Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets

Hal Alper¹, Kohei Miyaoku¹,² & Gregory Stephanopoulos¹

Identification of genes that affect the product accumulation phenotype of recombinant strains is an important problem in industrial strain construction and a central tenet of metabolic engineering. We have used systematic (model-based) and combinatorial (transposon-based) methods to identify gene knockout targets that increase lycopene biosynthesis in strains of *Escherichia coli*. We show that these two search strategies yield two distinct gene sets, which affect product synthesis either through an increase in precursor availability or through (largely unknown) kinetic or regulatory mechanisms, respectively. Exhaustive exploration of all possible combinations of the above gene sets yielded a unique set of 64 knockout strains spanning the metabolic landscape of systematic and combinatorial gene knockout targets. This included a global maximum strain exhibiting an 8.5-fold product increase over recombinant K12 wild type and a twofold increase over the engineered parental strain. These results were further validated in controlled culture conditions.

Optimization of metabolic phenotype often requires the simultaneous rerouting of metabolic intermediates and rewiring of regulatory networks. In prior work, this optimization has been accomplished by the modification of genes with well-defined structural or regulatory roles in the context of the particular metabolic pathway being considered¹⁻³. Distant genes affecting a metabolic phenotype either through redistribution of metabolite precursors or indirect kinetic and global regulatory effects have been particularly challenging to identify. Models are relatively ineffective in the search for such genes because of their inability to capture the genes’ complex, nonlinear kinetic and regulatory interactions. In general, methods for identifying genetic targets are not as powerful as the molecular biological tools that are effectively used to modify such targets. These issues become more involved when one considers the possibility of multiple gene modulations⁴. In general, the complex nature of the metabolic landscape raises significant challenges in the development of an optimal search strategy because varying genetic backgrounds and culturing conditions have a profound impact on the type of gene targets identified by various strategies.

Recently, we reported on a method for the rational design of strains that identifies single and multiple gene knockout targets based on a global stoichiometric analysis. The method was applied successfully to increase lycopene production in recombinant strains of *Escherichia coli*. Lycopene production was investigated in the context of the nonmevalonate pathway in which cells are recombinant, expressing the *crtEBI* operon to encode for the polymerization into the 40-carbon molecule product. The pre-engineered strain used for the study contained chromosomal overexpressions of *dxs*, *idi* and *ispFD* (Fig. 1a). There has been a significant effort to specifically engineer the isoprenoid pathway and downstream genes⁵⁻¹⁰; however, in the previous study² and this current one, we investigate genome-wide gene knockout targets. A total of seven single and multiple stoichiometric gene deletions, (*AgdhA*, *ΔaceE*, *ΔyjiD* (gpmB), *ΔfdhF*, *ΔgdhA*, *ΔaceE*, *ΔgdhA*, *ΔaceE* *ΔfdhF*), were predicted and experimentally validated to increase lycopene production through increasing the supply of precursors and cofactors that are important in the lycopene pathway⁵. These seven mutations along with the parental strain comprise the set of eight systematically designed genotypes. The left panel of Figure 1b depicts the methodology for identifying these systematic gene knockout targets.

Lycopene production in these systematically identified knockout strains was still below the stoichiometric maximum, presumably limited by unknown kinetic or regulatory factors that are unaccounted for in stoichiometric models. To identify additional knockout targets that affect the lycopene phenotype via regulatory, kinetic or other unknown mechanisms, we undertook a global transposon library search in the background of the pre-engineered parental strain. Screening this transposon library on glucose plates identified three gene targets that correlated with lycopene overproduction. Upon sequencing, these combinatorial targets were identified as *rssB* (also known as *hmr*), *yjfP* and *yjiD*. In the case of *yjiD*, the transposon was found to be inserted between the identified promoter region and the gene for *yjiD* and will henceforth be referred to as *ΔyjiD*. The right panel of Figure 1b shows the identity and annotated function of these selected gene targets along a representative location of the transposon insertion event. We note that none of the previously identified single stoichiometric genes surfaced in the combinatorial transposon search.

---

¹Department of Chemical Engineering, Massachusetts Institute of Technology, Room 56-469, Cambridge, Massachusetts 02139, USA. ²On leave from Mitsubishi Chemical Corporation. Correspondence should be addressed to G.S. (gregstep@mit.edu).

Published online 10 April 2005; doi:10.1038/nbt1083
because of the relatively high threshold of the lycopene accumulation level imposed in the selection of candidate strains. Using these three identified targets, it is possible to create a total of seven gene combinations of single, double and triple combinatorial target mutations (ΔrssB, ΔyjfP, ΔyiiD, ΔrssB ΔyjfP, ΔrssB ΔyiiD, ΔyjfP ΔyiiD and ΔrssB ΔyjfP ΔyiiD). These seven combinations along with the parental strain constitute the combinatorial strain set comprising a total of eight strains.

The previous results point to two distinct sets of stoichiometric and combinatorial gene targets. It is not clear how these targets interact when combined. To answer this question, we conducted an exhaustive study of the 64 strains comprising all combinations of the eight stoichiometric and eight combinatorial genotypes. These target genes were modified in the background of the pre-engineered recombinant E. coli strain. The resulting production profiles over the course of a 48-hour shake-flask fermentation process provided the information needed for the complete mapping of the lycopene metabolic landscape (Fig. 2).

Several interesting observations arise from the topology of this metabolic landscape. First, two global maxima exist, each with production levels around 11,000 p.p.m. (µg/g dry cell weight). The first strain contains the ΔgdhA ΔaceE ΔfthE genotype, which is a purely stoichiometrically designed strain. The other maximum is ΔgdhA ΔaceE ΔyiiD, which is created through the combination of stoichiometric and combinatorial targets. Second, several local maximum points are present with production levels ranging from 8,400 to 9,400 p.p.m., each formed from the combination of systematic and combinatorial targets. Third, the left quadrant of the graph indicates that the combination or stacking of more than one combinatorial knockout target greatly reduces lycopene levels to below 2,000 p.p.m., and as low as only 500 p.p.m. for some constructs, which is below the production level of a recombinant wild-type E. coli K12 strain. Finally, visual inspection of this landscape suggests a highly nonlinear function with many local optima.

Clustering methods have been routinely applied to the analysis of microarray (and other) data to determine sets of genes that exhibit similar expression profiles14. Likewise, the technique of hierarchical clustering may be applied to the metabolic landscape of Figure 2 to cluster gene knockout constructs exhibiting similar production profiles over the four time points. Presumably, strains clustering most
closely accumulate product by following similar modes of action in the mechanism of lycopene production. Upon clustering the entire set of 64 strains, two distinct organizations emerge for the two sets of gene targets previously identified.

Clustering lycopene profiles (across the four time points) for the eight stoichiometric knockout strains revealed a fairly close, stacked dendrogram (see abscissa of Fig. 3a). When these strains are plotted against the lycopene accumulation level, they reveal an expanding concentric bubble-plot suggesting an additive effect of accumulating gene deletions. This is in concert with the presumed mode of action in these strains, namely the increasing availability of precursors and cofactors that are needed for lycopene biosynthesis.

In contrast to Figure 3a, all combinatorial targets, as exemplified by rssB, force a split-tree shape in the dendrogram when performed in the background of each of the seven stoichiometric targets (Fig. 3b). Different time courses in lycopene accumulation suggest different modes of action for the effect of the combinatorial genes on this phenotype. Specifically, whereas each construct formed from the deletion of a single combinatorial target gene tends to exhibit similar behavior (increased production), the deletion of combinations of these genes yields phenotypes that are neither linear nor synergistic. In fact, double and triple knockout constructs arising from these combinatorial targets exhibit vastly different production profiles from the individual targets (Fig. 2). This nonlinearity suggests that the combinatorial targets are disrupting regulatory processes that are relatively incompatable, and in certain cases deleterious, when combined.

Biological differences are observed when combinatorial genes are deleted together with stoichiometric ones. Strains in cluster Y (Fig. 3b) all exhibit an extended lag phase, which extends to 16–18 h before reaching a typical cell density OD 3.5–4.0. In contrast, strains in cluster Z do not posses such a lag phase and exhibit a steady increase of lycopene production with time. The average, scaled
production profiles for the purely systematic cluster and the two clusters forced by an rssB deletion are compared in Figure 3c. It is noted that this branched pattern is exhibited by all strains constructed from the deletion of any combinatorial gene in the background of the stoichiometric targets, with different production profiles characterizing each of the clusters.

Drawing from this analysis, it appears that stacking (that is, deleting) combinatorial target genes upon stoichiometric ones leads to a decoupling of the stoichiometric logic. This decoupling is evident in analyzing the impact of the deletion of rssB or any other combinatorial gene, on the shape of the dendrogram obtained from hierarchical clustering of the lycopene accumulation profiles for the eight stoichiometric strains; it is also quantified by covariance analysis (Supplementary Fig. 1 and Supplementary Discussion).

The exhaustive exploration of the combinations of stoichiometric and combinatorial targets allowed the identification of several interesting strains on the basis of their performance in batch shake-flask cultivations. To better assess the production capacity of these knock-out strains, fed-batch cultivations were carried out in shake-flasks and controlled bioreactors with staged glucose feed (Fig. 4). Several strains were thus evaluated. Optimized shake-flask fermentations highlight the capability of the global maximum strains to produce upwards of 18,000 p.p.m. in 24–40 h (Fig. 4). These global maximum strains were also grown in 500-ml bioreactors with a similar glucose feeding profile and pH control and showed enhanced lycopene production producing upwards of 23,000 p.p.m. in only 60 h (data not shown). Further improvements are possible through iterative bioreactor optimization. The good correspondence between fermentor and shaker-flasks results from the deletion of any combinatorial gene in the background of the knock-out strains. Yet, it proved invaluable in the creation of some important strains in the landscape. Third, the presence of many local maxima complicates the nature of the landscape and raises questions about general sequential search strategies. Previously, sequential search strategies were found to be quite effective when applied to the space of stoichiometric genes, which is due to their overall additive effect on phenotype. Figure 2 suggests that this result does not hold when combinatorial genes are also included in the search space, necessitating exhaustive combinatorial searches of the type undertaken in this study. Although identification of optimal gene targets will continue to be a demanding undertaking, searches for gene targets will be significantly aided by advanced models of cell function accounting for kinetic and regulatory mechanisms.

It should be noted that the search of this study was limited to the overall cellular phenotype whereas the effect of regulatory targets is definitely nonlinear and more complex. This work serves as a case study aiming to understand the complex interaction of the genotype-phenotype space in the context of product overproduction phenotype.

This study underscores some important issues optimizing phenotype. First, high-throughput screening methods combined with detailed cellular models will aid in efficient strain optimization. Second, combinatorial targets influencing global cellular function should be invoked at later stages in the strain improvement process to avoid selecting those with limited utility or incompatible modes of action. Finally, metabolic genes seem to have a linear impact in the overall cellular phenotype whereas the effect of regulatory targets is definitely nonlinear and more complex. This work serves as a case study aiming to understand the complex interaction of the genotype-phenotype space in the context of product overproduction phenotype. The lessons gained from the exhaustive exploration of systematic and combinatorial gene knockout sets can help shape future strain improvement programs as they are tested in diverse systems for divergent products.

METHODS

Strains and media. E. coli K12 PTS-dxs, PTS-idi, PTS-ispFD, provided by DuPont, was used as the lycopene expression strain when harboring the PACLYC plasmid containing the crEBI operon. Overexpression of dxs, idi, and...
**LETTERS**

PCR product recombination using the pKD46 plasmid expressing the lambda Knockout construction and verification.

Consequently constructed by using PCR product recombination and tested for comparison. Strains identified through transposon mutagenesis were subsequently constructed by using PCR product recombination and tested for maintenance of the lycopene overproduction phenotype.

Gene target identity was determined through BLAST nucleotide sequence and the product of the TAIL3 reaction was sequenced using the primer TAIL-seq, 5'-CGTTGGCTACCCG-3' used as described in the reference. The product of the TAIL3 reaction was purified by a PCR cleanup kit (Qiagen) after gel visualization. This product was sequenced using the primer TAIL-seq, 5'-CATGCGCTTCTATCGCCTTCTT-3' as described in the reference. The product of the TAIL3 reaction was sequenced.

**Knockout construction and verification.** Gene deletions were conducted using PCR product recombination using the pKD46 plasmid expressing the lambda red recombination system and pKD13 as the template for PCR (see Supplementary Table 1 online for primer designs). Gene knockouts were verified through colony PCR. Phage transduction was used for creating multiple gene knockout strains. P1vir phage transduction was used to transfer knockout mutants between strains20. PCR primers used for knockout and verification may be found in Supplementary Table 1 online.

**Lycopene assay.** Intracellular lycopene content was extracted from 1 ml of bacterial culture at the point of total glucose exhaustion. The cell pellet was washed, and then extracted in 1 ml of acetone at 55 °C for 15 min with intermittent vortexing. The lycopene content in the supernatant was quantified through absorbance at 475 nm12 and concentrations were calculated through a standard curve. The entire extraction process was performed in reduced light conditions to prevent photobleaching and degradation. Cell mass was calculated by correlating dry cell with OD600 for use in p.p.m. (mg lycopene/g dry cell weight) calculations.

**Hierarchical clustering routines.** A complete linkage hierarchical clustering of the lycopene time profiles for the entire 8 × 8 strain matrix (containing values of the maximum lycopene production) using the Euclidean distance as the similarity metric was performed using Cluster Version 3.0. Dendrograms were visualized using Java TreeView Version 1.0.8.

**Acknowledgments**

We acknowledge financial support of this work by the DuPont-MIT Alliance. In particular, we would like to thank Wonchul Suh for providing the parental E. coli strain. We also thank Joel Moxley for providing thoughtful suggestions and Veronica Godoy for providing the initial phase stock.

**Competing interests statement**

The authors declare that they have no competing financial interests.

*Received 10 February; accepted 2 March 2005 Published online at http://www.nature.com/naturebiotechnology/


*Note: Supplementary information is available on the Nature Biotechnology website.*