	IC$_{50}$ (µM) (std dev)		Possible SL Targets[b]
	E. coli	*S. Typh.*	
SERA-126	88 (3)	41 (4)	HemF(+), HemN(++) SucC(−), LpdA(+) SerA(+++), GlyA(−) Mdh(+), Ppc(−)
SIM1-074	21% (2)	49% (3) ~155	HemF(+), HemN(++) SucC(+++), LpdA(−) SerA(++), GlyA(+) Mdh(−), Ppc(−)
SIM4-003	32 (1)	11 (0)	HemF(−), HemN(++) SucC(++), LpdA(+) SerA(+++), GlyA(++) Mdh(++), Ppc(−)

Table 3. IC$_{50}$ data for hit compounds in bacterial growth inhibition assays, n = 2 with standard deviations to one significant figure in parentheses. For the weak inhibitor SIM1-074, percent inhibitions at the top dose, 200 µM, are reported. One experiment provided sufficient data to calculate an IC$_{50}$ (155 µM). The possible SL targets for each compound were determined by docking the compounds against each of the eight proteins. Each "+" symbol is one standard deviation above the average of random compounds docked in the binding site. A "−" symbol implies that the compound did not dock into that site or was less than one standard deviation from the random compound average.

synergistic, additive, or antagonistic effect in the combination testing using the Loewe model[36]. Because the SerA/GlyA pair was the most likely target, we anticipated additivity as the most likely outcome.

All combinations exhibited a percent inhibition that was either similar to or lower than the calculated additive values in the combination studies against both strains (Fig. 3). Of note is that the SERA-126/SIM1-074 combination against *S.* Typhimurium resulted in lower than expected percent inhibition by 30–45% (Fig. 3). This effect was outside the experimental uncertainty (3–7%) of the difference of the measurements from the simple additive assumption, and suggests potential antagonism in this case. No synergy was observed.

Discussion

We present data here that evaluated how accurately systems biology-based metabolic models can predict SL gene pairs in four species of Enterobacteriaceae, especially *E. coli* and *S.* Typhimurium; a workflow for the discovery of combination inhibitors directed against SL pairs; and a set of compound pairs that inhibit growth of *E. coli* and *S.* Typhimurium in plate-based whole cell anti-bacterial assays in an additive manner, which was consistent with the predicted outcome for the presumed SL pair targeted by the compounds. These results highlight specific areas in which metabolic models can be improved and evaluate the ability of the models to guide the selection of anti-bacterial drug combinations.

a.

b.
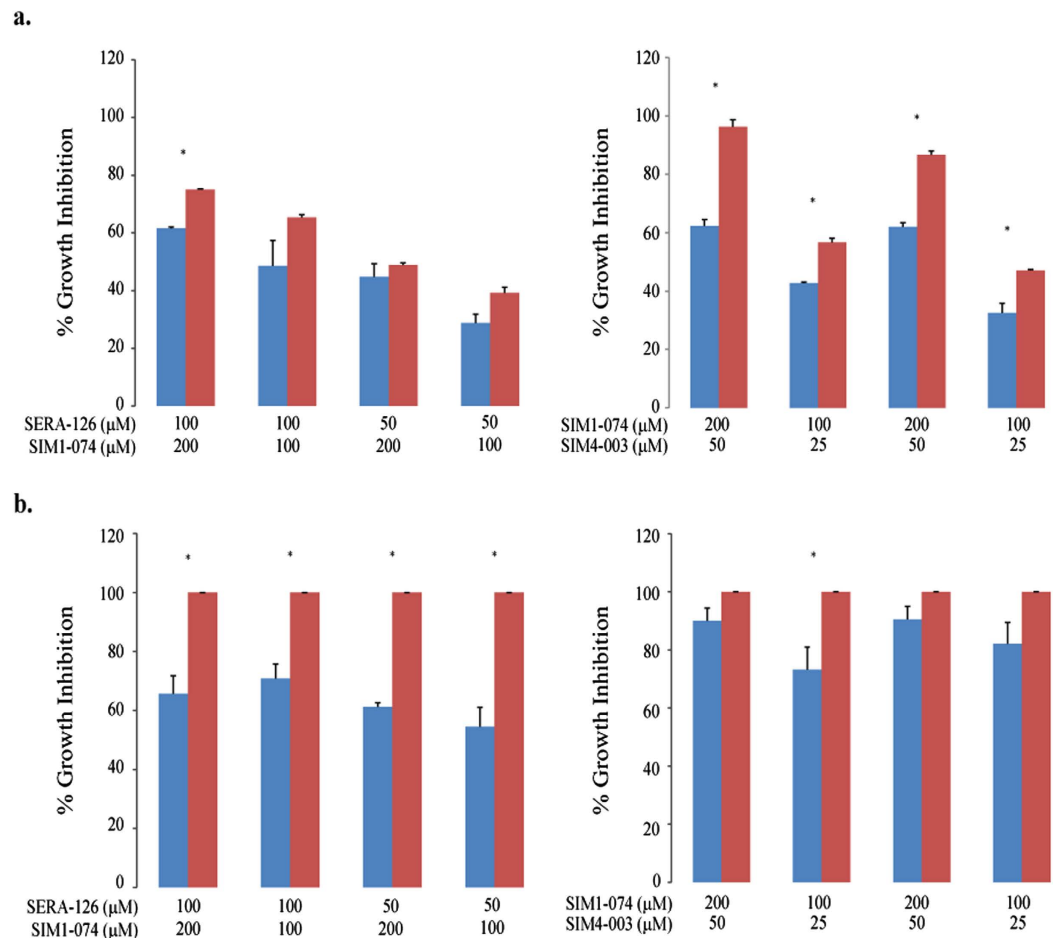
Figure 3. Percent growth inhibition in combination studies against (A) *E. coli* **and (B)** *S.* **Typhimurium.** The actual percent growth inhibitions achieved by testing two compounds in combination (blue bars) were compared with expected values based on simple additivity according to the Loewe model (red bars). The standard deviation of the single agent and combination studies ranged as high as 10% but was typically 2–5%. The calculated expected percent growth inhibition for *S.* Typhimurium was greater than 100%, but is represented here as 100%. Asterisks (*) denote statistically significant differences ($p < 0.05$) between the actual and the expected percent growth inhibition using *t*-test.

The overall accuracy rate in identifying SL pairs in this study was 25% to 43%, but the true value is likely different due to several factors. One such factor was the manner in which putative SL pairs were selected for experimental testing. The selections balanced two different goals: to be amenable to the discovery of new inhibitors and simultaneously to provide data about the ability of the models to predict synthetic lethality in an *a priori* manner. Based on these two criteria, the pairs were filtered as follows. First, pairs in which one of the two genes was known to be singly essential experimentally were removed from the list. These pairs are clear incorrect model predictions. Second, pairs in which one member was a membrane-bound protein were removed because such proteins are more difficult drug targets than their cytoplasmic counterparts. There were ten such pairs for the *E. coli* EDL933 model after simulation in M9 medium. Third, pairs already known to be synthetically lethal were removed so that this study remained focused on the discovery of new SL pairs rather than revisit known ones. Nine pairs fit this category for the *E. coli* EDL933 model after simulation in M9 medium, and we ruled out seven of them. We constructed one pair, *purT*/*purN*, because the reported double mutant had additional genetic modification besides the two deletions that might have impacted synthetic lethality. We constructed a second pair, *metE*/*metH*, to explore the limitations of the metabolic models (see below). In the end, we selected from a list of 52 remaining putative SL pairs in *E. coli* EDL933 for experimental testing, and the accuracy rate calculated here was based on this subset. A similar analysis was performed for the *S.* Typhimurium model.

Several specific cases highlight the limitations of metabolic models. For example, a cofactor not accounted for in the models plays an essential role in the reaction catalyzed by MetE/MetH. This pair was predicted by the *E. coli* EDL933 model, as well as the iJO1366 model for the K12 MG1655 strain[10], to be synthetically lethal in both LB and glucose M9. We tested this prediction by constructing the two single

mutants and the double mutant, finding that all three were viable in LB. In glucose M9, however, only Δ*metH* was viable; Δ*metE* and the double mutant Δ*metE*Δ*metH* were both non-viable. Therefore, the double mutant was not a true SL pair since *metE* is singly essential under this condition. Both enzymes catalyze the same essential reaction, but MetH absolutely requires cobalamin to function[37]. MetH is inactive in glucose M9 because cobalamin is not a component of the medium, resulting in the single essentiality phenotype of *metE* under this condition. The metabolic models used here do not explicitly account for details such as cofactors, a deficiency that led to the erroneous prediction of synthetic lethality for this pair. The latest iterations of metabolic models, however, do account for additional features such as gene expression, protein synthesis, and cofactor usage, and therefore correctly predict the requirement for cobalamin for this pair[38].

Several reasons might account for the low number of compounds identified for each SL protein derived from virtual screening, as only one pair of compounds targeting one SL pair was found. First, the docking methodology has a significant error rate even for well-explored systems, although virtual screening typically identifies compounds at a higher rate than random screening[39]. Second, there were two possible sites of competitive inhibition for many of the enzymes, a cofactor binding site and a second one for the substrate. In these cases, we docked to each site in order to identify a greater number of possible inhibitors. The modifications to the binding sites were intended to mimic reactants, products or transition states in order to fashion the binding site shape appropriately. Third, there were no crystal structures containing an inhibitor bound to the protein for any of the eight proteins, which increases the difficulty of the chemical search.

In a similar fashion, there are several reasons for the lack of activity observed in the growth inhibition assays during testing of both individual compounds and combination testing. The hit rate after testing of individual compounds was ~1%. Factors such as low compound permeability, efflux, off-target binding, and poor target engagement inside the bacterial cell likely contributed to the steep drop-off in the number of compounds that passed from the virtual to the experimental screening steps as these factors are not fully taken into account during virtual screening. Moreover, the screen was performed against wild-type cells. In this design, compounds should not display strong activity when present individually since they would inhibit only one of the two proteins in the SL pair, making it difficult to identify promising agents that might be active in combination. Future studies might approach this problem by screening against knockout mutants rather than the wild-type variant. For example, if two proteins X and Y form an SL pair, one would perform virtual screening against protein X to identify potential inhibitors, and then perform biological screening of those inhibitors using mutants containing deletions in gene Y. One would do the opposite case as well. Active compounds identified in these two screens would have a greater chance of having the expected mechanism of action, inhibition of the target protein in the SL pair, rather than inhibition arising from a different or non-specific mechanism. The compounds would then be subjected to combination studies.

Looking forward, the findings presented here pinpoint specific ways in which metabolic models can be improved, and thereby provide a guide for future experimental work. In turn, models that have greater biological accuracy would increase their utility across many different fields, and would serve as the foundation of a workflow that aims to discover combination inhibitors against SL protein pairs. This task would be difficult to implement without the use of metabolic models as a guide due to the extremely large number of possible protein pairs. Consequently, the discovery of combination therapeutics might become much more feasible as the models are improved and refined screening strategies are adopted.

Materials and Methods

Strains and media. All knockout mutants were created in *Escherichia coli* EDL933 (ATCC 700927) or *Salmonella enterica* serovar Typhimurium 14028s (ATCC 14028). Both strains are enteric pathogens. Compound screening was performed against *E. coli* Seattle 1946 (ATCC 25922), which is a Clinical and Laboratory Standards Institute (CLSI) control strain for antimicrobial susceptibility testing, and against *S.* Typhimurium 14028s.

All strains were grown in either Luria-Bertani (LB) broth/agar or glucose M9 media. The M9 medium contained either 2 g/L glucose, 100 μM $CaCl_2$, 2 mM $MgSO_4$, 6.8 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl, and 250 μL/L trace elements. The trace element solution consisted of (per liter): $FeCl_3•6H_2O$ (16.67 g), $ZnSO_4•7H_2O$ (0.18 g), $CuCl_2•2H_2O$ (0.12 g), $MnSO_4•H_2O$ (0.12 g), $CoCl_2•6H_2O$ (0.18 g) and $Na_2EDTA•2H_2O$ (22.25 g). Antibiotics were added as necessary at the following concentrations: ampicillin at 100 μg/mL, kanamycin at 50 μg/mL, and chloramphenicol at 25 μg/mL. LB powder was purchased from EMD Chemicals (Gibbstown, NJ) and used at the manufacturer's recommended concentration. All other chemicals were purchased from Fisher Scientific (Waltham, MA) or Sigma-Aldrich (St. Louis, MO).

Mathematical modeling and prediction of synthetic lethal gene pairs. Metabolic network reconstructions for *E. coli* K12 MG1655 and EDL933[10,25], *K. pneumoniae* MGH 78578[26], *Y. pestis* CO92[27], and *S.* Typhimurium LT2[28] were loaded into and simulated using the python version of the COBRA Toolbox[40]. These models can be downloaded from the BiGG database (bigg.ucsd.edu)[41]. All model reactions retained their default bounds[10]. Both glucose M9 and BCS defined media were simulated by setting

a lower bound of each exchange reaction as listed in Supplementary Dataset 1. LB media was likewise simulated by setting bounds on each exchange reaction as indicated in Supplementary Dataset 1.

Each single knockout strain was modeled by using the delete_model_gene function to constrain each reaction catalyzed by the corresponding enzyme to zero. Analogously, double gene deletions were calculated using the double_gene_deletetion_fba module and double reaction deletions were performed using the double_reaction_deletion module. Model growth phenotypes were determined using flux balance analysis (FBA) with the core biomass reaction as the objective. If a particular knockout resulted in a simulated growth rate equal to zero, that gene or set of genes was deemed to be singly essential or synthetically lethal, respectively. As an example, an IPython notebook with command line codes used to carry out this analysis for the *E. coli* K-12 MG1655 iJO1366 model is available at http://nbviewer.ipython.org/github/JonM4024/synthetic_lethals/blob/master/double_deletions_example.ipynb. The Gurobi (Gurobi Optimizer Version 5.6, Gurobi Optimization, Inc.) linear programming solver was used to perform FBA.

Construction of gene deletion mutants in *E. coli* and *S.* Typhimurium.

All gene knockouts were created using the protocol of Datsenko and Wanner[42]. Briefly, a kanamycin resistance cassette containing flanking FRT sites was generated by PCR using pKD13 as the template. The ends of the cassette comprised 60 nucleotides that contained the start or stop codon plus 57 bp that were homologous to the 57 bp immediately upstream and downstream of gene to be deleted. Correct insertion of the marker and subsequent removal from the chromosome were confirmed by PCR and Sanger sequencing. All PCR products were purified with the QIAGEN PCR clean-up kit (Valencia, CA).

Virtual and Compound Screening.

Crystal structures of six proteins were selected from the protein databank (www.rcsb.org) which were identical in sequence to the proteins found in the two bacteria. Two homology models were developed (HemF and LpdA) because no crystal structures were available. In total, thirteen sites were used for docking. Preference was given to those structures that contained a co-crystalized substrate or inhibitor so that we could select compounds that likely bind at a catalytic site. Standardized methods were used to prepare each binding site for docking, and to model the sites with compounds that confer appropriate shape and electrostatic interactions for potential inhibitory compounds. Diverse compounds from a library of 300,000 to 600,000 commercial compounds were docked using high throughput virtual screening (HTVS) precision. The best scoring compounds were subjected to atom pair similarity[43] calculations to determine chemically similar structures with a similarity cutoff of 70–80%. In addition, compounds that contained the maximum HierS scaffolds[44] of any compound on the HTVS list were also selected. The resulting combined list of compounds was docked again using standard precision. Compounds that docked well according to the Glide docking program and scored well based on a complementary scoring scheme were examined manually and purchased, if available.

Our selection choices were guided by the need for at least one or two single-agent inhibitors per protein target. Therefore, we set a goal of testing a minimum of 200 compounds per protein target based on reported success rates of selecting true active compounds in typical *in vitro* biochemical screens by docking methods. Docking methods provide a mean enrichment factor of 1–60 over random selections with an enrichment of 10 being a reasonable expectation, or approximately a 10% hit-rate[45]. We also assumed that only ~10% of compounds would permeate the Gram-negative cell walls yielding a net 1% hit rate in bacterial assays.

After identifying weak hits in the bacterial growth inhibition assay, we selected chemically similar compounds and likewise docked them to their putative binding site. The selection of similar compounds was done using atom pair similarity[43] calculations with a similarity cutoff of 70–80%. These similar compounds were then docked using the same methods as the original compounds. Any resulting compounds that met the same docking score criteria as the original compounds were purchased, if available.

The IC_{50} and percent inhibition for each purchased compound was tested in bacterial growth inhibition assays. All active compounds were re-confirmed by carefully reweighing them to measure percent inhibition and IC_{50} more precisely. Based on the percent growth inhibition data, the best concentrations for combination testing were chosen. A 2×2 matrix of concentrations was chosen such that the simple additive values would be in the 25–75% inhibition range. This choice allows synergy to be distinguished from the combined experimental uncertainty of each compound tested alone.

Tests were also conducted to assess compound turbidity in each of the studies to determine the interference (if any) from the test compounds that could affect the absorbance readings of the assay at 600 nm. None of the compounds had any interference with absorbance at 600nm.

Additional details describing the virtual and compound screening procedures can be found in Supplementary File 1.

References

1. Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. & Palsson, B. O. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology* **7,** 129–143, doi: 10.1038/nrmicro1949 (2009).
2. Palsson, B. *Systems biology: simulation of dynamic network states.* (Cambridge University Press, 2011).
3. Licona-Cassani, C., Marcellin, E., Quek, L. E., Jacob, S. & Nielsen, L. K. Reconstruction of the Saccharopolyspora erythraea genome-scale model and its use for enhancing erythromycin production. *Antonie Van Leeuwenhoek* **102,** 493–502, doi: 10.1007/s10482-012-9783-2 (2012).

4. Lee, S. J. *et al.* Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation. *Appl Environ Microbiol* **71,** 7880–7887, doi: 10.1128/AEM.71.12.7880-7887.2005 (2005).

5. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337,** 1101–1104, doi: 10.1126/science.1216861 (2012).

6. Almaas, E., Oltvai, Z. N. & Barabasi, A. L. The activity reaction core and plasticity of metabolic networks. *PLoS Comput Biol* **1,** e68, doi: 10.1371/journal.pcbi.0010068 (2005).

7. Chen, L. & Vitkup, D. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol* **7,** R17, doi: 10.1186/gb-2006-7-2-r17 (2006).

8. Reed, J. L. *et al.* Systems approach to refining genome annotation. *Proc Natl Acad Sci USA* **103,** 17480–17484, doi: 10.1073/pnas.0603364103 (2006).

9. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology* **3,** 121, doi: 10.1038/msb4100155 (2007).

10. Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Molecular systems biology* **7,** 535, doi: 10.1038/msb.2011.65 (2011).

11. Heavner, B. D., Smallbone, K., Price, N. D. & Walker, L. P. Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database (Oxford)* **2013,** bat059, doi: 10.1093/database/bat059 (2013).

12. Henry, C. S., Zinner, J. F., Cohoon, M. P. & Stevens, R. L. iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. *Genome Biol* **10,** R69, doi: 10.1186/gb-2009-10-6-r69 (2009).

13. Harrison, R., Papp, B., Pal, C., Oliver, S. G. & Delneri, D. Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci USA* **104,** 2307–2312, doi:10.1073/pnas.0607153104 (2007).

14. Bohni, E. [Comparative bacteriological investigations with the combination trimethoprim/sulfamethoxazole *in vitro* and *in vivo*]. *Chemotherapy* **14,** Suppl:1–21 (1969).

15. Bohni, E. Chemotherapeutic activity of the combination of trimethoprim and sulphamethoxazole in infections of mice. *Postgraduate medical journal* **45,** Suppl:18–21 (1969).

16. Stein, G. E. & Gurwith, M. J. Amoxicillin-potassium clavulanate, a beta-lactamase-resistant antibiotic combination. *Clinical pharmacy* **3,** 591–599 (1984).

17. Mathews Griner, L. A. *et al.* High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell-like diffuse large B-cell lymphoma cells. *Proc Natl Acad Sci USA* **111,** 2349–2354, doi: 10.1073/pnas.1311846111 (2014).

18. Ramon-Garcia, S. *et al.* Synergistic drug combinations for tuberculosis therapy identified by a novel high-throughput screen. *Antimicrobial agents and chemotherapy* **55,** 3861–3869, doi: 10.1128/AAC.00474-11 (2011).

19. Dorr, J. R. *et al.* Synthetic lethal metabolic targeting of cellular senescence in cancer therapy. *Nature* **501,** 421–425, doi: 10.1038/nature12437 (2013).

20. Kaelin, W. G. Jr. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews. Cancer* **5,** 689–698, doi: 10.1038/nrc1691 (2005).

21. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434,** 917–921, doi: 10.1038/nature03445 (2005).

22. Ashworth, A. A synthetic lethal therapeutic approach: poly(ADP) ribose polymerase inhibitors for the treatment of cancers deficient in DNA double-strand break repair. *J Clin Oncol* **26,** 3785–3790, doi: 10.1200/JCO.2008.16.0812 (2008).

23. Bryant, H. E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434,** 913–917, doi: 10.1038/nature03443 (2005).

24. Lehar, J. *et al.* Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nature biotechnology* **27,** 659–666, doi: 10.1038/nbt.1549 (2009).

25. Monk, J. M. *et al.* Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci USA* **110,** 20338–20343, doi: 10.1073/pnas.1307797110 (2013).

26. Liao, Y. C. *et al.* An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. *Journal of bacteriology* **193,** 1710–1717, doi: 10.1128/JB.01218-10 (2011).

27. Charusanti, P. *et al.* An experimentally-supported genome-scale metabolic network reconstruction for Yersinia pestis CO92. *BMC Syst Biol* **5,** 163, doi: 10.1186/1752-0509-5-163 (2011).

28. Thiele, I. *et al.* A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Syst Biol* **5,** 8, doi: 10.1186/1752-0509-5-8 (2011).

29. Suthers, P. F., Zomorrodi, A. & Maranas, C. D. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Molecular systems biology* **5,** 301, doi: 10.1038/msb.2009.56 (2009).

30. Guell, O., Sagues, F. & Serrano, M. A. Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. *PLoS Comput Biol* **10,** e1003637, doi: 10.1371/journal.pcbi.1003637 (2014).

31. Latif, H., Li, H. J., Charusanti, P., Palsson, B. O. & Aziz, R. K. A Gapless, Unambiguous Genome Sequence of the Enterohemorrhagic Escherichia coli O157:H7 Strain EDL933. *Genome Announc* **2,** doi: 10.1128/genomeA.00821-14 (2014).

32. Fowler, J. M. & Brubaker, R. R. Physiological basis of the low calcium response in Yersinia pestis. *Infection and immunity* **62,** 5234–5241 (1994).

33. Slager, J., Kjos, M., Attaiech, L. & Veening, J. W. Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. *Cell* **157,** 395–406, doi: 10.1016/j.cell.2014.01.068 (2014).

34. McClure, R. *et al.* Computational analysis of bacterial RNA-Seq data. *Nucleic acids research* **41,** e140, doi: 10.1093/nar/gkt444 (2013).

35. Nygaard, P. & Smith, J. M. Evidence for a novel glycinamide ribonucleotide transformylase in Escherichia coli. *Journal of bacteriology* **175,** 3591–3597 (1993).

36. Lehar, J. *et al.* Chemical combination effects predict connectivity in biological systems. *Molecular systems biology* **3,** 80, doi: 10.1038/msb4100116 (2007).

37. Banerjee, R. V., Johnston, N. L., Sobeski, J. K., Datta, P. & Matthews, R. G. Cloning and sequence analysis of the Escherichia coli metH gene encoding cobalamin-dependent methionine synthase and isolation of a tryptic fragment containing the cobalamin-binding domain. *The Journal of biological chemistry* **264,** 13888–13895 (1989).

38. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology* **9,** 693, doi: 10.1038/msb.2013.52 (2013).

39. McGaughey, G. *et al.* COMP 149-Comparison of topological, shape, and docking methods in virtual screening. *Abstr Pap Am Chem S* **234** (2007).

40. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* **7,** 74, doi: 10.1186/1752-0509-7-74 (2013).

41. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. O. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11,** 213, doi: 10.1186/1471-2105-11-213 (2010).

42. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci USA* **97,** 6640–6645, doi: 10.1073/pnas.120163297 (2000).
43. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom Pairs as Molecular-Features in Structure Activity Studies—Definition and Applications. *J Chem Inf Comp Sci* **25,** 64–73, doi: Doi 10.1021/Ci00046a002 (1985).
44. Wilkens, S. J., Janes, J. & Su, A. I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J Med Chem* **48,** 3182–3193, doi: 10.1021/Jm049032d (2005).
45. McGaughey, G. B. *et al.* Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model* **47,** 1504–1519, doi: Doi 10.1021/Ci700052x (2007).

Acknowledgements

Author Contributions

P.C. conceived the project. R.K.A., B.O.P. and P.C. designed the experiments. R.K.A., K.A.A., N.L.F. and H.J.L. constructed and tested the knockout mutants. J.M.M. implemented and carried out simulations involving the different metabolic models. D.B.K. designed the docking and compound selection protocols. R.M.L. performed the docking calculations and protein modeling. D.B.K., M.L., A.M., C.S. and S.D. designed the antibacterial combination experiments. S.L., A.M. and A.A.N. carried out the antibacterial and combination experiments. All authors contributed to data analysis. R.K.A., J.M.M., B.O.P., D.B.K., R.M.L., S.L., A.M., C.S., S.D. and P.C. contributed to manuscript preparation.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Aziz, R. K. *et al.* Systems biology-guided identification of synthetic lethal gene pairs and its potential use to discover antibiotic combinations. *Sci. Rep.* **5,** 16025; doi: 10.1038/srep16025 (2015).