Title

Characterizing Strain Variation in Engineered E. coli Using a Multi-Omics-Based Workflow

Journal

Cell Systems, 2(5)

ISSN

2405-4712

Authors

Brunk, Elizabeth
George, Kevin W
Alonso-Gutierrez, Jorge
et al.

Publication Date

2016-05-01

DOI

10.1016/j.cels.2016.04.004

Peer reviewed

Characterizing strain variation in engineered *E. coli* using a multi-omics based workflow

Elizabeth Brunk[a,b,c,†], Kevin W. George[a,c,d,†], Jorge Alonso-Gutierrez[a,c], Mitchell Thompson[a,e], Edward Baidoo[a,c], George Wang[a,c], Christopher J. Petzold[a,c] Douglas McCloskey[b], Jonathan Monk[b], Laurence Yang[b], Edward J. O'Brien[b], Tanveer S. Batth[a], Hector Garcia Martin[a,c], Adam Feist[b,c], Paul D. Adams[a,g], Jay D. Keasling[a,c,f,h,i], Bernhard O. Palsson[b,f,*], Taek Soon Lee[a,c,*]

[a] Joint Bioenergy Institute (JBEI), 5885 Hollis Street, Emeryville, CA 94608, USA

[b] Department of Bioengineering, University of California San Diego CA 92093, USA

[c] Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[d] Current address: Amyris, 5885 Hollis Street, Emeryville CA 94608, USA

[e] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

[f] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Horsholm, Denmark

[g] Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[h] Department of Chemical & Biomolecular Engineering, University of California, Berkeley, CA 94720, USA

[i] Department of Bioengineering, University of California, Berkeley, CA 94720, USA

†Authors contributed equally

*Correspondence: T.S.L (tslee@lbl.gov) and B.O.P (palsson@eng.ucsd.edu)

SUMMARY

Understanding complex metabolic interactions that occur between heterologous and native biochemical pathways represents a major challenge in metabolic engineering and synthetic biology. We present a workflow that integrates metabolomics, proteomics, and genome-scale models of *Escherichia coli* metabolism to study the effects of introducing a heterologous pathway into a microbial host. This workflow incorporates complementary approaches from computational systems biology, metabolic engineering, and synthetic biology, provides molecular insight into how the host organism microenvironment changes as a result of pathway engineering, and demonstrates how biological mechanisms underlying strain variation can be exploited as an engineering strategy to increase product yield. As a proof-of-concept, we present the analysis of eight engineered strains producing three biofuels: isopentenol, limonene, and bisabolene. Application of this workflow identified the roles of candidate genes, pathways, and biochemical reactions in observed experimental phenomena, and facilitated the construction of a mutant strain with improved isopentenol productivity. The contributed workflow is available as an open-source tool, in the format of three iPython notebooks.

INTRODUCTION

The confluence of high-throughput omics technologies and computational advances has dramatically changed our ability to probe biological phenomena across a vast range of chemical and biological scales (Berger et al., 2013; de Jong et al., 2012; Kuehnbaum and Britz-McKibbin, 2013; Tyo et al., 2007). Large-scale improvements in data coverage and measurement fidelity enable the quantitative tracking of dynamic changes in RNA transcripts, ribosome profiling, proteins, and metabolites in unprecedented detail (Aebersold and Mann, 2003; de Godoy et al., 2008; Fuhrer and Zamboni, 2015; Gross, 2011; Kahn, 2011; Metzker, 2010; Zhang et al., 2014). Yet, current computational tools for handling such data are rapidly becoming inadequate when compared to the amount of omic data that can now be generated (Stephens et al., 2015). This challenge, referred to as *Big Data to Knowledge* (Margolis et al., 2014), requires balancing the deluge of experimental "big data" with a solid, theoretical basis for its interpretation.

Some of the major impediments to realizing the potential impact of big data resources include: a lack of appropriate *in silico* tools, poor data accessibility, and insufficient cross-disciplinary training (Berger et al., 2013). Current computational methods are limited in their capabilities to accommodate the increasingly diverse range of experimental techniques and to contextualize new data within existing data sets (Berger et al., 2013). To make matters worse, the skillsets required of scientists in the era of big data now extend outside the traditional scope of biochemistry and molecular biology to include bioinformatics, biostatistics and computer science. Hence, despite the interest to collaborate or use tools from an orthogonal field of research, domain-specific jargon is yet another obstacle to overcome by the prospective practitioner in big data science (Rolfsson and Palsson, 2015).

In this work, we hope to lower the barrier of entry into computational systems biology by creating a framework upon which disparate biological data types can be analyzed and interpreted. We take advantage of three synergistic, accelerating domains of science- systems biology, metabolic engineering and synthetic biology- to develop a workflow that reconciles systems-level, multi-omics analysis and genome-scale modeling with synthetic pathway engineering. While the generation of large-scale omic data has already enabled numerous metabolic engineering efforts (Alonso-Gutierrez et al., 2015; George et al., 2014; Han et al., 2003, 2001; Kabir and Shimizu, 2003; Landels et al., 2015; Lee et al., 2005), the high-dimensionality of multi-omics data makes systematically extracting biologically meaningful information for a single strain, let alone a multi-strain comparison, a significant challenge (Kwok, 2010; Nielsen et al., 2014; Palsson and Zengler, 2010). In most cases, engineering strategies, such as the design–build–test–analyze (DBTA) cycle, (Bailey, 1991) are based on relatively few experimental outputs (e.g., product titer), which severely limits the "analyze" phase of the optimization cycle. This motivates the development of tools to better characterize the biological components of these complex systems, decrease the heavy reliance on iterative trial-and-error, and, ultimately, bring biological engineering closer to other, more rational, engineering disciplines.

To address this multi-layered challenge, our hierarchical workflow consists of three stages (Figure 1). In the first stage, basic strain differences are assessed through a global analysis of computationally -derived "dynamic difference profiles". The second stage uses multivariate analysis to identify relevant patterns and correlations in key metabolites and proteins. In the last stage, these inputs are reconciled with genome-scale models to identify perturbed metabolic nodes that are subsequently validated and investigated as engineering leads. We apply this framework to eight engineered strains of *E. coli* producing three isoprenoid-derived advanced

biofuels, and demonstrate that this strategy is capable of clarifying convoluted metabolic network responses, identifying potential bottlenecks, and elucidating the complex interplay between synthetic and endogenous *E. coli* metabolism.

RESULTS AND DISCUSSION

Pathway description, strain selection, and multi-omics data generation

In synthetic biology, the design of efficient "cell factories" typically involves the introduction of heterologous genes and metabolic pathways into a microbial host. In the last decade, broad classes of chemicals including isoprenoids, polyketides, branched chain alcohols, and fatty acids have been successfully produced using a variety of microbial hosts and renewable, bio-based materials (Janßen and Steinbüchel, 2014; Jullesson et al., 2015; Jung et al., 2010; Peralta-Yahya et al., 2012; Ro et al., 2006; Runguphan and Keasling, 2014; Sarria et al., 2014; Shen and Liao, 2013; Steen et al., 2010; Trinh et al., 2011; Yim et al., 2011). The native mevalonate pathway from *Saccharomyces cerevisiae*, which consists of six reactions that convert acetyl-CoA into the isoprenoid precursor isopentenyl diphosphate (ipdp or IPP), has been heterologously expressed in *E. coli (Martin et al., 2003)* and adapted to produce a variety of terpene fuels and chemicals (George et al., 2015a). By expressing additional genes, this core pathway has been modified to produce C_5 (hemiterpene) isopentenol (Chou and Keasling, 2012), C_{10} (monoterpene) limonene (Alonso-Gutierrez et al., 2013), and C_{15} (sesquiterpene) bisabolene (Peralta-Yahya et al., 2011), terpenes that serve as drop-in replacements for gasoline, jet fuel, and diesel, respectively.

Optimization of each of these heterologous pathways has yielded strains with significantly improved titers through methods such as codon optimization of poorly-expressed genes, promoter supplementation, altered operon order, and changes in plasmid copy number (Alonso-Gutierrez et al., 2015; George et al., 2015b, 2014; Peralta-Yahya et al., 2011). Though the titer of

each fuel target has consistently improved, the impact of these optimizations on endogenous *E. coli* metabolism has yet to be comprehensively explored (Supplementary Figure S1). Given that previous strain optimization has focused primarily on the mevalonate pathway itself, we suspected that a systematic exploration of the interplay between heterologous pathway engineering and endogenous metabolism could better characterize strain variation, identify perturbed metabolic nodes, and ultimately yield new engineering targets. To explore this issue, we selected representative strains for each biofuel (Figure 2(a)) with different levels of optimization (Figure 2 (b)) and collected extensive omics data (Figure 2(c)) for both heterologous and endogenous metabolism in a fermentation time-course.

Our analysis included three isopentenol-producing strains (I1-I3), three limonene-producing strains (L1-L3), two bisabolene-producing strains (B1-B2), and wild-type *E. coli* DH1 (WT) (9 strains total; Supplementary Figure S2 and Table S1). The numbering of the strains in each set represents their overall performance (product yield) and evolution of the optimization process (i.e., "1" represents non-optimized pathway and "2" or "3" represents variants with better performance). Samples were collected to measure cell growth, product titer, intracellular and extracellular metabolites, and selected proteins at multiple time-points (0 to 72 hours post-induction) in the batch fermentation. Altogether, our analysis included the absolute quantification of more than 80 metabolites and the relative quantification (via a targeted SRM method (Picotti and Aebersold, 2012)) of more than 50 proteins or protein complexes spanning key "nodes" in heterologous (i.e., mevalonate pathway) and endogenous metabolism (Supplemental Tables "Heterologous Proteomic Data Analysis.xls" and "Heterologous Metabolomics Data Analysis.xls").

Our goal over the next three stages of the contributed workflow is to elucidate the effect that optimization has on host metabolism through a combined data-driven and hypothesis-driven approach, in which the generation of multi-omics data is complemented with both statistical and bottoms-up, metabolic modeling methods.

Stage one: Integrating multi-omics data and profiling batch fermentation dynamics

Stage one of the workflow (Figure 3) integrates raw, multi-dimensional omics data to identify basic differences between strains. First, we assign test (e.g., engineered strain) and control (e.g., WT) conditions, which can vary depending upon the question being addressed. We take the difference of measured metabolite concentrations in the test and control conditions at each time point to "bin" the pairwise differences into one of six "dynamic difference profiles." These profiles describe the behavior of the test condition relative to the control (e.g., "no change", "constant", "deviation", "return", "shift", and "transient"; Figure 3), and systematically identify global trends to characterize broad changes between any two strains. With this framework in place, thousands of omics inputs can be rapidly "filtered" into distinct profiles to facilitate large-scale strain comparisons and statistical analysis.

Using the above schema, we generated dynamic difference profiles for the 8 biofuel-producing strains to examine which metabolites were among the most perturbed nodes with maximal changes relative to WT. Strains I1, L1, and B1 consistently secrete acetate at similar levels to WT (e.g. "no change" or "constant" profiles; Supplementary Figure S3), whereas strains I2, L2 and I3 strongly deviate (concentrations 14, 15 and 18 fold lower than WT). Dynamic difference profiles also highlight changes in less-abundant, intracellular ('_c') metabolites, where differences between strains are often more subtle. Certain strains show large-scale "transient" changes in the intracellular concentrations of citrate (cit_c), alpha-ketoglutarate (akg_c),

158 glycolate (glyclt_c) and amino acids, such as glutamate and lysine, which are most dramatic for
159 isopentenol producers - the strains that produce the most biofuel (Figure 2(b)).

160 Our findings generally point to a global pattern: the profiles of low-producing strains tend
161 to "cluster" with WT rather than high-producing strains of the same fuel target. Despite the
162 introduction of different heterologous pathways, the metabolite profiles of the poorly-optimized
163 strains (i.e., I1, L1, and B1) show minimal deviations from WT. Similar to WT, these strains do
164 not consume all available glucose and the concentrations of intracellular central carbon
165 metabolites, like succinate (succ_c) and phosphoenolpyruvate (pep_c), match WT levels (e.g.
166 "no change" profile). In contrast, profiles of top producing strains show large-scale deviations
167 from WT, especially for citric acid cycle (TCA) metabolites (strains I2, I3, and L3 with 16-30
168 fold changes in concentration of succ_c and between 5-13 fold changes in concentration of pep_c
169 for strains I2, I3, L2 and B2; Supplementary Figure S3). These findings suggest that the level of
170 pathway optimization, rather than the identity of the target biofuel, tends to dictate the
171 endogenous metabolic response. While this is not entirely unexpected given the common
172 mevalonate pathway "backbone" of each strain (Figure 2(a)), it suggests that the role of
173 potentially confounding factors such as biofuel toxicity (Dunlop et al., 2011) or FPP (frdp_c)
174 feedback inhibition (Primak et al., 2011), which vary markedly for each fuel target or pathway, is
175 minimal in these strains compared to impact of overall product titer.

176 In summary, the first stage of this workflow provides a rapid means to filter complex omics
177 data into categorical "dynamic difference profiles" and facilitate strain comparisons. The main
178 understanding gained from this stage of the workflow is that, for the current group of strains,
179 optimization level (i.e., overall product yield), rather than chosen fuel target, dictates degree of
180 metabolic perturbation. While this analysis also provides valuable insight into the broad

metabolic response to engineering by highlighting maximally perturbed nodes, additional analyses are needed to: (1) explore how these changes are correlated over time (stage two) and (2) contextualize these perturbations within a biochemical network (stage three).

Stage two: correlations in key metabolic fingerprints distinguish strain behavior

Despite the high dimensionality of multi-omics datasets, unsupervised learning methods, like PCA, capture much of the variation in a few key metabolites. In stage two of the workflow, we use standard multivariate analyses to reduce the dimensionality of multi-strain metabolomics data and identify common patterns in changing metabolite concentrations over time (Figure 4, steps (1) and (2)). Specifically, we carry out Principal Component Analysis (PCA) on the aggregate metabolomics data set (9 strains, 13 time points, and 86 metabolites) to identify key metabolites that drive strain variation, determine how these drivers change over time, and uncover unique features for strain characterization.

Using PCA on this data set shows that the first, second, and third singular vectors account for more than 80% of the variance in the dataset (Supplementary Figure S4). For the top-producing strains, coefficients (factor loadings) for the fuel products tend to be the most significant, coinciding with the increased production yields for these strains. Not surprisingly, certain extracellular metabolites, including lactate, pyruvate, formate, and acetate, also have higher coefficients. Performing PCA on extracellular versus intracellular metabolites, we find that the first two eigenvectors sum to more than 60% and 70%, respectively, indicating that (i) changes in intracellular concentrations are correlated over time and (ii) the uptake and secretion of extracellular metabolites are also correlated processes.

Plotting the first two singular vectors of PCA on the exometabolome shows a distinct three-state behavior in all 9 strains. We find that these three phases correspond to distinct time

intervals in the data set: (i) phase I (0-6 hours); (ii) phase II (6-20 hours); and (iii) phase III (20-72 hours), as illustrated in a simplified depiction in Figure 4. The variation in each phase is driven by changes in extracellular metabolites, such as, in the case of WT, glucose in phase I, lactic acid, formate, and pyruvate in phase II, and acetate in phase III, which is consistent with what is commonly observed in exponential, early stationary, and late stationary growth phases of *E. coli*. These same metabolites show completely different behavior in top-producing strains (e.g. acetate becomes a driver of phase II in strains I3, L3, and B2 and formate and lactic acid drive phase III; Supplementary Figure S5). The shift in acetate is interesting because its assimilation, or uptake, following its secretion is a key differentiator between optimized strains and non-optimized derivatives. By assimilating acetate in phase II, optimized strains such as I3 can recapture "lost" carbon and reform acetyl-CoA through the action of acetyl-CoA synthetase (Wolfe, 2005).

Intriguingly, changes in key intracellular metabolites also appear to coincide with this three phase behavior. As expected, amino acids are the main drivers of variation during the first phase. In the second phase, variation in low-producing strains is driven by glycolate (glyclt_c), glyoxylate (glx_c), and isocitrate (icit_c), which is consistent with glyoxylate metabolism and wild-type behavior. In top-producing strains, however, phosphoenolpyruvate (pep_c), citrate (cit_c), and α-ketoglutarate (akg_c) become the main drivers of phase II, which suggests the metabolic use of *other* TCA cycle reactions in these strains and corroborates the respective dynamic difference profiles from stage two.

To summarize, stage two of our workflow provides a means for correlating changes in metabolite concentrations over time. Using PCA, we identified three phases in time-course metabolomics data that are driven by the uptake and secretion of key metabolites, in addition to

specific intracellular metabolite changes. The identification of these three metabolic phases motivates a more in-depth characterization of each of these states by genome scale-modeling (stage three of our workflow). In the following section, we seek to understand whether the perturbed nodes discovered in the first stage of this workflow impact genome-scale flux networks. As described below, we use the findings from stage two to model pseudo-metabolic steady states.

Stage three: genome-scale modeling provides mechanistic insights into strain variation

In stage three, genome-scale models provide contextual basis for the analysis of multi-scale omics data sets (Feist and Palsson, 2008; Oberhardt et al., 2009; O'Brien et al., 2015). Instead of only looking at one reaction, metabolite, or protein at a time, multiple reactions are modeled and assessed simultaneously, which helps in gaining insight into the reaction system as a whole. It is important to note that, while reduction of multidimensional data is an important principles of stage two, reduction of network-level information can be non-informative and misleading (e.g. if an important metabolic lead lies in a peripheral pathway not in the core metabolism). Here, we use the comprehensive biochemical content of the metabolic network reconstruction of *E. coli (Orth et al., 2011)* and the predictive capability of constraint-based modeling approaches (O'Brien et al., 2015; Orth et al., 2010) to elucidate metabolic perturbations through the chemical connections contained in the reconstruction.

The identification of the three phases from PCA implies that each phase is a different metabolic state with a unique phenotype. While all nine strains share a similar characteristic three-phase behavior, the metabolites driving the variation in a given phase differ greatly (see stage two). This supports the hypothesis that even small variations in pathway engineering could lead to significant changes in endogenous metabolism. To investigate this, we carried out flux

balance analysis (FBA) together with a Markov chain Monte Carlo-based (MCMC) sampling approach (Almaas et al., 2004) on each of the three phases for each strain. As discussed in detail below, this analysis shows that the exometabolome causes significant shifts in key reaction fluxes (p-value < 0.05 using an empirical test) relative to WT (Figure 4 steps (3) and (4)). Most importantly, these shifting reactions cluster around the highly perturbed nodes that are observed in both metabolomic and proteomic data sets.

Significantly changing reaction fluxes indicate an increase or decrease in the flux (or flow of metabolites through a reaction), relative to WT. For each phase, we identified the most perturbed reaction fluxes, clustered the shifting reactions to find any common links between these nodes, and visualized the clusters graphically. The majority of shifting pathways include reactions in the pentose phosphate pathway (PPP), glycolysis/gluconeogenesis, and TCA (Figure 5 (a-c), respectively), with the exception of some peripheral reactions (e.g. phosphopentomutase-2 deoxyribose). For high-producing strains, most of the significant shifting reactions (relative to WT) occur either in late exponential phase (phase I, Figure 5 (d)) or early stationary phase (phase II, Figure 5 (e)). Interestingly, we see strain-specific groupings in the types of reactions that shift significantly in these phases. For example, in strains L2 and B2 we observe increased flux through specific reactions in PPP, namely phosphogluconate dehydrogenase (GND) and 6-phosphogluconolactonase (PGL), whereas for isopentenol-producing strains we see large-scale perturbation in flux networks surrounding triose phosphate isomerase (TPI), sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase (FBA3), transaldolase (TALA), and transketolase (TKT1, TKT2) (see Figure 5 (d), (e)). Other phenotypic changes become pronounced in certain strains during early stationary phase (phase II), such as flux diverting to TCA pathway reactions including alpha-ketoglutarate dehydrogenase (AKGDH), aconitase

(ACONTa, ACONTb), citrate synthase (CS), and isocitrate dehydrogenase (ICDHyr) (Figure 5 (e) and Supplementary Figure S6).

Visualization of these perturbed reaction nodes brings about a striking commonality that many are NADPH-producing reactions, such as ICDHyr, GND, and those related to specific amino acid biosynthetic pathways (e.g., AKGDH; Figure 6 (a)). Ultimately, modeling indicates that the cumulative flux to NADPH-producing reactions is significantly elevated for higher-producing strains (Supplementary Figure S7). This observation is consistent with previous work that identified NADPH availability as a limiting factor in isoprenoid production in *Saccharomyces cerevisiae*, the native host for the mevalonate pathway (Asadollahi et al., 2009). One explanation for an apparent depletion of NADPH is related to the NADPH-dependent HMG-CoA reductase (HMGR), which catalyzes the second step of the heterologous mevalonate pathway in each engineered strain. Due to the action of this enzyme, two molecules of NADPH are consumed for each molecule of mevalonate that is produced, coupling high biofuel titers with increased demand for NADPH.

In summary, we turn to genome-scale modeling as a diverse tool for elucidating the underlying biology of pathway optimization. The third stage of this workflow serves as a predictive method to connect various perturbed nodes in mechanistic detail and highlight a common function link, as opposed to a static network or statistical analysis, which mainly provides descriptive purpose. While the apparent NADPH limitation highlighted by these simulations seems obvious in retrospect, it is important to note that understanding how networks re-route to accommodate such bottlenecks is less trivial. In the section that follows, we describe how tracing perturbations through a genome-scale flux network helps elucidate experimentally

observed metabolic perturbations that, upon first glance, have no apparent connection with the NADPH node.

Model-aided predictions of engineered metabolic phenotypes are consistent with experiments

Perturbations in the intracellular flux networks, identified through modeling, can be cross-validated with complementary data sets, such as intracellular metabolomics and proteomics data. As mentioned in the above section, modeling intracellular flux networks makes use of uptake and secretion rates of glucose, organic acids, amino acids, and the fuel product. Therefore, the consistency of model predictions can be evaluated by comparing them to significantly perturbed nodes observed in the data. In this section, we demonstrate how three different data types, metabolomics, proteomics, and genome-scale flux predictions, are reconciled and corroborate our model-driven hypothesis that specific metabolic pathways re-route to meet the demands of pathway-induced NADPH depletion.

Our findings suggest that NADPH is depleted in engineered strains and that heterologous expression of HMGR is the main source of this behavior. The largest perturbations in reactions linked to a common NADPH node are found in strains expressing high levels of HMGR protein (e.g. strain L2, the sole engineered strain with pathway genes on a high-copy plasmid, has HMGR levels 10-20 fold higher than any other strain). Consistent with increased flux through the HMGR reaction, intracellular concentrations of $NADP^+$ in strain L2 are significantly elevated compared to other strains, (Figure 6 (b) box A), linking HMGR expression with NADPH depletion. Furthermore, the cellular demand for NADPH appears to perturb several reactions in glycolysis and the TCA cycle, in accordance with both modeling and experiments related to

317 strain L2: phosphoglycerate, citrate, α-ketoglutarate, and malate levels increase by nearly 10-fold,
318 5-fold, and 3-fold, respectively, during the time-course (see Figure 6 (b) boxes B and C).

319 While strain L2 is useful in establishing a clear link between HMGR expression and
320 NADPH depletion, strain I3, the top performing strain on the basis of yield and product titer,
321 provides even more convincing evidence of a metabolic response to pathway optimization, and
322 consequently, NADPH depletion. Model predictions for strain I3 indicate that key nodes in
323 glycolysis/gluconeogenesis, such as glyceraldehyde-3-phosphate dehydrogenase (GAPD),
324 triphosphate isomerase (TPI), and enolase (ENO), divert flux to provide the cell with routes to
325 NADPH regeneration. One route for regenerating NADPH is through the PPP (e.g., GND and
326 glucose-6-phosphate dehydrogenase, or G6PDH2r). Constructing dynamic difference profiles
327 (stage one) from proteomics data, we find perturbations in key PPP proteins (e.g., G6PDH2r,
328 GND, TALA, and TKT1) that are consistent with model predictions (Figure 6 (c) and
329 Supplementary Table S2). Another route the cell uses for regenerating NADPH is through the
330 TCA cycle (e.g., ICDHyr). Model predictions of increased flux through the TCA cycle are
331 consistent with metabolomic measurements for strain I3: intracellular citrate and aconitase levels
332 increase by 2-3 fold over WT (Figure 6 (b) box C) and dynamic difference profiles for proteins
333 involved in these reactions increase by 2 to 3-fold over WT (e.g. CS and ACONTb protein levels;
334 Figure 6 (c)). In this context, a previously perplexing observation - the apparent "shunting" of
335 assimilated acetate into the TCA cycle rather than the mevalonate pathway (Supplementary
336 Figure S8) is succinctly explained as a means to regenerate NADPH through ICDHyr rather than
337 deplete it through HMGR.

338 While we do not observe significant shifts in the levels of some PPP proteins that would
339 play an active role in NADPH regeneration, such as GND (perhaps partially due to low signal

intensity or noise), we do find that many have significant increases (p-value < 0.05 using an empirical test) in RNA levels; particularly >5-fold changes (over WT) for the expression of GND (unpublished data). Intriguingly, ENO and AKGDH expression levels are also 5-fold over WT, which coincides with increases in metabolite levels of α-ketoglutarate-derived amino acids (e.g., glutamate, glutamine, histidine, arginine), which were found to be enriched in high-producing isopentenol strains in stage one of this workflow.

Taken together, reconciliation of metabolomics and proteomics with genome-scale modeling proposes a general mechanism by which the cell responds to HMGR-mediated cofactor depletion by redirecting flux through the TCA cycle and/or PPP to regenerate NADPH. Importantly, the workflow highlights NADPH regeneration as a potential engineering strategy to improve mevalonate pathway function and product yields. As a validation, we attempted to address NADPH depletion not through mevalonate pathway engineering (Ma et al., 2011), but by identifying single gene knockouts (SKOs) that re-route flux to produce higher product yield.

Identifying metabolic properties relevant to re-engineering

Using the knowledge gained from the three-stage workflow, we reevaluated our constraint-based modeling simulations in the presence of single gene knockouts (SKOs). We were interested in discovering SKOs that re-route flux in pathways that compete with the mevalonate pathway and are related to NADPH production/depletion. As a proof-of-concept, we generated model-driven predictions of SKO candidates using the genome-scale metabolic model of strain I3, which produced the most biofuel and showed strong evidence for NADPH depletion.

Using flux variability analysis, we rank-ordered SKO candidates using several metrics that were identified to be important factors underlying strain variation in stage three of this workflow: (i) minimized flux through NADPH-consuming reactions; (ii) maximized flux through NADPH-

producing reactions; and (iii) maximized flux through PPP reactions. The SKO candidates identified using these metrics are provided in Table 1 and in Supplementary Information (see Table S3 and an iPython notebook titled, "Engineering_SKOs.ipynb").

To test the effects of the model-predicted SKOs, we experimentally tested the top four SKO candidates in strain I3 (Supplementary Figure S9). We found that one of the SKOs, *ΔydbK* (blattner code b1378), shows especially intriguing characteristics related to enhanced biofuel production. This gene is predicted to act as a pyruvate synthase (reaction POR5 in the iJO1366 (Orth et. al. 2011)), which converts pyruvate to acetyl-CoA. When this gene is knocked out from strain I3 (I3 Δ ydbK), we observe an almost 2-fold increase in the specific (i.e., growth-normalized) production of isopentenol (Figure 7). In I3 Δ ydbK, the specific production is increased at every time point during batch fermentation, and by 48 hours, I3ΔydbK shows a higher absolute titer of isopentenol (920 mg/L versus 800 mg/L for strains I3ΔydbK and I3, respectively). Intriguingly, Δ ydbK also significantly increases the specific production of limonene in strain L3, which also demonstrated strong evidence for NADPH depletion (Supplementary Figure S10), suggesting a commonality between isopentenol and limonene producing strains. In contrast, this SKO has minimal effects on bisabolene production in strain B2.

CONCLUSION

To date, the majority of metabolic engineering efforts serve as demonstrations of future potential rather than industry-ready technology. Achieving large-scale, economical production of microbial-derived products requires production strains to be exhaustively optimized for high yields and productivities. The challenges associated with this goal are numerous and massive in scope, including pathway "balancing" with appropriate protein expression and activity, product

and metabolite toxicity, feedback inhibition, strict energetic requirements, cofactor imbalances, and competition with endogenous pathways (Paddon and Keasling, 2014). Thus, the inherent complexity of biological systems makes them difficult to effectively design and control (Endy, 2005).

Greater synergy between systems biology, metabolic engineering, and synthetic biology would greatly benefit all three disciplines, given their complementary, yet classically separate, approaches to bioengineering (Nielsen et al., 2014). A major challenge in both metabolic engineering and synthetic biology is understanding how the introduction of engineered or non-native components into a biochemical network influences the behavior of the entire system. To meet this challenge, we have developed a three-stage workflow to interpret complex multi-omics data for multi-strain characterization. Each of the three stages of the workflow works together as a concerted pipeline to efficiently process highly dimensional datasets.

The first two stages of the workflow act as a flexible framework to interpret raw, multi-omics data by sorting strain phenotypes based on their dynamic difference profiles and correlating measurements based on distinct patterns derived from thousands of measurements. These two stages of the workflow in particular are well-suited for integration with high-throughput strain engineering and analysis pipelines, where "manual" assessment of convoluted omics data is not feasible. While evaluating multi-omics data using statistics-based approaches, such as multivariate analyses, has been shown to be useful as a stand-alone method for making sense of highly over-expressed heterologous pathways(Alonso-Gutierrez et al., 2015), unraveling the global response of the cell to pathway engineering requires moving beyond statistics-based approaches and incorporating system-wide analyses. To account for the systems-level response

409 to pathway engineering, the third stage of this workflow leverages these statistics-based
410 approaches with genome-scale metabolic modeling.

411 Here, we demonstrate that through the pairing of synthetic pathway construction and a
412 systems-level, model-driven analysis, our multi-omics-based workflow successfully reconciles
413 metabolomics data, proteomics data, and predictions from genome-scale models. Using
414 mevalonate pathway engineering as a case study, we demonstrate that our approach is capable of
415 elucidating the complex interplay between heterologous pathway engineering and endogenous
416 metabolism in a microbial host. The utility of such a workflow is expected to become
417 increasingly important with the parallel, accelerating advances in technologies related to strain
418 generation and high-throughput analyses (e.g., on the order of thousands of strains and thousands
419 of measurements).

420

421 **ACKNOWLEDGEMENTS**

422 This work was funded by the Joint BioEnergy Institute (http://www.jbei.org/), which is
423 supported by the US Department of Energy, Office of Science, Office of Biological and
424 Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley
425 National Laboratory and the US Department of Energy (K.W.G., J.A-G., M.T. H.G.M. E.B.,
426 C.J.P., P.D.A., J.D.K., T.S.L.), by the Swiss National Science Foundation (grant p2elp2_148961
427 to E.B.), and by the National Institutes of Health (grant GM057089 to B.O.P.). This research
428 used resources of the National Energy Research Scientific Computing Center, a DOE Office of
429 Science User Facility supported by the Office of Science of the U.S. Department of Energy
430 under Contract No. DE-AC02-05CH11231. We also gratefully acknowledge Dr. Daniel Zielinski,
431 Dr. Aarash Bordbar, Chris Shymansky, Jennifer Gin, Dr. Josh Lerman, and Professor Vassily

Hatzimanikatis for early input in the project as well as Ali Ebrahim for technical support. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

AUTHOR CONTRIBUTIONS

Conceptualization, K.W.G, J.A-G., E.B., and T.S.L.; Methodology, E.B., K.W.G., J.A-G., C.J.P., T.S.B., L.Y., D.M., and J.M.; Investigation, E.B., K.W.G, J.A-G., M.T., and E.B.; Writing – Original Draft, E.B. and K.W.G.; Writing – Review & Editing, E.B., K.W.G., T.S.L., B.O.P., J.D.K. P.D.A. H.G.M., A.F., E.J.O., C.J.P.; Funding Acquisition, T.S.L and J.D.K; Resources, T.S.L and J.D.K.; Supervision, T.S.L, J.D.K., and B.O.P.

EXPERIMENTAL PROCEDURES

All chemicals and media components were purchased from Sigma-Aldrich (St. Louis, MO), VWR (West Chester, PA), or Fischer Scientific (Pittsburgh, PA) and used without modification unless noted otherwise. The *E. coli* strains used in the work, DH10B and DH1, were purchased from Invitrogen (Carlsbad, CA) and ATCC, respectively. For proteomics experiments, mass spectrometric-grade trypsin was purchased from Sigma-Aldrich.

Plasmid and strain construction

E. coli DH10B was the host for pathway cloning and plasmid manipulations, while *E. coli* DH1 was used as the production host. Plasmids were assembled according to the BglBrick standard(Anderson et al., 2010) using standardized vectors(Lee et al., 2011) with the exception

of pTrc99A(Amann et al., 1988). Transformations were performed with chemically-competent cells as described previously (Chung et al., 1989). A list of plasmids and strains used in this study is provided in Supplementary Table S1.

Growth conditions and production of advanced biofuels

Seed cultures of 8 biofuel-producing strains (Supplementary Table S1) and *E. coli* DH1 were grown overnight in 5 mL volumes of LB medium with appropriate antibiotics at 37°C. For production, 100 mL volumes of EZ-Rich defined medium with 1% glucose in a 1 L Erlenmeyer flask were inoculated with seed cultures to an initial OD_{600} of 0.1. Production cultures were grown in rotary shakers (200 rpm) at 30°C to an OD_{600} of 0.6 and induced with 500 µM isopropyl β-D-1-thiogalactopyranoside (IPTG). For limonene- and bisabolene-producing strains, a 10% overlay of dodecane was added at induction. Proof-of-concept gene deletion experiments were carried out in 25 mL of EZ-Rich defined media in 250 mL Erlenmeyer flasks with the same induction parameters.

Metabolomics sampling and analysis

Samples were collected for metabolomics and proteomics at set time-points during the batch fermentation following induction (note that time "0" corresponds to immediately prior to induction rather than the initial inoculation). For metabolite sampling, 1.8 mL of culture was harvested by centrifugation in a 2 mL microfuge tube. From this sample, 0.1 mL was used to measure OD_{600} and 0.2 mL was frozen at -20°C and extracted with ethyl acetate to measure isopentenol (see(George et al., 2014) for GC-FID methods). For limonene- and bisabolene-producing strains, 0.1 mL of dodecane overlay was collected and diluted into ethyl acetate (see(Peralta-Yahya et al., 2011) and(Alonso-Gutierrez et al., 2013) for GC-MS methods). The

480 remaining volume (1.5 mL) was pelleted (14,000 x g) in a tabletop centrifuge at 4°C. For
481 organic acid analysis, 0.25 mL of supernatant was collected and frozen at -20°C until analysis.
482 Another 0.25 mL of supernatant was mixed 1:1 with ice cold MeOH and stored at -20°C for the
483 quantification of extracellular metabolites. The remaining supernatant was decanted, and the
484 pellet was resuspended in 0.3 mL of ice cold MeOH and stored at -20°C for the quantification of
485 intracellular metabolites. The intracellular metabolite sample was vortexed thoroughly and
486 centrifuged (8000 x g) for 10 minutes at 4 C. The supernatant was collected, mixed with a 1:1
487 volume of water, and filtered through a Millipore™ Amicon Ultra 3kD MW cut-off filter (14000
488 x g for 60 minutes at -2°C). Extracellular metabolite samples were filtered in an identical
489 manner. Water was added to the flow-through to a final volume of 1 mL and the samples were
490 lyophilized overnight. Samples were reconstituted in 90 µL MeOH:H_2O (1:1) prior to analysis.
491 Extracellular organic acids were analyzed by an Agilent 1200 Series HPLC system equipped
492 with a photodiode array detector set a 210, 254, and 280 nm. Metabolites were separated on an
493 Aminex HPX-87H column with 8% cross linkage (150 mm length, 7.8 mm internal diameter, 9
494 µm particle size; Bio-Rad, Richmond, CA, USA). An isocratic elution was performed using 4
495 mM sulfuric acid with a flow rate of 0.6 mL/min. A refractive index detector (RID) was used to
496 detect organic acids and glucose, while pyruvate was detected with a diode array detector (DAD)
497 at 210 nm.
498 Intracellular and extracellular metabolites were analyzed by liquid chromatography mass
499 spectrometry (LC-MS) on a ZIC-HILIC column (150 mm length, 2.1 mm internal diameter, 2.5
500 µm particle size) using an Agilent 1200 Series HPLC. The HPLC system was coupled to an
501 Agilent 6210 time-of-flight mass spectrometer by a 1/2 post-column split. All metabolites were
502 quantified via eight-point calibration curves ranging from 781.25 nM to 200 µM. A variety of

methods were used to quantify different classes of metabolites. Please see previous references for details on the quantification of glycolysis and TCA cycle intermediates(Juminaga et al., 2012), amino acids(Bokinsky et al., 2013), organic acids(Juminaga et al., 2012), nucleotides and CoAs(Bokinsky et al., 2013), and mevalonate pathway intermediates(Weaver et al., 2015).

Proteomics sampling and analysis

For proteomics, 5 mL of culture was collected by centrifugation in a 15 mL falcon tube (5000 x g at 4°C). Supernatant was decanted, and cell pellets were stored at -80°C prior to analysis. Cell pellets were extracted with chloroform/methanol and protein samples were prepared as described previously (Redding-Johanson et al., 2011). Following drying in a vacuum concentrator (ThermoSavant), the protein pellet was resuspended in ammonium bicarbonate and quantified using DC Protein reagent (BioRad, Hercules, CA). Protein was diluted to 0.5 mg/mL and disulfide bonds were reduced with tris(2-carboxyethyl)phosphine (TCEP) for 30 minutes at room temperature followed by disulfide bond blocking with 10 mM iodoacetamide at room temperature in the dark for 30 minutes. Samples were analyzed using an AB Sciex (Foster City, CA) 5500Q-Trap mass spectrometer operating in MRM (SRM) mode coupled to an Agilent 1100 system. For method details, please see references (George et al., 2014) and (Batth et al., 2014).

Constraint-based modeling

Constraint-based modeling and analysis of metabolic networks have been extensively reviewed and described elsewhere (Bordbar et al., 2014; Feist and Palsson, 2008; Price et al., 2004). To summarize, all of the reactions in a metabolic network can be described mathematically by a stoichiometric matrix, *S*, which has dimensions $m \times n$ *(*the number of total metabolites and reactions in a model, respectively). Each element in *S* represents a stoichiometric coefficient of

527 the metabolite in its respective reaction. The mass balance equations at steady state are
528 represented as

529 $$S \cdot v = 0 \quad [1]$$

530 where *v* is a flux vector, indicating the direction of flux through a reaction. Constraints on the
531 system can be imposed such that fluxes range between a defined maximum and minimum and
532 can be reversible or irreversible,

533 $$v_{min} \leq v \leq v_{max} \quad [2]$$

534 Once the topology and set of constraints is known, the model can be used with various
535 constraint-based methods (Bordbar et al., 2014) to understand or predict cellular phenotypes.
536 In this work, *S* matrix was constructed from a previous reconstruction (Orth et al., 2011).
537 Heterologous genes and non-native mevalonate pathway intermediates were added to the model
538 in the form of mass and charge balanced reactions. Select metabolites, known to cross the cell
539 membrane (based on extracellular measurements), were added as exchange reactions, allowing
540 those metabolites to leave or enter the extracellular space in the model. When available, uptake
541 and secretion rates were used from new or published data to constrain the upper and lower
542 bounds of the exchange reactions (George et al., 2014). All parameters are detailed in
543 Supplementary Files, Metabolomics and Proteomics Data Analyses.
544
545 *Monte Carlo sampling*
546 Markov chain Monte Carlo (MCMC) sampling was used to generate a set of feasible
547 distributions of fluxes in the genome-scale network. The method uses the artificially centered hit-
548 and-run algorithm with slight modifications. In the initial step, the algorithm generates a set of
549 non-uniform, pseudo-random points. Through several iterations, each of the flux points in the
550 network is randomly modified such that they remain within the feasible solution space. The

551 random modifications follow specific rules: (i) the direction in which the point is moved is
552 random, (ii) the amount of space a point travels is limited and (iii) a new random point is then
553 chosen along the new line. If carried out for enough time, the set of points will distribute
554 uniformly throughout flux space and will provide a distribution (or range) in fluxes through a
555 given reaction. This range represents most likely flux for a given reaction in the metabolic
556 network, and depends on the network topology and model constraints. For more details, see the
557 Supplementary Information.
558
559 *Predicting phenotypic differences between wild-type and engineered strains*
560 The phenotypic differences between wild-type and engineered strains were computed based on
561 published data (George et al., 2014; Peralta-Yahya et al., 2012) as well as newly generated
562 datasets. Using Principal Component Analysis (PCA), we determined the phases of growth (i.e.
563 exponential, late exponential, early stationary phase and late stationary phase) and used the
564 average uptake and secretion of extracellular metabolites as input to the genome-scale models.
565 For single measured values of a given metabolite or peptide, we estimated the variance using
566 triplicate measurements taken for the same metabolite or peptide from wild-type cultures at the
567 initial time-point. The original dataset for 9 strains provided secretion and uptake measurements
568 for glucose and many of the major organic acids (e.g., formate, succinate, pyruvate, acetate, etc.).
569 A validation set provided amino acid uptake measurements and additional measurements in the
570 exponential phase of growth. See Supplementary Information for more details.
571 Simulations were conducted using the *i*JO1366 model of *E. coli* K12-MG1655. Comparisons to a
572 different *E. coli* metabolic reconstruction, the DH1 strain, did not show significant changes in
573 model topology. The models were allowed to take up the same substrates provided
574 experimentally at rates consistent with the data (Supplementary Files, "Proteomics Data Analysis"

575 and “Metabolomics Data Analysis”). Monte Carlo sampling was used to identify all feasible flux
576 states. This was done for different phases of growth, ranging late exponential to early and late
577 stationary phases. A *z* score-based analysis was carried out to determine the most significantly
578 shifting fluxes between wild-type and engineered strains as well as between various regions of
579 growth (i.e. exponential versus stationary phase). The z-score calculation was repeated 1,000
580 times and the mean value is reported. Comparisons with the experimental data were done by
581 calculating differences in concentration and peak areas for the metabolomics and proteomics data
582 sets. Similarly, z-scores were computed to determine which shifts were significant over others.
583 See Supplementary Notes for more details.
584

585 **REFERENCES**

586 Aebersold, R., Mann, M., 2003. Mass spectrometry-based proteomics. Nature 422, 198–207.

587 Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z.N., Barabási, A.-L., 2004. Global organization of
588 metabolic fluxes in the bacterium Escherichia coli. Nature 427, 839–843.

589 Alonso-Gutierrez, J., Chan, R., Batth, T.S., Adams, P.D., Keasling, J.D., Petzold, C.J., Lee, T.S.,
590 2013. Metabolic engineering of Escherichia coli for limonene and perillyl alcohol production.
591 Metab. Eng. 19, 33–41.

592 Alonso-Gutierrez, J., Kim, E.-M., Batth, T.S., Cho, N., Hu, Q., Chan, L.J.G., Petzold, C.J.,
593 Hillson, N.J., Adams, P.D., Keasling, J.D., Garcia Martin, H., Lee, T.S., 2015. Principal
594 component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. Metab. Eng.
595 28, 123–133.

596 Amann, E., Ochs, B., Abel, K.J., 1988. Tightly regulated tac promoter vectors useful for the

expression of unfused and fused proteins in Escherichia coli. Gene 69, 301–315.

Anderson, J.C., Dueber, J.E., Leguia, M., Wu, G.C., Goler, J.A., Arkin, A.P., Keasling, J.D., 2010. BglBricks: A flexible standard for biological part assembly. J. Biol. Eng. 4, 1.

Asadollahi, M.A., Maury, J., Patil, K.R., Schalk, M., Clark, A., Nielsen, J., 2009. Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering. Metab. Eng. 11, 328–334.

Bailey, J.E., 1991. Toward a science of metabolic engineering. Science 252, 1668–1675.

Batth, T.S., Singh, P., Ramakrishnan, V.R., Sousa, M.M.L., Chan, L.J.G., Tran, H.M., Luning, E.G., Pan, E.H.Y., Vuu, K.M., Keasling, J.D., Adams, P.D., Petzold, C.J., 2014. A targeted proteomics toolkit for high-throughput absolute quantification of Escherichia coli proteins. Metab. Eng. 26C, 48–56.

Berger, B., Peng, J., Singh, M., 2013. Computational solutions for omics data. Nat. Rev. Genet. 14, 333–346.

Bokinsky, G., Baidoo, E.E.K., Akella, S., Burd, H., Weaver, D., Alonso-Gutierrez, J., García-Martín, H., Lee, T.S., Keasling, J.D., 2013. HipA-triggered growth arrest and β-lactam tolerance in Escherichia coli are mediated by RelA-dependent ppGpp synthesis. J. Bacteriol. 195, 3173–3182.

Bordbar, A., Monk, J.M., King, Z.A., Palsson, B.O., 2014. Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. 15, 107–120.

Chou, H.H., Keasling, J.D., 2012. Synthetic pathway for production of five-carbon alcohols from isopentenyl diphosphate. Appl. Environ. Microbiol. 78, 7849–7855.

618 Chung, C.T., Niemela, S.L., Miller, R.H., 1989. One-step preparation of competent Escherichia
619 coli: transformation and storage of bacterial cells in the same solution. Proc. Natl. Acad. Sci. U.
620 S. A. 86, 2172–2175.

621 de Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C.,
622 Mann, M., 2008. Comprehensive mass-spectrometry-based proteome quantification of haploid
623 versus diploid yeast. Nature 455, 1251–1254.

624 de Jong, B., Siewers, V., Nielsen, J., 2012. Systems biology of yeast: enabling technology for
625 development of cell factories for production of advanced biofuels. Curr. Opin. Biotechnol. 23,
626 624–630.

627 Dunlop, M.J., Dossani, Z.Y., Szmidt, H.L., Chu, H.C., Lee, T.S., Keasling, J.D., Hadi, M.Z.,
628 Mukhopadhyay, A., 2011. Engineering microbial biofuel tolerance and export using efflux
629 pumps. Mol. Syst. Biol. 7, 487.

630 Endy, D., 2005. Foundations for engineering biology. Nature 438, 449–453.

631 Feist, A.M., Palsson, B.Ø., 2008. The growing scope of applications of genome-scale metabolic
632 reconstructions using Escherichia coli. Nat. Biotechnol. 26, 659–667.

633 Fuhrer, T., Zamboni, N., 2015. High-throughput discovery metabolomics. Curr. Opin.
634 Biotechnol. 31, 73–78.

635 George, K.W., Alonso-Gutierrez, J., Keasling, J.D., Lee, T.S., 2015a. Isoprenoid Drugs, Biofuels,
636 and Chemicals—Artemisinin, Farnesene, and Beyond.

637 George, K.W., Chen, A., Jain, A., Batth, T.S., Baidoo, E.E.K., Wang, G., Adams, P.D., Petzold,
638 C.J., Keasling, J.D., Lee, T.S., 2014. Correlation analysis of targeted proteins and metabolites to

assess and engineer microbial isopentenol production. Biotechnol. Bioeng. 111, 1648–1658.

George, K.W., Thompson, M.G., Kang, A., Baidoo, E., Wang, G., Chan, L.J.G., Adams, P.D., Petzold, C.J., Keasling, J.D., Lee, T.S., 2015b. Metabolic engineering for the high-yield production of isoprenoid-based C5 alcohols in E. coli. Sci. Rep. 5, 11128.

Gross, M., 2011. Riding the wave of biological data. Curr. Biol. 21, R204–6.

Han, M.-J., Jeong, K.J., Yoo, J.-S., Lee, S.Y., 2003. Engineering Escherichia coli for increased productivity of serine-rich proteins based on proteome profiling. Appl. Environ. Microbiol. 69, 5772–5781.

Han, M.J., Yoon, S.S., Lee, S.Y., 2001. Proteome analysis of metabolically engineered Escherichia coli producing Poly(3-hydroxybutyrate). J. Bacteriol. 183, 301–308.

Janßen, H.J., Steinbüchel, A., 2014. Fatty acid synthesis in Escherichia coli and its applications towards the production of fatty acid based biofuels. Biotechnol. Biofuels 7, 7.

Jullesson, D., David, F., Pfleger, B., Nielsen, J., 2015. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. Biotechnol. Adv. doi:10.1016/j.biotechadv.2015.02.011

Juminaga, D., Baidoo, E.E.K., Redding-Johanson, A.M., Batth, T.S., Burd, H., Mukhopadhyay, A., Petzold, C.J., Keasling, J.D., 2012. Modular engineering of L-tyrosine production in Escherichia coli. Appl. Environ. Microbiol. 78, 89–98.

Jung, Y.K., Kim, T.Y., Park, S.J., Lee, S.Y., 2010. Metabolic engineering of Escherichia coli for the production of polylactic acid and its copolymers. Biotechnol. Bioeng. 105, 161–171.

659 Kabir, M.M., Shimizu, K., 2003. Fermentation characteristics and protein expression patterns in
660 a recombinant Escherichia coli mutant lacking phosphoglucose isomerase for poly(3-
661 hydroxybutyrate) production. Appl. Microbiol. Biotechnol. 62, 244–255.

662 Kahn, S.D., 2011. On the future of genomic data. Science 331, 728–729.

663 Kuehnbaum, N.L., Britz-McKibbin, P., 2013. New advances in separation science for
664 metabolomics: resolving chemical diversity in a post-genomic era. Chem. Rev. 113, 2437–2468.

665 Kwok, R., 2010. Five hard truths for synthetic biology. Nature 463, 288–290.

666 Landels, A., Evans, C., Noirel, J., Wright, P.C., 2015. Advances in proteomics for production
667 strain analysis. Curr. Opin. Biotechnol. 35, 111–117.

668 Lee, S.Y., Lee, D.-Y., Kim, T.Y., 2005. Systems biotechnology for strain improvement. Trends
669 Biotechnol. 23, 349–358.

670 Lee, T.S., Krupa, R.A., Zhang, F., Hajimorad, M., Holtz, W.J., Prasad, N., Lee, S.K., Keasling,
671 J.D., 2011. BglBrick vectors and datasheets: A synthetic biology platform for gene expression. J.
672 Biol. Eng. 5, 12.

673 Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., Green, E.D.,
674 2014. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing
675 on biomedical big data. J. Am. Med. Inform. Assoc. 21, 957–958.

676 Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., Keasling, J.D., 2003. Engineering a
677 mevalonate pathway in Escherichia coli for production of terpenoids. Nat. Biotechnol. 21, 796–
678 802.

679 Ma, S.M., Garcia, D.E., Redding-Johanson, A.M., Friedland, G.D., Chan, R., Batth, T.S.,
680 Haliburton, J.R., Chivian, D., Keasling, J.D., Petzold, C.J., Lee, T.S., Chhabra, S.R., 2011.
681 Optimization of a heterologous mevalonate pathway through the use of variant HMG-CoA
682 reductases. Metab. Eng. 13, 588–597.

683 Metzker, M.L., 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31–46.

684 Nielsen, J., Fussenegger, M., Keasling, J., Lee, S.Y., Liao, J.C., Prather, K., Palsson, B., 2014.
685 Engineering synergy in biotechnology. Nat. Chem. Biol. 10, 319–322.

686 Oberhardt, M.A., Palsson, B.Ø., Papin, J.A., 2009. Applications of genome-scale metabolic
687 reconstructions. Mol. Syst. Biol. 5.

688 O'Brien, E.J., Monk, J.M., Palsson, B.O., 2015. Using Genome-scale Models to Predict
689 Biological Capabilities. Cell 161, 971–987.

690 Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M., Palsson, B.Ø., 2011. A
691 comprehensive genome-scale reconstruction of Escherichia coli metabolism?2011. Mol. Syst.
692 Biol. 7.

693 Orth, J.D., Thiele, I., Palsson, B.Ø., 2010. What is flux balance analysis? Nat. Biotechnol. 28,
694 245–248.

695 Paddon, C.J., Keasling, J.D., 2014. Semi-synthetic artemisinin: a model for the use of synthetic
696 biology in pharmaceutical development. Nat. Rev. Microbiol. 12, 355–367.

697 Palsson, B., Zengler, K., 2010. The challenges of integrating multi-omic data sets. Nat. Chem.
698 Biol. 6, 787–789.

699 Peralta-Yahya, P.P., Ouellet, M., Chan, R., Mukhopadhyay, A., Keasling, J.D., Lee, T.S., 2011.
700 Identification and microbial production of a terpene-based advanced biofuel. Nat. Commun. 2,
701 483.

702 Peralta-Yahya, P.P., Zhang, F., del Cardayre, S.B., Keasling, J.D., 2012. Microbial engineering
703 for the production of advanced biofuels. Nature 488, 320–328.

704 Picotti, P., Aebersold, R., 2012. Selected reaction monitoring-based proteomics: workflows,
705 potential, pitfalls and future directions. Nat. Methods 9, 555–566.

706 Price, N.D., Reed, J.L., Palsson, B.Ø., 2004. Genome-scale models of microbial cells: evaluating
707 the consequences of constraints. Nat. Rev. Microbiol. 2, 886–897.

708 Primak, Y.A., Du, M., Miller, M.C., Wells, D.H., Nielsen, A.T., Weyler, W., Beck, Z.Q., 2011.
709 Characterization of a feedback-resistant mevalonate kinase from the archaeon Methanosarcina
710 mazei. Appl. Environ. Microbiol. 77, 7772–7778.

711 Redding-Johanson, A.M., Batth, T.S., Chan, R., Krupa, R., Szmidt, H.L., Adams, P.D., Keasling,
712 J.D., Lee, T.S., Mukhopadhyay, A., Petzold, C.J., 2011. Targeted proteomics for metabolic
713 pathway optimization: application to terpene production. Metab. Eng. 13, 194–203.

714 Ro, D.-K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A.,
715 Eachus, R.A., Ham, T.S., Kirby, J., Chang, M.C.Y., Withers, S.T., Shiba, Y., Sarpong, R.,
716 Keasling, J.D., 2006. Production of the antimalarial drug precursor artemisinic acid in
717 engineered yeast. Nature 440, 940–943.

718 Rolfsson, Ó., Palsson, B.O., 2015. Decoding the jargon of bottom-up metabolic systems biology.
719 Bioessays 37, 588–591.

720 Runguphan, W., Keasling, J.D., 2014. Metabolic engineering of Saccharomyces cerevisiae for
721 production of fatty acid-derived biofuels and chemicals. Metab. Eng. 21, 103–113.

722 Sarria, S., Wong, B., Martín, H.G., Keasling, J.D., Peralta-Yahya, P., 2014. Microbial Synthesis
723 of Pinene. ACS Synth. Biol. 3, 466–475.

724 Shen, C.R., Liao, J.C., 2013. Synergy as design principle for metabolic engineering of 1-
725 propanol production in Escherichia coli. Metab. Eng. 17, 12–22.

726 Steen, E.J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., Del Cardayre, S.B.,
727 Keasling, J.D., 2010. Microbial production of fatty-acid-derived fuels and chemicals from plant
728 biomass. Nature 463, 559–562.

729 Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz,
730 M.C., Sinha, S., Robinson, G.E., 2015. Big Data: Astronomical or Genomical? PLoS Biol. 13,
731 e1002195.

732 Trinh, C.T., Li, J., Blanch, H.W., Clark, D.S., 2011. Redesigning Escherichia coli metabolism
733 for anaerobic production of isobutanol. Appl. Environ. Microbiol. 77, 4894–4904.

734 Tyo, K.E., Alper, H.S., Stephanopoulos, G.N., 2007. Expanding the metabolic engineering
735 toolbox: more options to engineer cells. Trends Biotechnol. 25, 132–137.

736 Weaver, L.J., Sousa, M.M.L., Wang, G., Baidoo, E., Petzold, C.J., Keasling, J.D., 2015. A
737 kinetic-based approach to understanding heterologous mevalonate pathway function in E. coli.
738 Biotechnol. Bioeng. 112, 111–119.

739 Wolfe, A.J., 2005. The acetate switch. Microbiol. Mol. Biol. Rev. 69, 12–50.

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H.B., Andrae, S., Yang, T.H., Lee, S.Y., Burk, M.J., Van Dien, S., 2011. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat. Chem. Biol. 7, 445–452.

Zhang, Z., Wu, S., Stenoien, D.L., Paša-Tolić, L., 2014. High-throughput proteomics. Annu. Rev. Anal. Chem. 7, 427–454.

FIGURE LEGENDS

Figure 1: A workflow for bridging the genotype-phenotype relationship with multi-omics data and genome-scale models of *E. coli* metabolism expressing heterologous pathways. (a) Multi-scale data types that are generally collected to elucidate changes in metabolic phenotypes of engineered strains: information on gene expression, protein abundance, metabolite concentrations, and predicted fluxes in a genome-scale metabolic network. (b) Our workflow involves a hierarchical staging of computational analysis methods. In the first stage, basic strain differences are binned based on characteristic changes. The second stage seeks to find relevant patterns and correlations in the data. The last stage builds on knowledge gained from the previous stages to elucidate mechanisms of action in the context of a genome-scale network that can explain apparent differences in strain behavior.

Figure 2: Pathway assembly, strain selection and multi-omics data generation. (a) This study focuses on the characterization of three versions of a heterologous mevalonate pathway engineered to synthesize isopentenol, limonene, and bisabolene. (b) Over a 72-hour time course, the engineered strains show various levels of fuel production due to changes in heterologous pathway architecture and expression. Each strain is indicated by its respective color, shown in the legend to the right. (c) The nine (eight engineered and one wild-type) strains were further analyzed using a multi-omics approach to generate measurements for 86 metabolites and 55 protein complexes at 9 different time-points during batch fermentation to generate detailed omics profiles. The complete data set is available in Supplementary Information.

Figure 3: Systems-level multi-omics integration and analysis of batch fermentation dynamics. (a) The first stage of the workflow filters, maps, and identifies system level

differences between control (e.g., WT) and test (e.g., engineered strain) conditions through the construction of dynamic difference profiles. The filtered multi-omic dynamic profiles are mapped onto the metabolic network for quick identification of trends in different strain phenotypes. (b) A categorization of the differences between control and test conditions for each data type was filtered into dynamic profiles. The differences for each data point relative to the control were calculated, and the errors of the measurements were propagated to determine the range of change (from significant to not changing) between the control and test conditions. The plots in the left column refer to positive ("+") shifts, or points where the test condition is greater than the control in terms of concentration or flux. Those in the right column refer to negative ("-") shifts. Standard deviations for the test and control condition for each data point were calculated from triplicate measurements or estimated based on the percent root-squared deviation (%RSD) of representative triplicate measurements. (c) A cartoon depiction of the data types included in this analysis, namely endo- and exo- metabolomics, proteomics, and sampled flux distributions. Top left refers to protein level measured by proteomics, top right refers to the product measured through metabolomics, bottom left refers to the substrate measured through metabolomics and bottom left refers to predicted flux computed for a specific reaction.

Figure 4: Integrating multi-omics data with genome-scale models of metabolism. Stages two and three of the workflow combine multivariate analyses and genome-scale models of metabolism to understand how correlated a data set is, and establish patterns, or phases, that can then be used as a basis for metabolic modeling. (1) Standard metabolomics and growth measurements for over 80 metabolites were taken for nine different strains over a 72-hour time course. (2) Applying PCA on this dataset, we dramatically reduce the dimensionality of this dataset and find three distinct metabolic phases that align with different phases of the growth:

exponential, early, and late stationary phases. (3) These metabolic phases can be modeled using constraint-based methods, such as Markov-chain Monte Carlo based sampling, by taking the average of the extracellular measurements within a given phase as inputs to the model. Using constraint-based modeling, we observe perturbed reactions in host metabolism resulting from pathway engineering (illustrated by the red colored nodes in the network). These perturbed reactions can be clustered to determine common links, such as cofactor usage. (4) Model predictions, used to identify which reactions in the network are expected to shift between phases or across different strains, can then be validated with other omics datasets, such as proteomics.

Figure 5: Genome-scale modeling revealed perturbations in TCA cycle and pentose-phosphate pathway activity associated with certain engineered phenotypes. Reactions colored by the shift (absolute value) in flux in a top-producing strain, I3, compared to wild-type in different pathways in central carbon metabolism: (a) the pentose-phosphate pathway; (b) glycolysis/gluconeogenesis; (c) TCA cycle. Shown in (d) are significant reaction flux shifts ($p < 0.05$) corresponding to various reactions in these pathways in phase I (0-6 hrs) and those for phase II (6-20 hrs) are displayed in (e). Here, shifts in metabolic flux represent overall changes (both positive and negative perturbations) from wild-type behavior. All metabolic maps generated using Escher maps.

Figure 6: Constraint-based modeling elucidates pathways that allow for coupling of NADPH metabolism and biofuel production. Displayed in (a) are the main NADPH-producing and consuming reactions in the genome-scale model of *E. coli* that carry the majority of flux. The sum of flux through these reactions is significantly higher in top-producing strains over WT. This coincides with the increased expression of HMGR (a NADPH-dependent reaction) in these

strains as well as increased accumulation of intracellular NADP concentrations. In (b), increases in cofactor (box A), glycolysis/gluconeogenesis (box B) and TCA (box C) metabolite concentrations (relative to wild-type *E. coli*) indicate which regions in metabolism are perturbed in different engineered strains. In (c), the "dynamic difference profiles" identify changes in protein levels for isopentenol-producing strains. Proteins were clustered into dynamic difference profiles and further categorized by whether protein levels increase (green) or decrease (lavender) during the time-course. As shown in the lower left panel, key glycolysis (yellow), PPP (orange), and TCA (red) proteins shift above WT levels in higher producing strains (I2 and I3). On the lower right panel is an example of how progressive engineering efforts change the dynamic difference profile for acetate synthase (ACS). In this case, progressing from minimal engineering to optimized strain (I1 to I2 to I3), the dynamic difference profile changes from having decreasing protein levels early on in the time course to increased protein levels at the end of the time course. Shifts in protein levels of other strains are given as Supplementary Information.

Figure 7: Model-driven predictions discover a gene-knockout that increases the specific production of isopentenol. Growth-normalized isopentenol titer (mg/L/OD600) is displayed for strain I3 (black) and I3 with *Δ*ydbK knockout (gray). At every non-zero time point, the knockout variant produces significantly more isopentenol than the highest producing strain, I3 (stars denote p-values: 4 hrs $p = 0.0058$ (**), 8 hrs $p < 0.0001$ (****), 24 hrs $p = 0.002$ (***), 48 hrs $p = 0.0037$ (**), using an unpaired two-tail t-test). At 48 hours, absolute isopentenol titers are 920 mg/L versus 800 mg/L for strains I3*Δ*ydbK and I3, respectively.

Table 1: Model-based predictions select gene-knockouts that produce a desired phenotypic state. Single gene knock-out simulations were performed using the genome-scale model of *E. coli* (iJO1366) to identify candidate targets that increase the production of isopentenol. Candidates were selected based on three specific characteristics: (1) the flux through all NADPH producing reactions was maximized; (2) the flux through the pentose-phosphate pathway was maximized, which is a characteristic of higher-producing strains; (3) the total flux through all NADPH consuming reactions was minimized. Genes in bold were experimentally tested and the gene in bold and underlined, ydbK, increases specific production of isopentenol. Genes in italics are transporters. The list of genes and their respective biological roles are displayed in Supplementary Table S3.

Fitness characteristic	Gene knock-outs	Gene names
maximized flux through NADPH producing reactions	*b0197*, b0198, *b0199*, *b4238*	*metQ*, metL, *metN*, nrdD
maximized flux through the pentose-phosphate pathway	**b1378,** *b0197*, b0198, *b0199*, *b4238*	**ydbk**, *metQ*, metL, *metN*, nrdD
minimized flux through NADPH consuming reactions	**b4209**, **b1748,** b1747, *b2501*, b4468	**ytfE, astC,** astA, ppk, glcE

Figure1

Click here to download Figure Fig1.tif

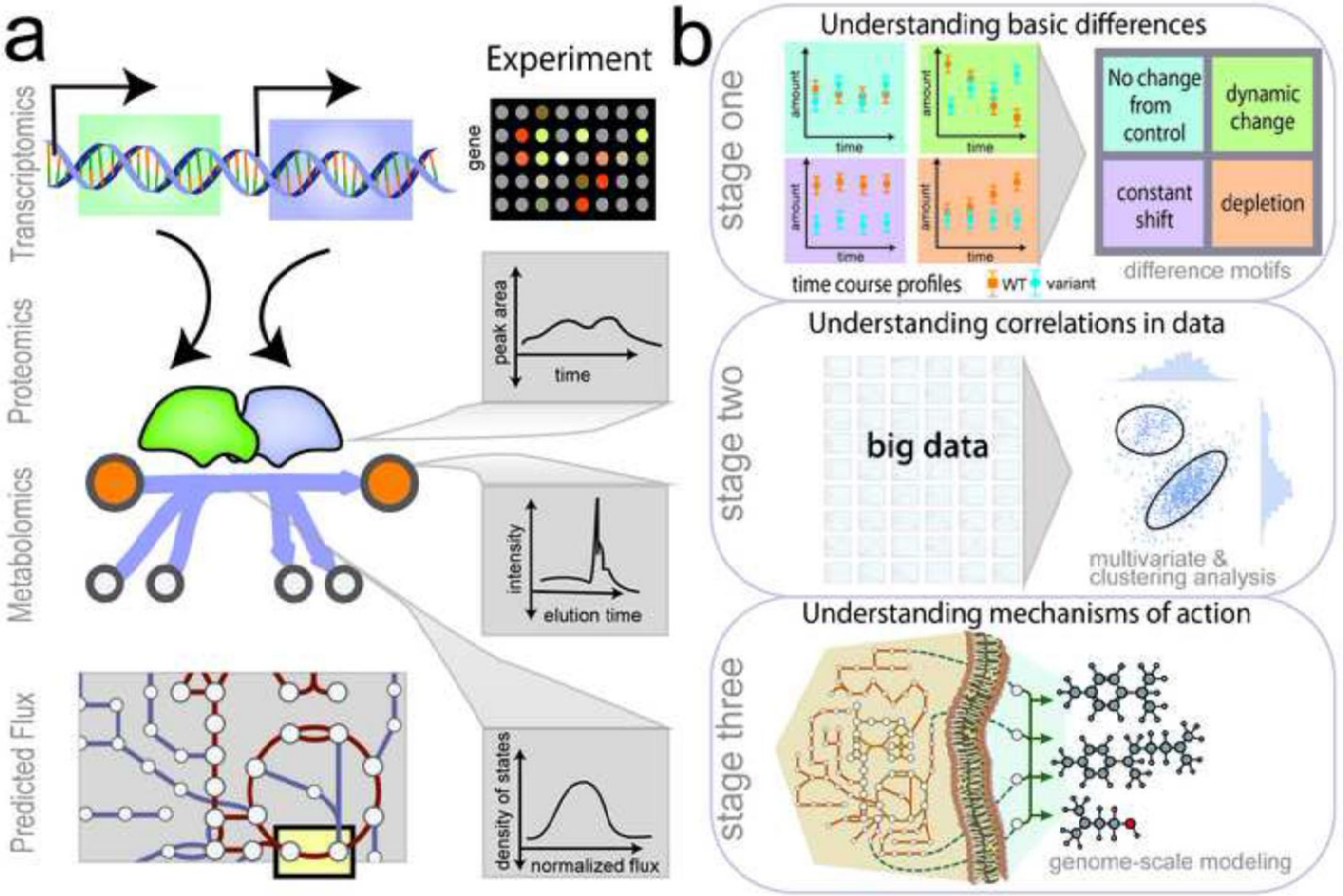


Figure2

Click here to download Figure Fig2.tif

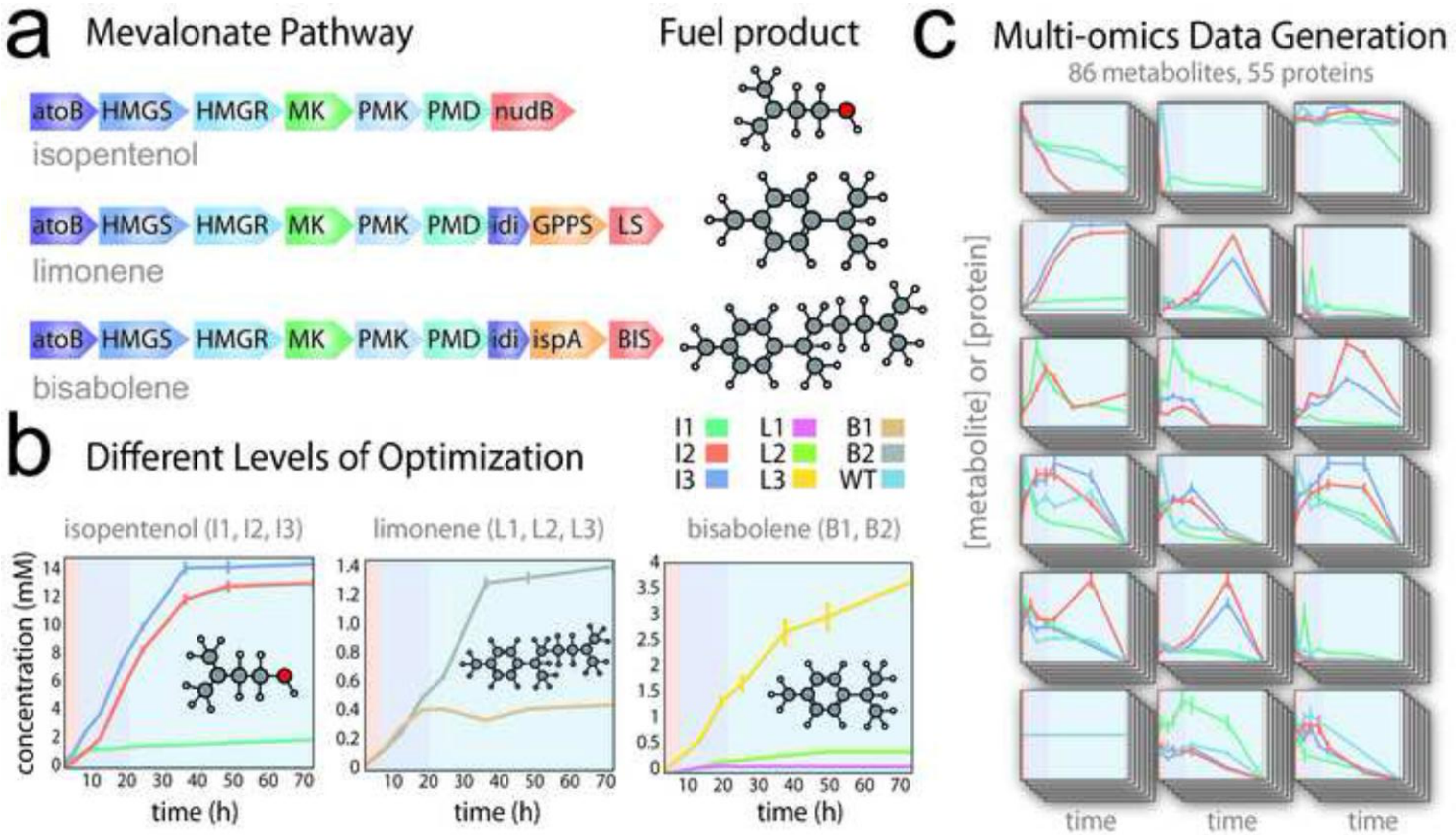
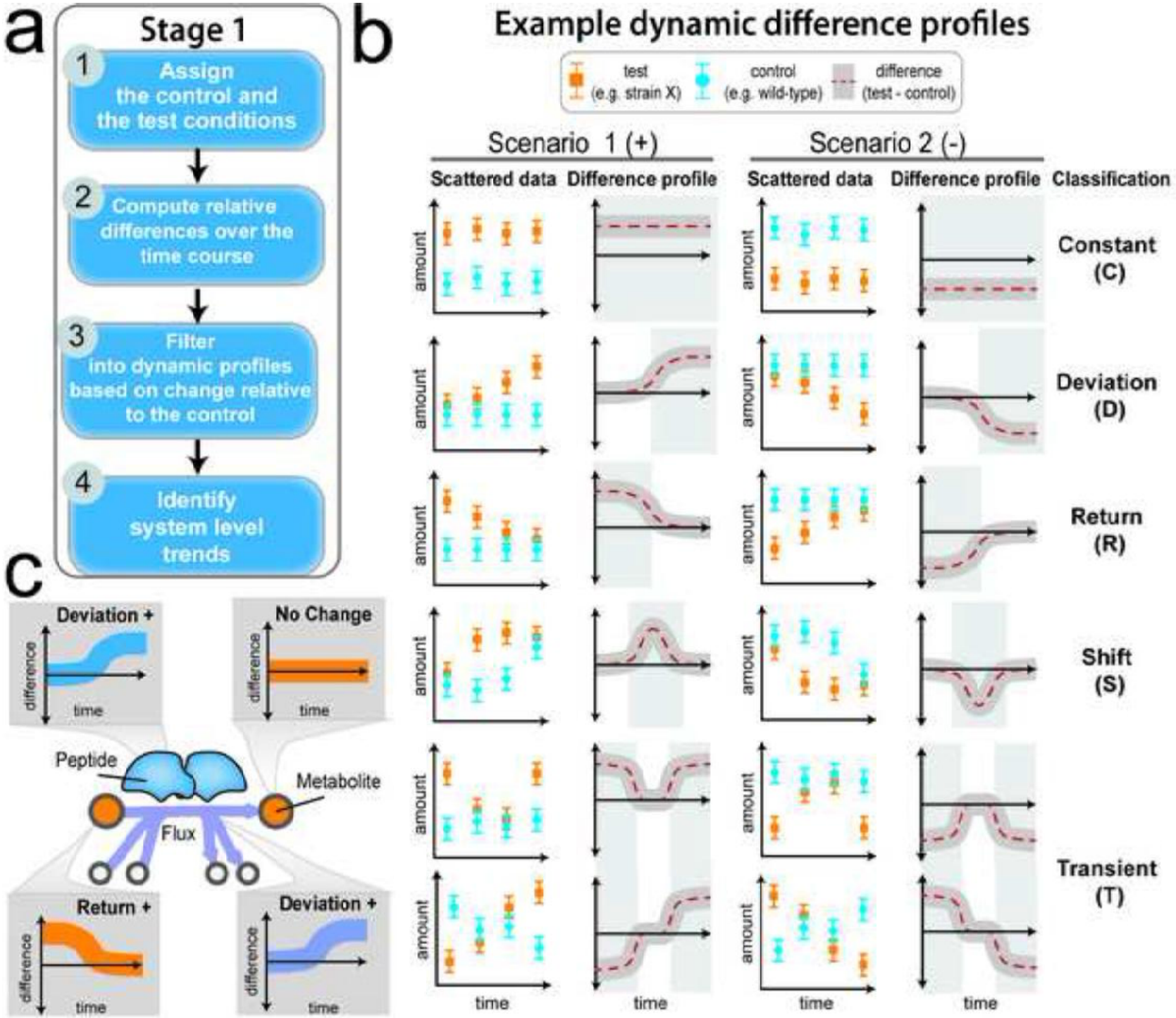


Figure3

Click here to download Figure Fig3.tif



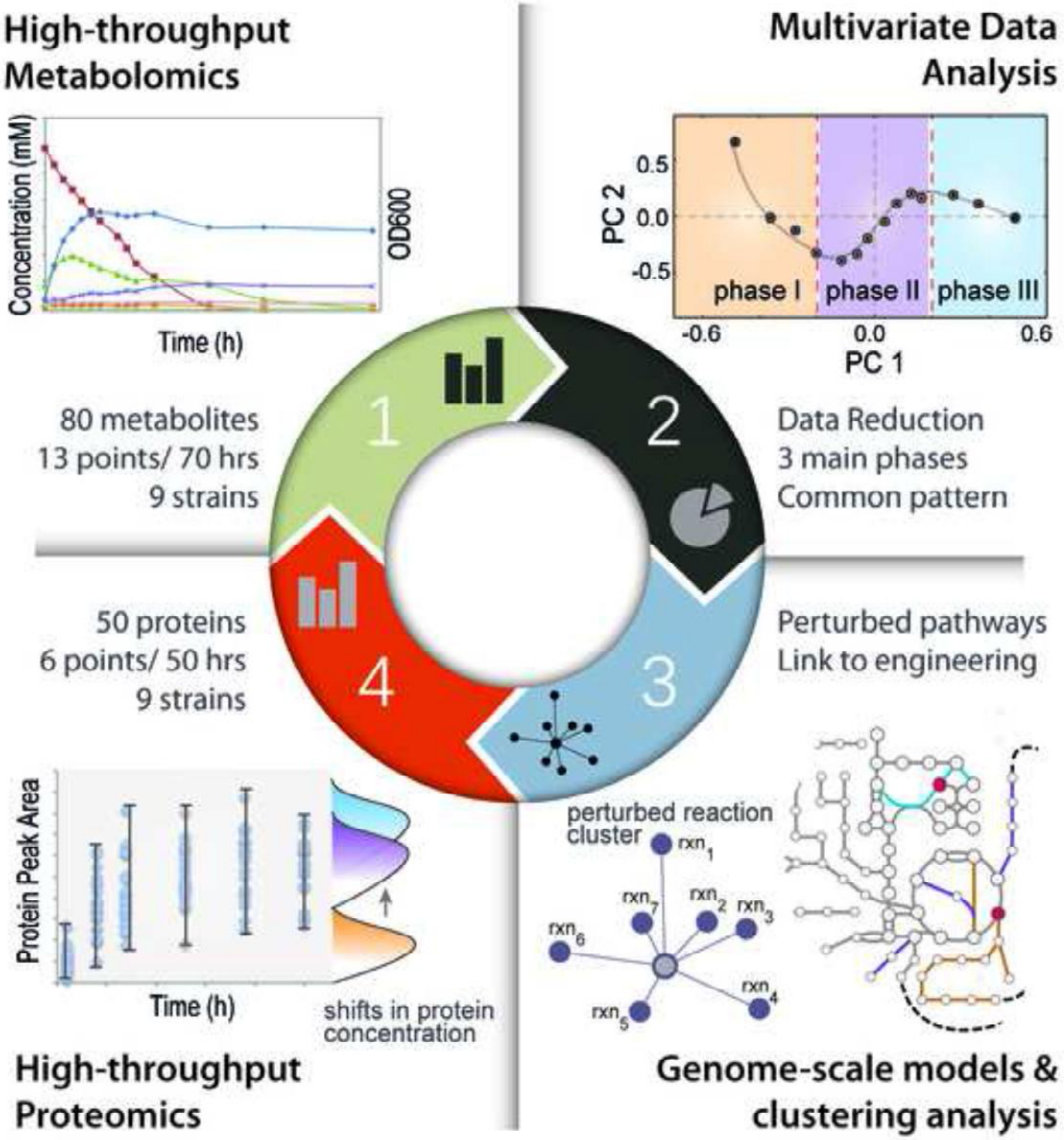


Figure5

Click here to download Figure Fig5.tif

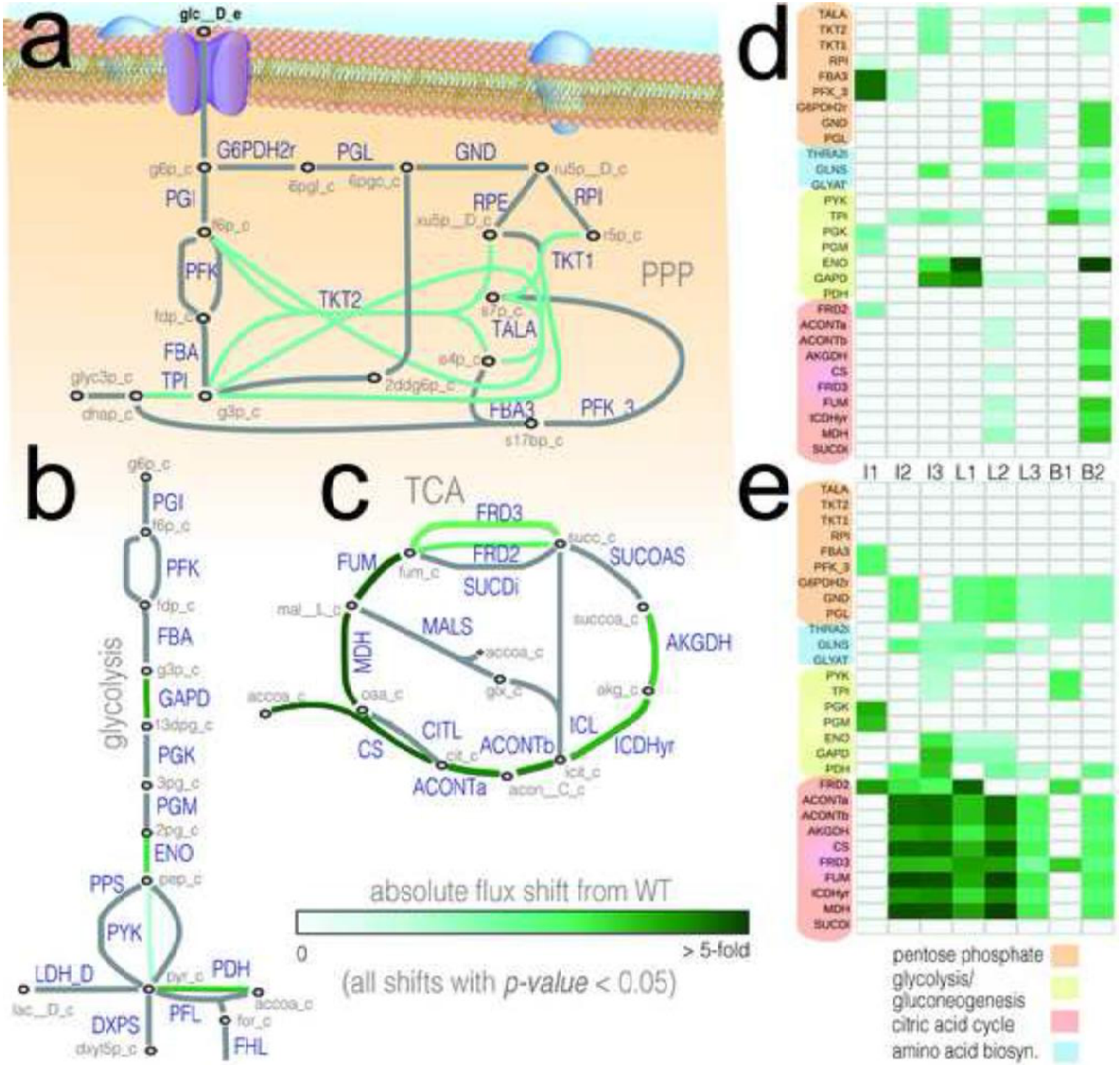


Figure6

Click here to download Figure Fig6.tif

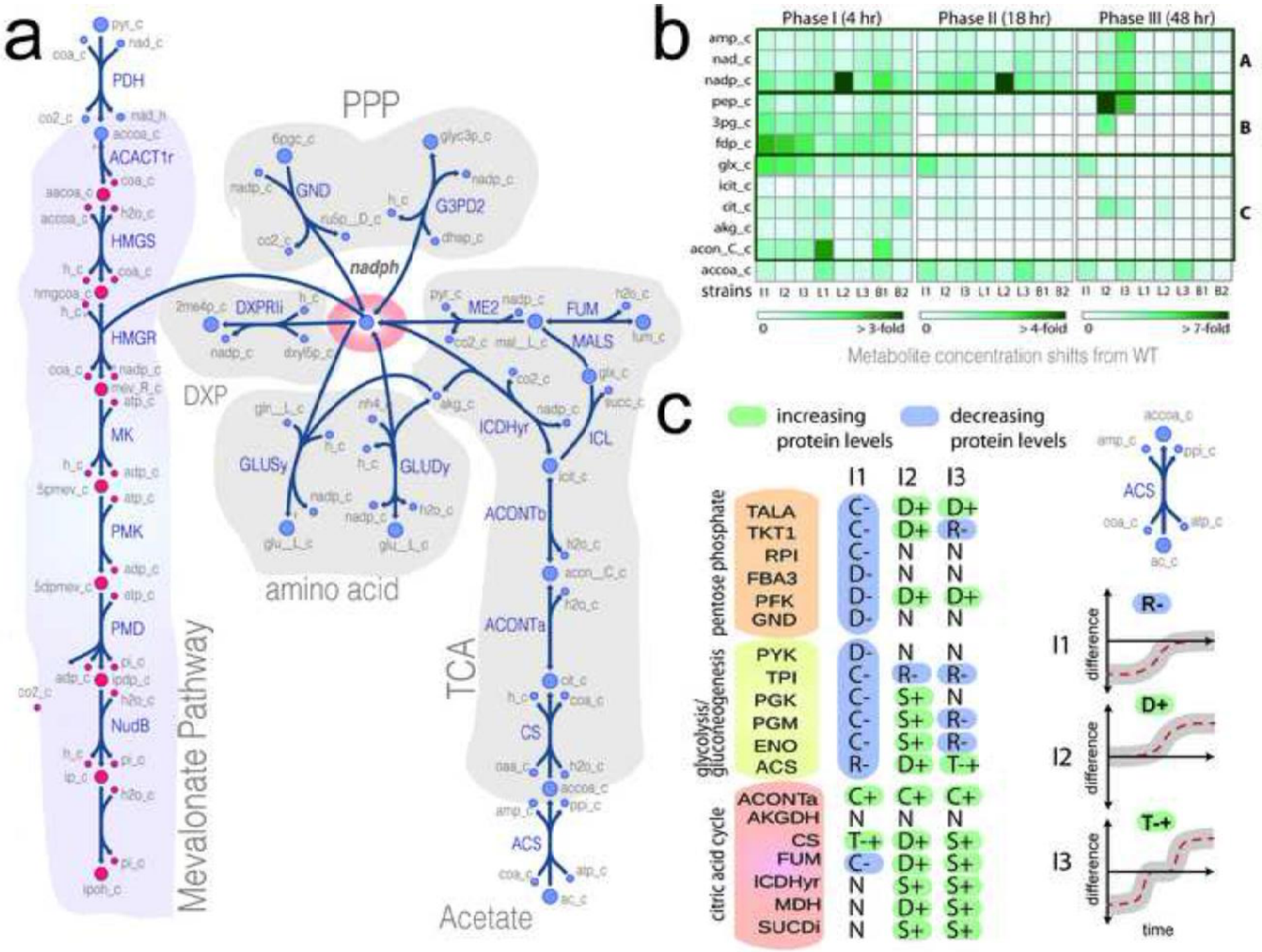
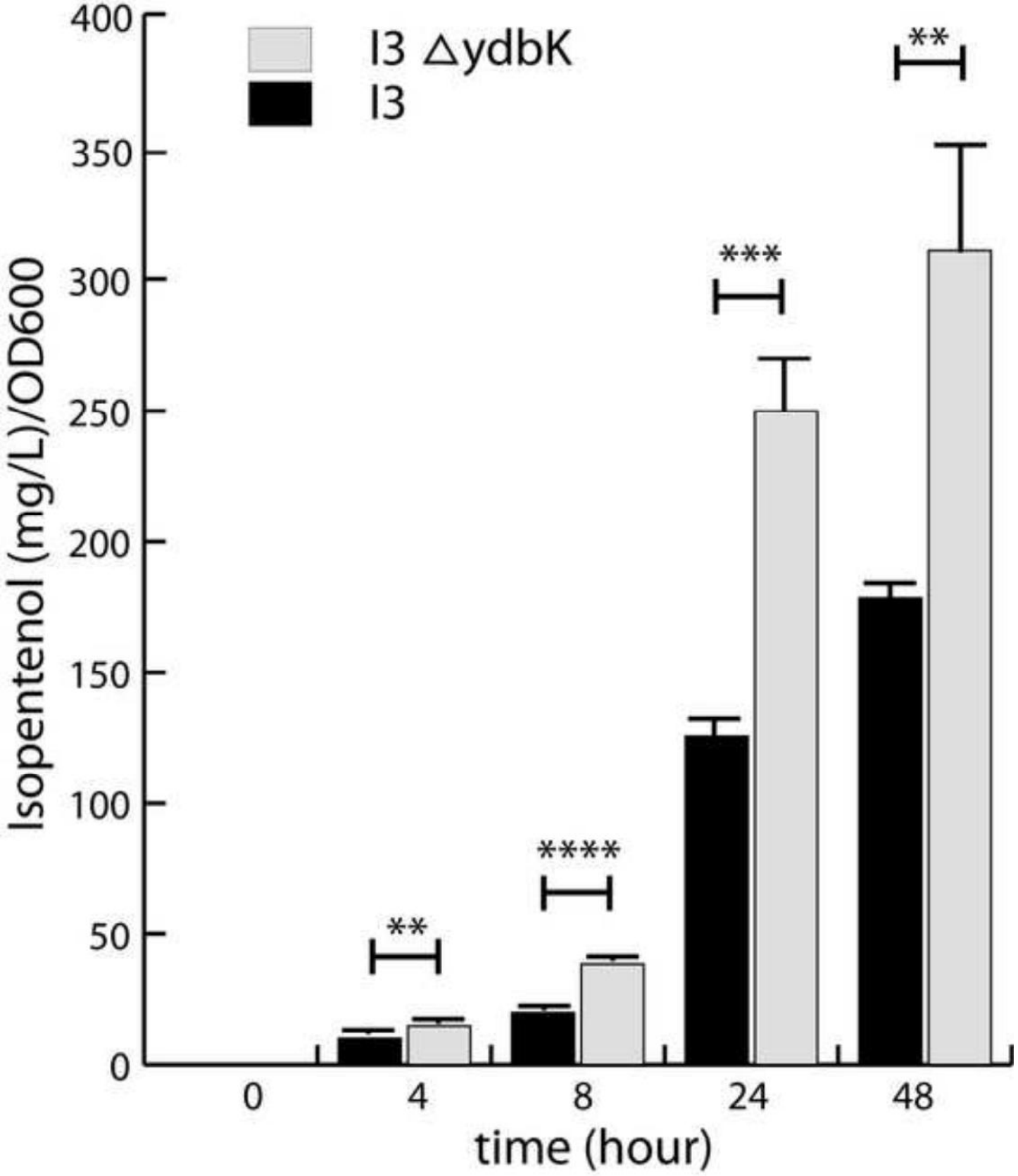


Figure7

Click here to download Figure Fig7.tif



Supplemental Information

Characterizing strain variation in engineered *E. coli* using a multi-omics based workflow

Elizabeth Brunk[a,b,c,†], Kevin W. George[a,c,d,†], Jorge Alonso-Gutierrez[a,c], Mitchell Thompson[a,e], Edward Baidoo[a,c], George Wang[a,c], Christopher J. Petzold[a,c] Douglas McCloskey[b], Jonathan Monk[b], Laurence Yang[b], Edward J. O'Brien[b], Tanveer S. Batth[a], Hector Garcia Martin[a,c], Adam Feist[b,c], Paul D. Adams[a,g], Jay D. Keasling[a,c,f,h,i], Bernhard O. Palsson[b,f,*], Taek Soon Lee[a,c,*]

[a] Joint Bioenergy Institute (JBEI), 5885 Hollis Street, Emeryville, CA 94608, USA

[b] Department of Bioengineering, University of California San Diego CA 92093, USA

[c] Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[d] Current address: Amyris, 5885 Hollis Street, Emeryville CA 94608, USA

[e] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

[f] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Horsholm, Denmark

[g] Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[h] Department of Chemical & Biomolecular Engineering, University of California, Berkeley, CA 94720, USA

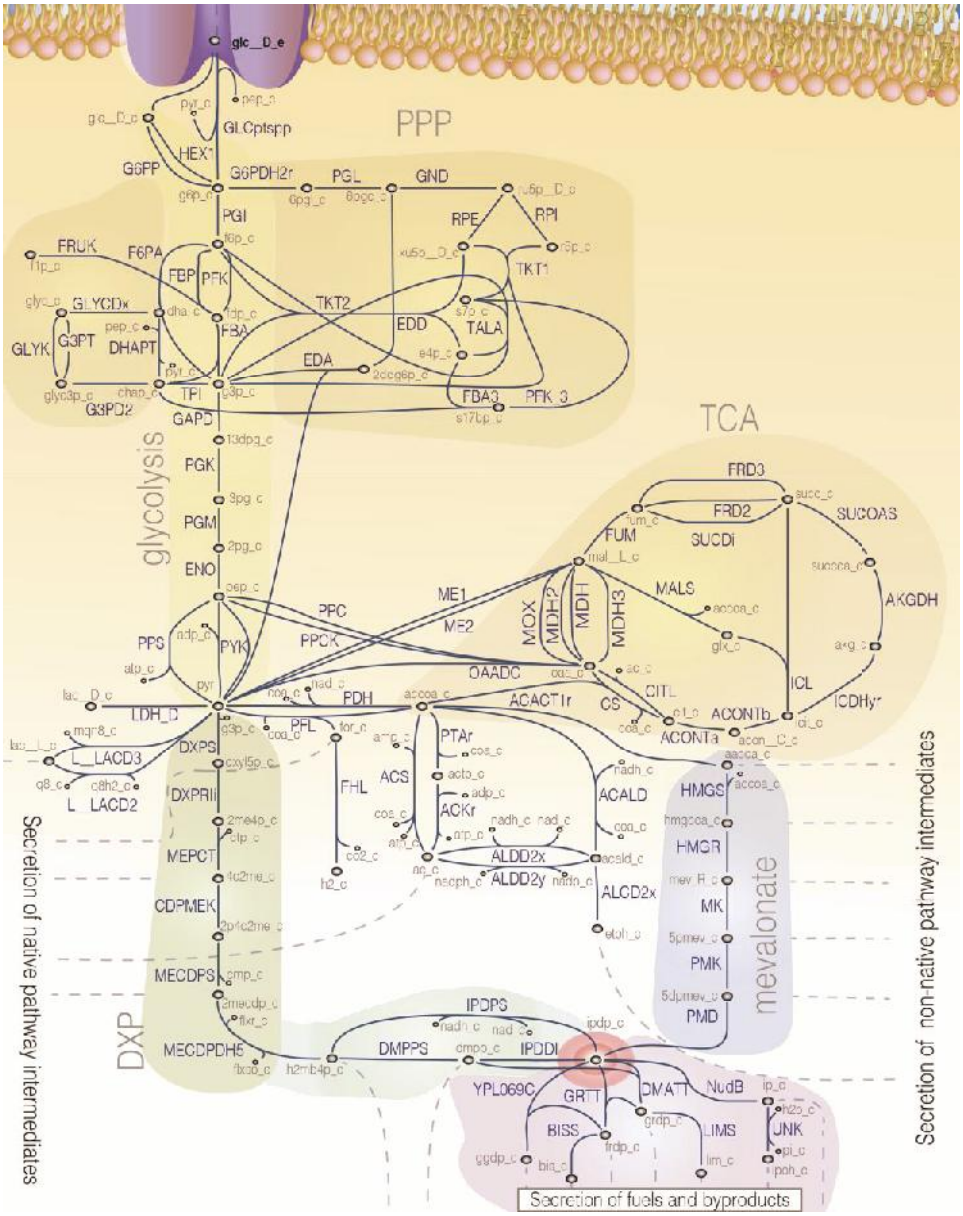
[i] Department of Bioengineering, University of California, Berkeley, CA 94720, USA

†Authors contributed equally

*Correspondence: T.S.L (tslee@lbl.gov) and B.O.P (palsson@eng.ucsd.edu)

Table of Contents

Figure S1. Core metabolic network of *E. coli* model together with the heterologous pathway
This figure depicts the core metabolic network of *E. coli* and the integration of the heterologous non-native mevalonate pathway (in blue). Shown in green is the native isoprenoid pathway in *E. coli* (DXP pathway). In cyan is the part of the network where these two pathways intersect, shown by the red sphere, highlighting the metabolite isopentenol diphosphate (ipdp_c). In purple are the reactions downstream of both the native and non-native pathways, which convert the pathway intermediates to fuel products, such as bisabolene (bis_c), limonene (lim_c) and isopentenol (ipoh_c).



Experimental Results

Pathway organization, strain selection, and multi-omics data generation

Table S1. Strains and plasmids used in this study.

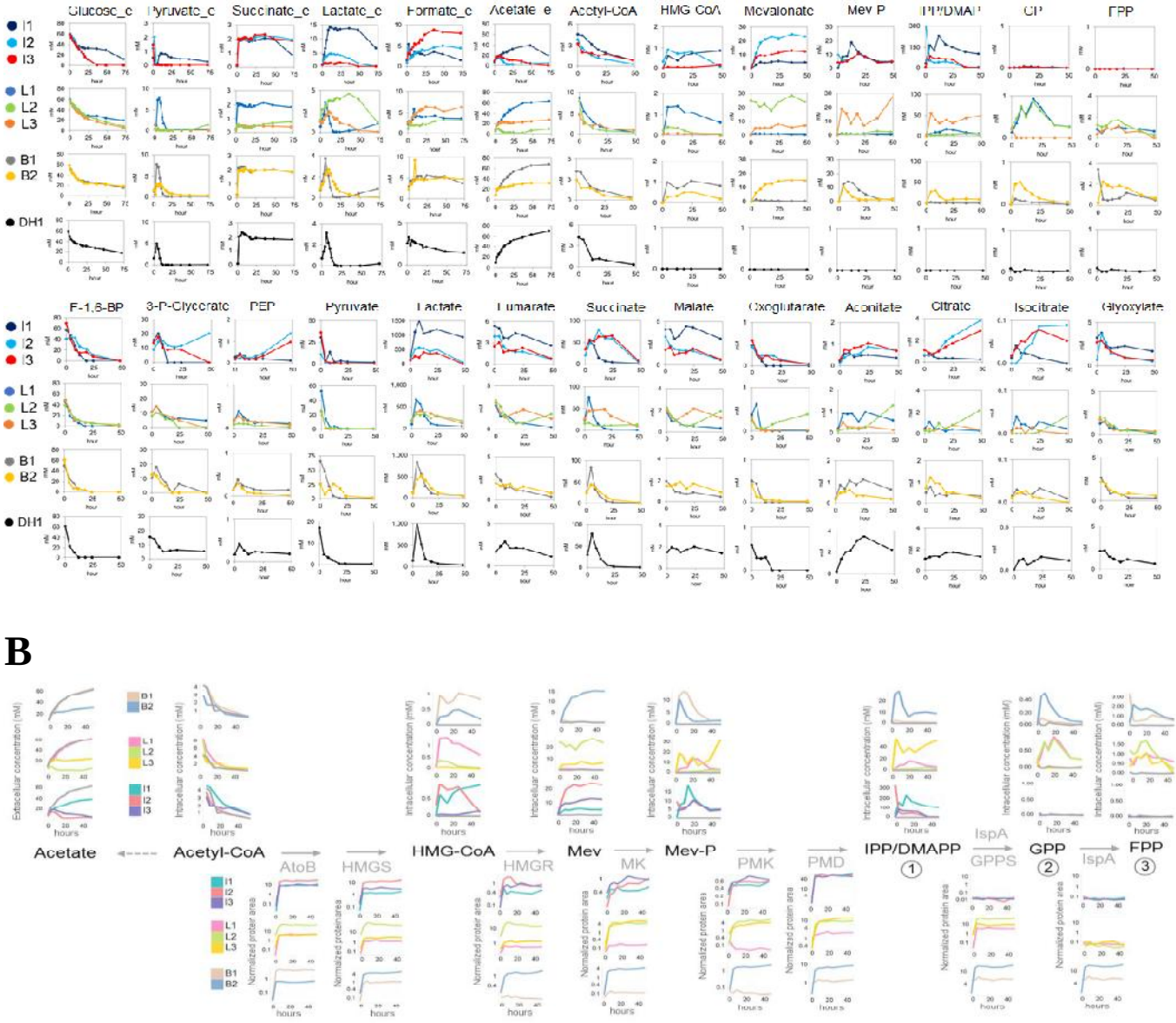
Previous optimization of the pathway plasmids for each biofuel has included codon optimization of selected genes, the insertion of supplemental promoters (e.g., a "trc" promoter) to divide the pathway into separate operons, and altered operon gene order (e.g., *PMK-MK* rather than *MK-PMK*). The primary result of these optimizations is the altered expression of key pathway proteins (see Figure S2), which in turn influences product yield and numerous aspects of host metabolism. A key focus in the optimization process has been the modulation of protein expression in the "top" portion of the mevalonate pathway (i.e., *atoB, HMGS,* and *HMGR, aka* "MevT") since the activities of these enzymes dictate flux to mevalonate and ultimately modulate downstream flux to the desired fuel product. Above, "MevTo" refers to the "original", non-optimized versions of *HMGS* and *HMGR*, "MevTco" refers to codon-optimized *HMGS* and *HMGR*, and "MevTsa" refers to *HMGS* and *HMGR* that is derived from *Staphylococcus aureus.* For each fuel product, we choose strains with various "levels" of optimization, while ensuring that a non-optimized variant (e.g., I1, L1, B1) was included for a baseline comparison. Please see the references provided in the table for additional detail. References were taken from ((George et al. 2014; Alonso-Gutierrez et al. 2013; Chubukov et al. 2015; Peralta-Yahya et al. 2011; Hanahan 1983))

Plasmids	Description	Reference
JPUB_006200	pBbA5c-MevTo-MK-PMK	George et al., 2014
JPUB_006210	pBbA5c-MevTco-PMK-MK	George et al., 2014
JPUB_004937	pBbA5c-MevTsa-MK-PMK	George et al., 2014
JPUB_004938	pTrc99A-nudB-PMD	George et al., 2014
JPUB_004921	pBbA5c-MevTo-MK-PMK-PMD-idi	Alonso-Gutierrez et al., 2013
JPUB_002471	pBbE1a-GPPS-LS-atoB-HMGSsa-HMGRsa-trc-MK-PMK-PMD-idi	Alonso-Gutierrez et al., 2015
JPUB_002470	pBbA5c-atoB-HMGSsa-HMGRsa-trc-MK-PMK-PMD-idi-trc-GPPS-LS	Alonso-Gutierrez et al., 2015
JPUB_002464	pTrc99A-GPPS-LS	Alonso-Gutierrez et al., 2015
JPUB_002473	pTrc99A-LS	Alonso-Gutierrez et al., 2015
JBx_000323	pBbA5c-MevTo-MK-PMK-PMD-idi-ispA	Peralta-Yahya et al., 2011
JPUB_002460	pBbA5c-MevTco-trc-MK-PMK-PMD-idi-ispA	Peralta-Yahya et al., 2011
JPUB_002466	pTrc99A-BIS	Peralta-Yahya et al., 2011

Strains	Description	Reference
I1	JPUB_006200 + JPUB_004938	George et al., 2014
I2	JPUB_006210 + JPUB_004938	George et al., 2014
I3	JPUB_004937 + JPUB_004938	George et al., 2014
L1	JPUB_004921 + JPUB_002464	Alonso-Gutierrez et al., 2015
L2	JPUB_002471	Alonso-Gutierrez et al., 2015
L3	JPUB_002470 + JPUB_002473	Alonso-Gutierrez et al., 2015
B1	JBx_000323 + JPUB_002466	Peralta-Yahya et al., 2011
B2	JPUB_002460 + JPUB_002466	Peralta-Yahya et al., 2011
DH1		Hanahan 1983

Figure S2 Representative set of metabolomics data and aggregate mevalonate pathway metabolomics and proteomics

A. Time course profiles of selected metabolites characterized by pathway. **B**. Multi-omics view of mevalonate pathway metabolomics and proteomics.



Representative data from the fermentation time-course is shown above. The aggregate dataset is available as a separate file. As described in the main text, strains tended to "cluster" based on optimization level rather than chose fuel target. Indeed, the measured profiles of a variety of metabolites in strains I1, L1, and B1 mirrored WT. Though still more similar to WT than optimized strains, strain I1 deviated from WT more than L1 or B1, potentially due to the accumulation of IPP, which is known to be

toxic (George et al., 2014). Changes pathway protein expression in optimized strains resulted in large changes in nearly every measured metabolite, including secreted organic acids and intracellular products of central carbon metabolism. Strains with high HMGS expression accumulated higher steady state levels of mevalonate and secreted less acetate than those with low HMGS. Careful analysis of pathway metabolomics and proteomics supports the notion that a balance between HMGS and MK expression directs flux towards IPP and downstream products. Strains with weak HMGS expression maintained low steady-state concentrations of mevalonate, but achieved high flux to downstream products (e.g. strain I1). This relationship likely stems from the substrate inhibition of MK by high concentrations of mevalonate. The importance of pathway balance is illustrated most dramatically by strain L2, where pathway genes were on a high copy plasmid. "Top" portion pathway genes were particularly enriched in strain L2: levels of AtoB, HMGS, and HMGR were more than 10-fold higher than L1 and L3 (Figure S2 (B)). Due to this enrichment, strain L2 accumulated high levels of intracellular and extracellular mevalonate and secreted the least acetate of any engineered strain. Despite strong pathway "pull" and rapid flux through the top portion of the pathway, flux to limonene was severely reduced. Given the poor kinetics of each terminal enzyme (i.e. NudB, Limonene synthase, Bisabolene synthase), we expected to observe the accumulation of IPP, GPP, or FPP depending on the biofuel target. While IPP did indeed accumulate to high levels in isopentenol strains (i.e. I1 – I3), levels of GPP and FPP were surprisingly low in limonene and bisabolene producers, perhaps indicative of GPP and FPP depletion to make quinones and other essential compounds.

Computational Results

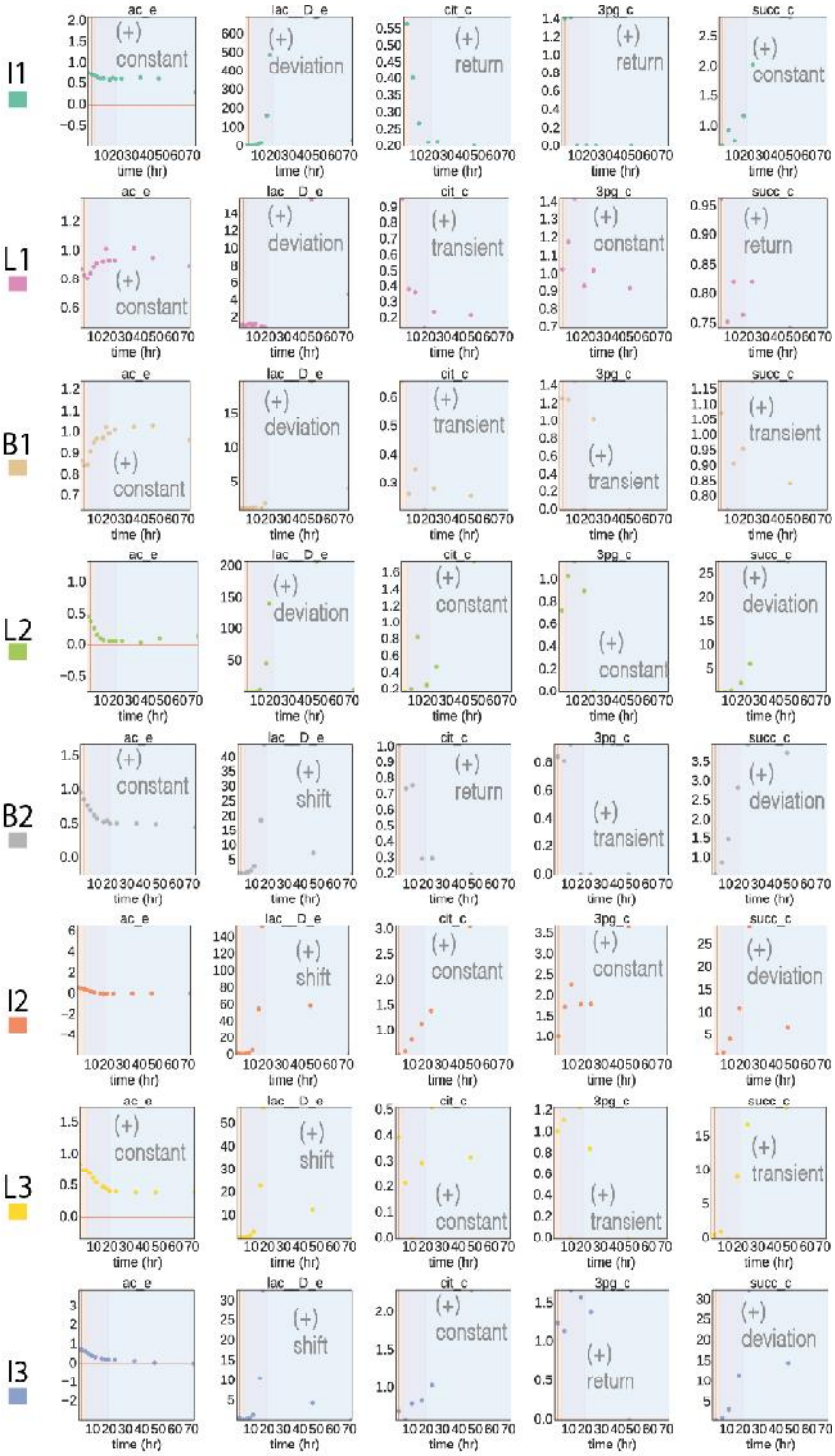
The three-stage iPython notebook series files can be found at the following links:
Stage one:
https://github.com/ebrunk/Strain_characterization_workflow/blob/master/ipython_notebook/Stage_one_Dynamic_Differences.ipynb
Stage two:
https://github.com/ebrunk/Strain_characterization_workflow/blob/master/ipython_notebook/Stage_two_multivariate.ipynb
Stage three:
https://github.com/ebrunk/Strain_characterization_workflow/blob/master/ipython_notebook/Stage_three_GEM.ipynb

Stage one: Integrate multi-omics data and profile the batch fermentation dynamics

For metabolites or peptides that did not have a triplicate measurement, we estimated the variance using the average variance for all metabolites or peptides measured. The standard deviation was scaled to the mean of the measured value according to the calculated %RSD for either triplicate measurements or the average of all measurements as stated above. The standard deviation was scaled using the following formula: standard deviation = 100 * %RSD * measured value.
See iPython notebook entitled "Stage_one_Dynamic_Differences.ipynb" for detailed scripts.

Figure S3. Difference profiles for various organic acids

Increased intracellular concentrations of glutamate, lysine and tyrosine were observed in the top-producing isopentenol strains; increases in intracellular arginine and phenylalanine concentrations were observed in all isopentenol strains as well as in the top-producing bisabolene strain; and increases in intracellular serine concentrations were observed in all strains, but most significantly in the top-producing isopentenol strains.



The process we took to construct the dynamic difference profiles was the following:

(1) subtract metabolite and protein concentrations or normalized peak areas, respectively, of the engineered strain from that of the WT strain (based on time point)

(2) For each time point during batch fermentation, track whether the difference measurement is greater than zero or less than zero, indicating a change in the engineered strain from WT

(3) Bin the clustering from (2) into several groups to establish common motifs: (i) difference measurement is constantly at zero (no change); (ii) difference measurement starts off at zero but increases (+) or decreases (-) at a later time point (shift/deviation) and remains shifted over WT; (iii) difference measurement does not start off at zero (+/-), but does return to zero at a later time point (return); (iv) difference measurement undergoes dynamic changes throughout the time series (cross).

(4) Once the clusters have been formed, analysis of engineered strains show similarities and differences based on (i) fuel product and/or (ii) pathway optimization level.

These six dynamic difference profiles are used to identify global patterns in the data, indicating whether the test condition: (i) matches the control throughout the time course (no change); (ii) is shifted above or below the control throughout the time course (constant (+/-)); (iii) is shifted above or below the control at the end of the time course (deviation (+/-)); (iv) is shifted above or below the control at the beginning of the time course (return (+/-)); (v) is shifted above or below the control at one point during the time course, but matches the control at the beginning and end of the time course (shift (+/-)); or (vi) is transiently shifting, or oscillating, both above and below the control at multiple time points (transient ++/--, -+/-+). As discussed in this section and the sections below, this analysis is applied to both metabolomics and proteomics data.

Stage two: Identify correlations in key metabolic fingerprints that distinguish strain behavior

Singular Value Decomposition (SVD) and principal component analysis (PCA) are multivariate analysis techniques that have been successfully applied in the reduction of highly dimensional data sets to find biological meaning. These approaches are generally used to examine the relationship among a set of *p* correlated variables. The first step in this analysis is organize all metabolomic and proteomic data into a matrix with each column containing the values of a different property, such as metabolite concentration, and each row a given the time that measurement was taken. Performing SVD and PCA in the context of analyzing metabolomic, proteomic and other datasets can provide answers to different questions.

The eigenvalues, λ_i of the matrix represent the variances (the degree of correlated change in the data set) associated with each new axis formed as a result of performing SVD. The eigenvector with the largest eigenvalue is called the first principal component (in the case of PCA analysis) and the singular vector with the second largest eigenvalue is the second principal component. The coefficients of an eigenvector indicate the contribution of the original variables to the vector and are referred to in the text as factor loadings. The new coordinates, Y, are called scores. If the variance along some of the axes is very small, then it can be ignored and the data can be represented in less than *p* dimensions. If the properties are completely independent, such that there is very little correlation between measurements, then all the property axes are needed to describe the dataset in its entirety. On the other extreme, if all the measurements are perfectly correlated, then a minimal subset of axes are needed to fully describe the data

set. Moreover, for a highly correlated set, given the value of one property, the values of other properties are also known. That is, the intrinsic dimensionality of the data set is now 1 instead of *p*, thus a huge reduction of the dimensionality takes place. Normally, the situation is somewhere in between these two extremes.

Overall, exo- and endo-metabolomics time series datasets were pre-processed, normalized and subjected to SVD analysis. To align all metabolite measurements to the same timepoints, we imputed "missing" timepoints with the average concentration between immediately adjacent timepoints. Subsequently, each metabolite concentration was mean-centered and standardized (i.e., divided by the standard deviation) across timepoints. This data normalization allowed us to compare the variation of metabolites even when their absolute magnitudes were different. We then applied SVD to the normalized data using the Numpy function, svd(). The three time phases are robust against the choice of data preprocessing (i.e., normalized or not) for all strains, although strains L3, B2, and DH1 show a potentially less distinct phase 2.

We identified the major bioprocess time phases based on the first two eigenvectors, as described earlier. Finally, we tested the robustness of our chosen time phases against the choice of data normalization, data imputation, and whether exometabolites were used exclusively versus the use of both exo- and endo-metabolites. We tested normalized (i.e., mean-centered and standardized) and imputed, normalized but not imputed, mean-centered but not standardized nor imputed, mean-centered and imputed but not normalized. Our chosen time phases were similar in all cases.

For the majority of strains, the first principal component represents 40% of the data, the second principal component represents 20% and the third principal component represents 20%, which sums to 80% of the variance explained in the first three principal components. Since the data is normalized, we expect that additional axes are required to explain the variance in the dataset. However, these findings demonstrate that the dataset is indeed highly correlated.

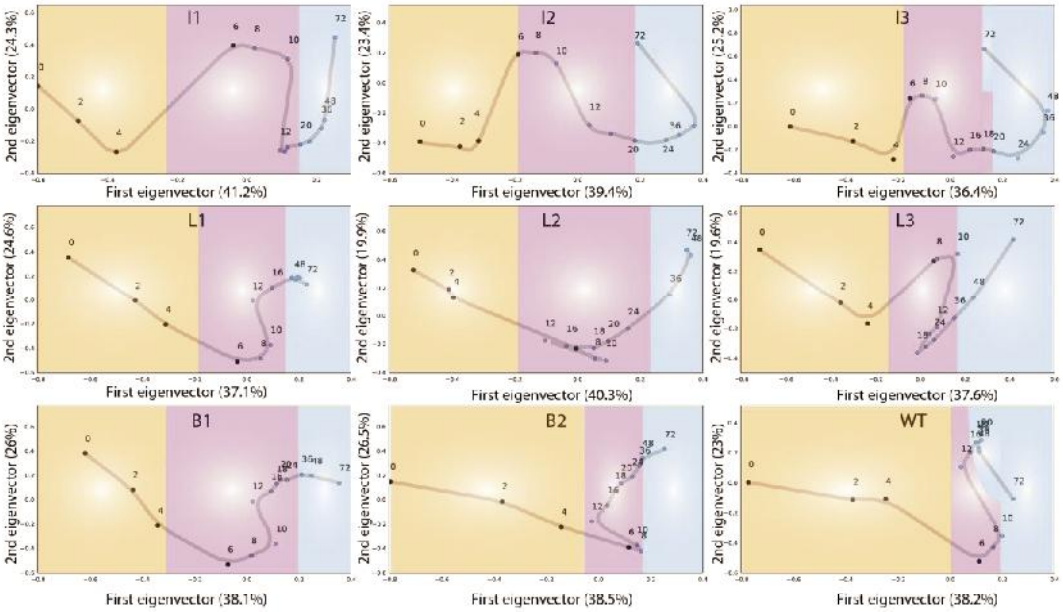
We also performed PCA on the proteomics data sets to understand how correlated proteome allocations are over 72 hours for different strains versus that of wild-type DH1. In general, PCA on normalized peak areas (normalized to BSA) indicates that the data is highly correlated (for most strains the first singular vector explains 50% or more of the variation) but we do not see a similar three state behavior as seen for the metabolomics data sets. For strains I2, I3, L2, L3 and B2, many of the proteins that have the highest coefficients on the first singular vector are indeed the mevalonate pathway proteins (as they are overexpressed).

Native *E. coli* proteins that also have large coefficients on the first singular vector are typically directly interacting with the mevalonate pathway (e.g. Acetyl-coa acetyltransferase, ACACT1r). This is to be expected, physiologically, as acetyl-coa acetyltransferase is expected to play a large role in redirecting carbon flux to the mevalonate pathway through the acetyl-coA node. In general, we do not expect to see the same degree of variation (e.g. the three-state behavior) in the native *E. coli*. As induction takes place at an OD600 of 1.5-2.0 (late exponential growth phase), most of the proteome in the cell has been established. See iPython notebook "Stage_two_multivariate.ipynb" for more information on this stage.

Figure S4. Singular Value Decomposition of Metabolomics Data

A. Singular Value Decomposition of all extracellular (phenotypic) and endometabolomic data from all 8 strains. Each point in the graph is indexed by the time (in hour) that the sample was taken. Missing measurements were imputed using the mean between timepoints. **B.** The relative variance explained by eigenvectors (components) from SVD of the normalized metabolomics data.

A


I1
I2
I3
L1
L2
L3
B1
B2
WT
2nd eigenvector (24.3%)
First eigenvector (41.2%)
2nd eigenvector (23.4%)
First eigenvector (39.1%)
2nd eigenvector (25.2%)
First eigenvector (36.1%)
2nd eigenvector (24.6%)
First eigenvector (37.1%)
2nd eigenvector (19.9%)
First eigenvector (40.3%)
2nd eigenvector (19.6%)
First eigenvector (37.6%)
2nd eigenvector (26%)
First eigenvector (38.1%)
2nd eigenvector (26.5%)
First eigenvector (38.5%)
2nd eigenvector (23%)
First eigenvector (38.2%)

B

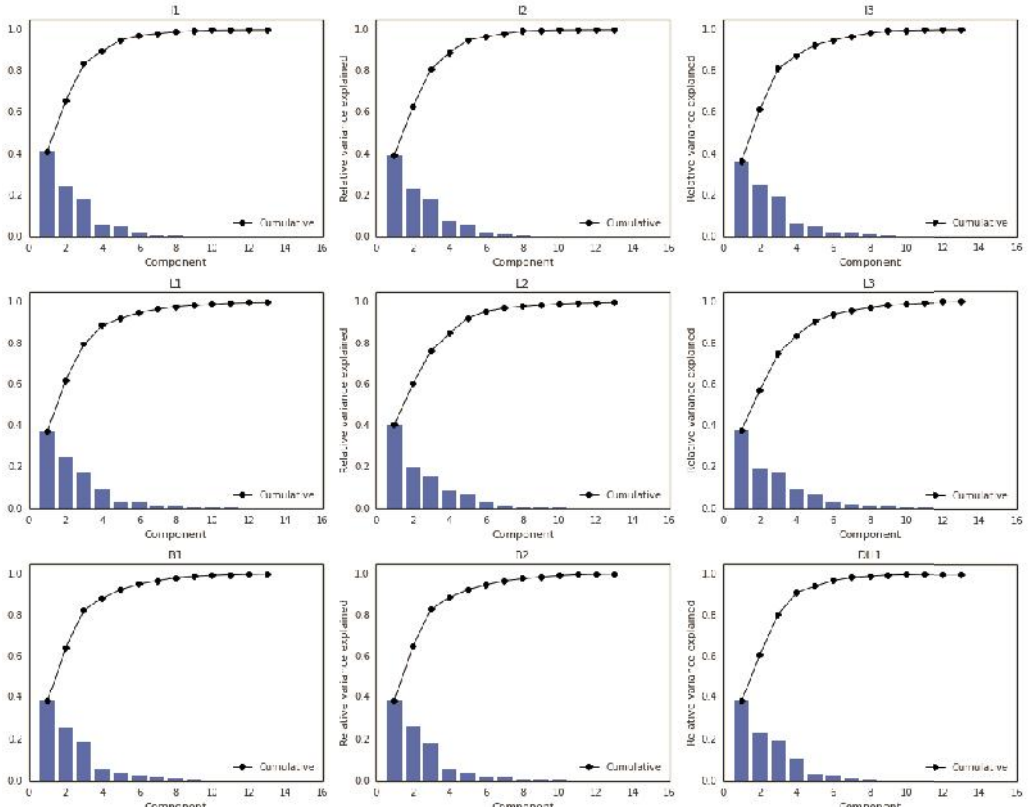
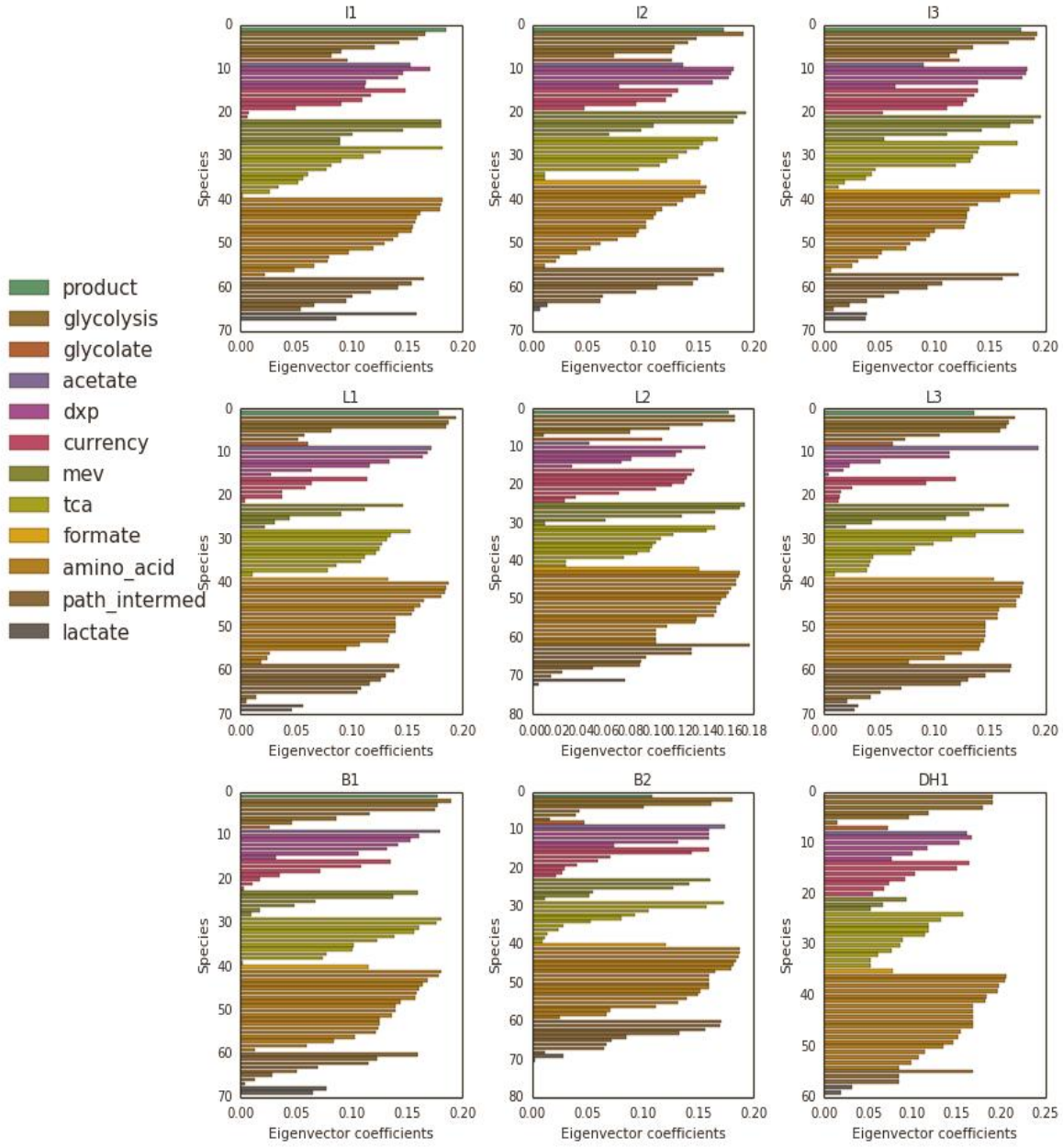

I1
I2
I3
L1
L2
L3
Component
Cumulative

Figure S5. Distinct behavior distinguishes phenotypic behavior between different fuel producers and different levels of pathway optimization

Singular Value Decomposition of all extracellular (phenotypic) and endometabolomic data from all 8 strains (I1, I2, I3, L1, L2, L3, B1, and B2), producing isopentenol, limonene and bisabolene fuel products. Larger eigen coefficients indicate metabolites that are main drivers in the variation of a particular strain's data.



Stage three: Perform genome-scale modeling to gain mechanistic insights into strain behavior

Building metabolic models for different fuel product producing strains

While multivariate analyses provide an efficient means to reducing the highly dimensional nature of high-throughput datasets, '-omics data sets can still be unwieldy and challenging to interpret owing to the unpredictable nature of metabolic circuitry. These challenges have recently been approached through the construction of large mechanistic models for various organisms, tissues and cell types (Becker and Palsson 2008; Shlomi et al. 2008; Lewis et al. 2010; Mo, Palsson, and Herrgård 2009) that serve as a context for further analysis. A workflow for generating metabolic models of biofuel-producing strains of *E. coli* consists of the following three steps. See iPython notebook entitled "Stage_three_GEM.ipynb" for more information.

Step 1. Reconstruct a metabolic network for the organism of interest from genome annotation, lists of biomolecular components and literature (Thiele and Palsson 2010).

Published metabolic models are continually being updated through many iterations of manual curation, generating hypotheses, validation through experiments and incorporation of new knowledge. Here, we use a recently updated metabolic model of *E. coli* (iJO1366).

Step 2. Identify reactions specific to the heterologous pathway of interest which will not be in the metabolic network reconstruction of the organism.

Characteristic stoichiometries for the various heterologous pathway intermediates are generally found in online databases, such as EcoCyc(Keseler et al. 2005) and Brenda(Schomburg et al. 2004; Chang et al. 2009).

Step 3. Simulation and analysis (Oberhardt, Palsson, and Papin 2009; Feist and Palsson 2008).

Once the network is accurately reconstructed and converted into an *in silico* model, it is used to generate hypotheses and to obtain insight into systems-level biological functions. This workflow was used to build three different biofuel producing models of *E. coli* metabolism (isopentenol, limonene and bisabolene). Genome-scale models use uptake and secretion rates of metabolites to constrain the flux solution space of the entire metabolic network. We apply our findings from PCA on the extracellular metabolomics to assign *pseudo*-steady states (i.e. exponential, early stationary and late stationary phases) to carry out a constraint-based modeling approach. Metabolite concentrations taken from each of the three phases were used to constrain the solution space of the flux network.

Markov-chain Monte Carlo Sampling of Flux space (MCMC)

We utilized the Artificial Centering Hit-and-Run (ACHR) Monte Carlo (MC) sampling algorithm (Thiele et al. 2005; Price, Schellenberger, and Palsson 2004) to uniformly sample the metabolic flux solution space defined by the set of constraints described above. This approach allowed biasing the sampling towards physiologically relevant parts of the solution space without imposing the requirement of strictly maximizing a predetermined objective function.

We assumed an error of 20% to set the lower and upper bounds of the constraint on uptake, secretion, thus inherently accounting for the sampling calculation sensitivity. In order to study more physiologically relevant portions of the flux space we restricted the sampling to the part of the solution space where the growth rate was at least 20% of the measured growth rate for the condition as determined by FBA and OD600 measurements. This assures that cellular growth remains an important overall objective by the *E. coli* cells even in batch cultivation conditions, but that the intracellular flux distributions may not correspond to maximum biomass production (Schuetz, Kuepfer, and Sauer 2007).

Carrying out this MC sampling procedure results in a distribution of a range of flux values for each

reaction in the metabolic network. Typically, the most likely flux state for a given reaction is represented by the mean of this distribution. Due to the overall shape of the metabolic flux solution space, most of the values in the sampled flux distributions are close to the minimally allowed growth rate (i.e. biomass production). The following sections describe the approaches that were used in the analysis of this data set. Post-processing of the flux distributions considers removal of reactions and their participating metabolites which are found to participate in intracellular loop reactions (Price, Thiele, and Palsson 2006) as they have been shown to have arbitrary flux values.

The three phases taken from the SVD analysis on the metabolomic datasets for each strain were used to constrain the solution space of the flux network. Relative levels of quantitative extracellular metabolome (EM) data were averaged across a given phase (e.g. 0 to 6 hours, 8-20 hours and 24 to 72 hours) to simulate the three different pseudo steady states for each strain. Different input constraints were used for each phase of each strain, thus the calculated solution spaces between the conditions differed based only on variations in the experimental secretion measurements.

The excreted mevalonate and DXP pathway intermediates together with the secreted organic acids were incorporated into the constraint-based framework as overflow secretion exchange fluxes to simulate the low-level production of experimentally observed excreted metabolites. The rates of secretion and uptake were approximated based on a normalization of the metabolite concentration with respect to the biomass (OD600 measurement) such that the flux is given in units of mmol/hr/gDW.

Z-score based analysis of Flux shifts from WT phenotype

We were interested in characterizing the flux shifts across various conditions, for example, a shift in time (e.g. between two phases) as well as the shift between strains (e.g. wild-type DH1 versus strain I3). As explained in the section above, the output of the MC sampling procedure is a distribution of flux for every reaction in the *E. coli* metabolic network model. Certain flux distributions are highly constrained (i.e. their distributions are not broad) whereas other flux distributions are not constrained and can vary across a wide range of flux values. Comparing changes between any two conditions must address the characteristic flux distributions to determine whether or not a shift can be considered significant.

We used a Z-score based approach based on a previously reported MC sampling based analysis (Mo, Palsson, and Herrgård 2009). To account and correct for background distribution, the Z-score was normalized by computing $\mu_{reaction,Nj}$ and $\sigma_{reaction,,Nj}$ corresponding to the mean $m_{reaction}$ and its standard deviation for 1,000 randomly generated reaction sets of size N_j. *Z*-scores for subsystems were calculated similarly by considering the set of reactions that belong to the given subsystem. Similarly, Z-scores were computed on the metabolite level by computing a normalized Z-score based on the number of reactions a metabolite is involved in. Z-scores for reactions that have a value greater than 1.74 indicate that they are significantly shifting (i.e. highly perturbed regions of metabolic space), such that the overlap between two flux distributions across two different conditions is significantly less, given a *p*value of less than 0.05%. Perturbation of subnetworks of reactions and connecting metabolites were visualized using Escher maps (http://escher.io.github).

Significantly perturbed regions of space were determined for each phase for each strain based on the Z-score methodology described above. The most significantly perturbed regions were compared across strains and those that were shared among the majority of strains are referred to as “global shifts.” A number of “unique shifts” have also been detected as a result of this part of the workflow, which are potentially interesting mechanisms that arise from the engineering of the strains.

Highly perturbed nodes in glycolysis/gluconeogenesis include phosphofructokinase (PFK), fructose-bisphosphate aldolase (FBA), glyceraldehyde-3-phosphate dehydrogenase (GAPD) and pyruvate dehydrogenase (PDH), which is linked directly to the synthesis of acetyl-coA. Highly perturbed nodes in pentose-phosphate pathway include triose-phosphate isomerase (TPI), sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase (FBA3), transaldolase (TALA) and transketolase (TKT1, TKT2). Finally, in the citric acid cycle, the most perturbed reactions are alpha-ketogluterate dehydrogenase (AKGDH), aconitase a and b (ACONTa, ACONTb), citrate synthase (CS) and isocitrate dehydrogenase (ICDHyr). Other highly perturbed nodes in glycolysis/gluconeogenesis include phosphofructokinase (PFK), fructose-bisphosphate aldolase (FBA), glyceraldehyde-3-phosphate dehydrogenase (GAPD) and pyruvate dehydrogenase (PDH), which is linked directly to the synthesis of acetyl-coA (see Supplementary Figures 19 and 20). To test the sensitivity of the results to the sampling times, separate Monte Carlo samples were run for each of the strains and convergence was confirmed. We also tested the sensitivity of the results to the relative magnitude of the extracellular metabolite secretion rates by performing the sampling for isopentenol-producing strains using different uptake/secretion data (Supplementary File “Heterologous Metabolomics Data Analysis.xls”).

Figure S6. Reactions that are significantly perturbed as a result of Engineering

A. Computed reaction z scores, indicated the perturbed metabolic regions, detected by constraint-based modeling and ranked significant if z-scores are greater than 1.74. **B.** Venn diagram clusters of highly perturbed reaction nodes shared or unique in fuel product groupings.

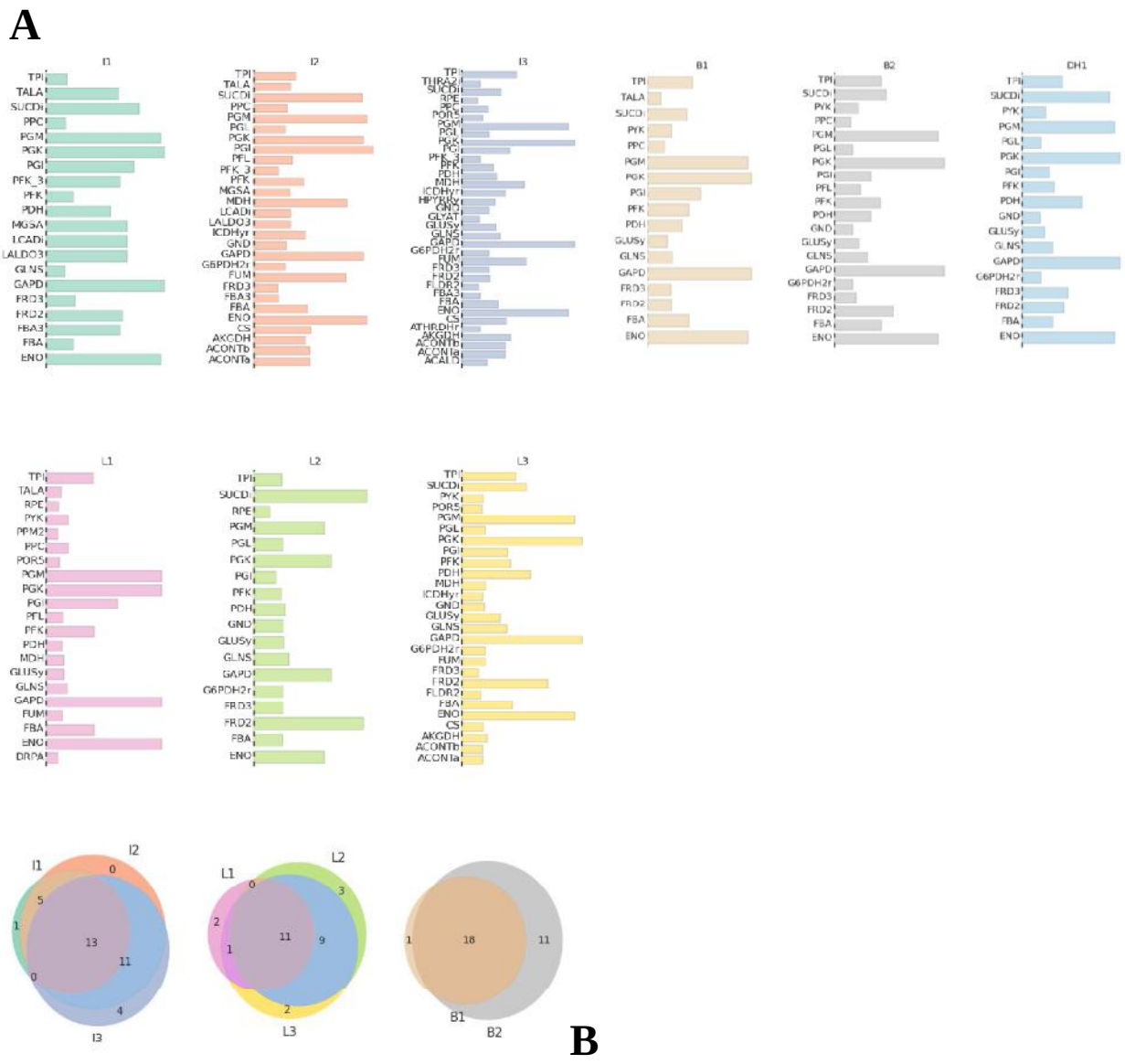
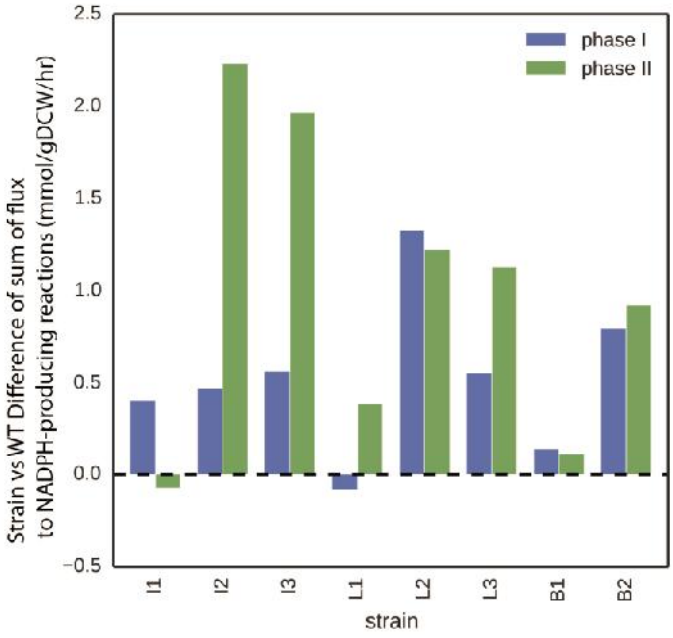


Figure S7. Sum of fluxes to NADPH-producing reactions.
Model predictions of phase I vs phase II fluxes through NADPH-producing reactions. Given is the difference of the sum of all fluxes producing NADPH in engineered versus WT strains.



A general, iterative workflow

The main text Figure 4 serves mainly as a conceptual demonstration of the detailed procedure carried out in Stages 2 and 3. In the first step (top right), we illustrate the type of data flowing into this stage of the workflow, using a typical metabolomics (e.g. organic acid) profile. The cartoon stacks of data/graphs indicate the amount of data we are dealing, since we are using high-throughput metabolomics data (86 metabolites, 13 time points and 9 strains in total). In the second step, data reduction using PCA unveils certain correlations from this highly dimensional dataset. The example discussed in the main text is the occurrence of the three phase behavior, determined from metabolite variation. In the third step, we can use knowledge of these patterns to define pseudo-steady states during the time course to carry out genome scale modeling. It is important to note that, while data reduction in phase two is important for pattern recognition and identifying correlations, reduction of network-level information can be non-informative if solutions lie in peripheral pathways in metabolism (i.e. not in central carbon metabolic pathways). The inputs into the model at this step are the metabolites that are consumed or secreted at each of the phases determined from PCA. At the end of this step, we compare all perturbed reaction fluxes due to pathway engineering and cluster them to find common links, such as cofactor usage. Finally, genome scale modeling provides insights into possible metabolic changes (e.g. endogenous flux changes) that we can compare to disparate omic data: RNAseq or, in this case, proteomics. It is important to note that genome-scale models predict optimality and, as input to the model, we use metabolite level changes (both secretion and uptake rates). Consequently, we see discrete flux changes for different levels of pathway optimization, as a direct result of an increased “pull” (due to codon optimization) through the mevalonate pathway. In this way, we don’t actually explicitly account for codon usage in the model, but we can

observe its effects (i.e. flux changes) through the higher/lower levels of measured secreted product and/or pathway intermediates that are used as inputs into the model simulation. Thus, in the fourth step, we compare changes in protein levels to changes detected in predicted fluxes (which can be linked to a protein, through a gene-protein-reaction relationship). This workflow is iterative in nature, since the knowledge gained from step 4 can feed back into the workflow and guide the focus of which changes (metabolites and/or flux along pathways) can be reconciled with changes in proteomics. In the end, we're essentially using genome scale models to model all possible states of metabolism, both in wild-type and engineered strains. In total, we model three specific products: isopentenol, limonene and bisabolene.

Model-aided prediction of engineered metabolic phenotypes is consistent with experiments

In general, we find that certain key metabolic phenotypes are consistent with the measurements. For example, flux through the mevalonate pathway, and, in particular, through HMGS and HMGR is significantly enriched for the higher producing strains (e.g. I2 and I3) versus the lower producing strains (e.g. I1). This is to be expected, as the HMGS and HMGR genes in I2 and I3 strains are codon-optimized and have been shown to produce 2-10 fold the amount of protein of I1. Also, the model is consistent with observed phenotype correlations from the second step of this workflow (raw data curation and analysis). For example, as more flux is observed through the mevalonate pathway for strains I2 and I3, we also observe an increased flux through certain reactions in the TCA cycle (e.g. citrate synthase (CS) and alpha-keto glutarate dehydrogenase (AKGDH)).

The most perturbed regions of metabolic space between phase I and phase II can be classified into five main subsystems: (i) upper glycolysis; (ii) lower glycolysis; (iii) TCA; (iv) PPP; and (v) amino acid metabolism. In the upper and lower glycolysis node, the reactions that are most significantly shifting are (TPI), glycerol-3-phosphate dehydrogenase (GAPD), and enolase (ENO). In TCA, citrate synthase (CS), alpha ketoglutarate dehydrogenase (AKGDH), aconitase (ACONTa/b) and fumerase (FUM) are detected to be significantly shifting in certain strains. In PPP, GND, TALA and TKT1/TKT2 are significantly shifting. Finally, in the amino acid biosynthetic changes we see changes in glutamate and glutamine synthetases.

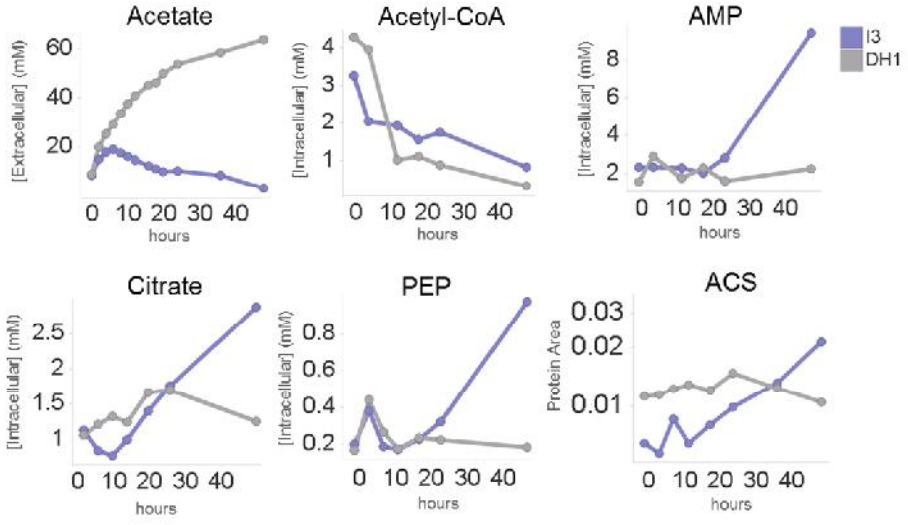
Table S2. Fold differences between WT and engineered strains in central carbon protein levels (at 48 hours)

strain	ACONTa	ACONTb	ACS	AKGDH
I1		2.307297	1.025409	1.324224
I2		1.992266	2.237486	1.494973
I3	1.012421	2.079424	3.098786	1.336769
L1	1.096099	1.235977	1.3773	1.423511
L2		1.277886	2.251004	1.598908
L3	1.044519	2.221093		1.62796
B1				1.167407
B2	1.280888	1.558578	1.842991	1.555346

strain	CS	FUM	G6PDH2r	GAPD
I1	1.413036			
I2	1.939038	1.667288	1.325948	1.204843
I3	2.114611	1.861303	1.149866	
L1	1.703191	1.208672		1.103652
L2	3.078821	1.910281	1.836256	1.14185
L3	2.372887	1.214664	1.094854	1.047619
B1	1.128843	1.029352	1.062454	1.080027
B2	2.003581	1.351933	1.160594	1.115617
strain	**GLYCL**	**ICDHyr**	**MDH**	**MDH2**
I1	1.283383	1.064527	1.576199	72.93218
I2	1.232722	1.616791	1.885506	81.65983
I3	1.328493	1.341796	1.715059	93.90099
L1	1.134186	1.463079	1.58885	56.7601
L2	1.358191	1.312792	1.252991	55.62739
L3	1.424241	1.515198	1.771878	86.59948
B1	1.091344	1.08939	1.496675	44.42694
B2	1.175677	1.263758	1.507556	48.73817
strain	**ME1**	**ME2**	**PDH**	**SUCDi**
I1	1.34884	1.100605	1.283383	1.431546
I2	1.857113	1.340254	1.232722	1.993585
I3	1.662339	1.39373	1.328493	1.391514
L1	1.789447		1.134186	1.708613
L2	2.008588	1.740199	1.358191	2.83125
L3	2.105288	1.363425	1.424241	2.054385
B1	1.483183		1.091344	1.213318
B2	2.377251	1.145139	1.175677	1.959825

Figure S8. Acetate assimilation links to TCA cycle.

Applying multi-omics to track changes in the acetate cycle between DH1 WT (gray) and strain I3.



One initially perplexing observation was the preferential "shunting" of acetate-derived acetyl-CoA into the TCA cycle rather than the mevalonate pathway, which are both possible routes for the carbon. In strain I3, acetate assimilation, which begins in phase II and continues until 48 hours post-induction, coincides with a 5-fold increase in protein levels of ACS and a 4-fold increase in the concentration of AMP, a by-product of the ACS-catalyzed reaction. Even though the acetyl-CoA generated from this reaction should provide additional carbon for the mevalonate pathway, no further increase in isopentenol titer (or steady state levels of any mevalonate pathway intermediate) was observed after 36 hours. The fact that intracellular citrate concentration and levels of TCA cycle proteins (such as CS and ACONTb) steadily increase during the time-course is consistent with the notion that acetyl-CoA is preferentially shunted into the TCA cycle (to regenerate NADPH) instead of to the mevalonate pathway.

Table S3. Single gene knock outs (SKOs) predicted from Flux variability analysis

Shown in red and bold is gene ***ydbK***, which was experimentally tested to have higher production yields compared to the top-producing strain, I3. The other two genes in red, ytfE and astC, were also experimentally tested but did not have higher production yields compared to strain I3. See iPython notebook, entitled "Engineering_SKOs.ipynb" for detailed scripts and analysis.

gene	gene_name	uniprot	role
b0197	metQ	P28635	transporter
b0198	metI	P31547	aspartate kinase
b0199	metN	P30750	transporter
b4238	nrdD	P28903	ribonucleoside-triphosphate reductase
b1378	**ydbK**	P52647	Probable pyruvate-flavodoxin oxidoreductase
b4209	ytfE	P69506	iron-sulfur cluster, repair
b1747	astA	P0AE37	arginine catabolism
b1748	astC	P77581	transaminase
b2501	ppk	P0A7B1	component of RNA degradosome
b4468	glcE	P52073	glycolate oxidase

A detailed workflow (with all scripts and output) is given in the iPython notebook. In summary, model-driven predictions of SKOs are generated using constraint-based modeling simulations of the wildtype *E. coli* strain constrained by phenotypic data (extracellular metabolomics), similar to stage 3 of this workflow. Here, we ran flux variability analysis (FVA) to screen the effects of genome-wide single gene knockouts. We computed three separate metrics to identify candidate "fitness" level, based on our observations from stage 3 of this workflow: (i) min flux through NADPH consuming reactions; (ii) max flux through NADPH producing reactions and (iii) max flux through PPP pathway. We then rank-ordered the SKOs that maximized or minimized these criteria without reducing the growth rate. A subset of the final list of candidate SKOs were tested experimentally for their effects.

For constructing the single gene knockouts (SKOs), we performed the following procedure: allele replacement to create mutants in E. coli DH1 was performed following the method of Detsenko *et al*

(Datsenko and Wanner 2000). Briefly, a ydbK deletion allele interrupted with a kanamycin cassette was amplified from the KEIO collection ydbK mutant using primers ydbKF 'GGTAATGCACACATCCCAATC' and ydbKR 'GGCCATCAACTTTGCCATAC'. PCR products were introduced into E. coli DH1 harboring pKD46 and transformed as previously described (Datsenko and Wanner 2000). Mutagenesis was confirmed by sequencing.

Identifying metabolic properties relevant to re-engineering

Figure S9. Growth and Isopentenol production for strain I3 (WT) and mutant (*Δ*ydbK)
(Top left) A time-course growth profile of wild type (Strain I3) and mutant *Δ*ydbK strains measured by optical density as well as isopentenol production over the course of the fermentation. The *Δ*ydbK I3 strain produced similar absolute yields of isopentenol with less growth compared to the WT I3 producer. (Top right) shows glucose uptake and acetate production in the context of growth and isopentenol production for both WT and mutant *Δ*ydbK I3 strains. While both strains produced nearly identical amounts of acetate and isopentenol, the *Δ*ydbK I3 strain consumed glucose more slowly and appeared to accumulate less biomass as estimated by optical density. (Bottom) Production and growth profiles for all tested knockouts.

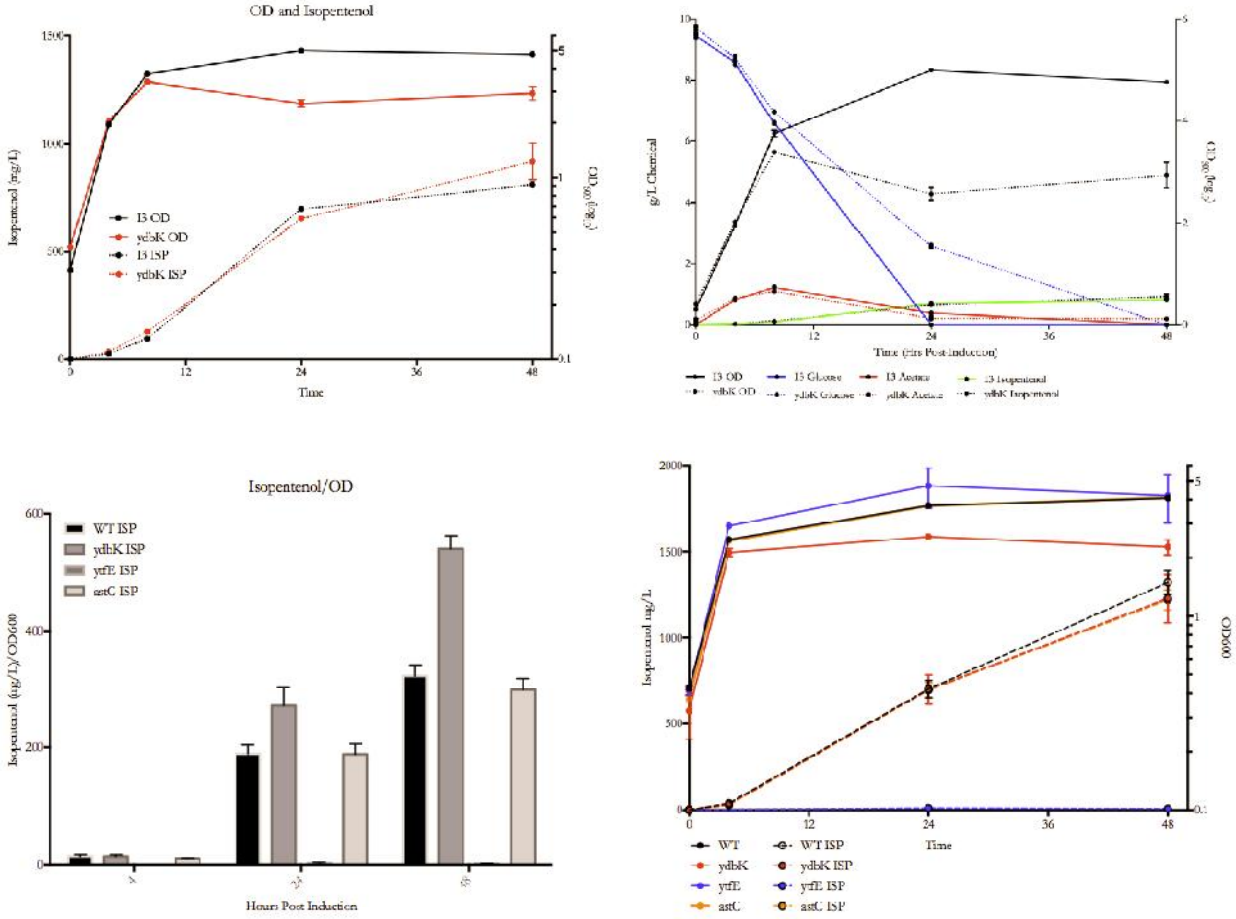
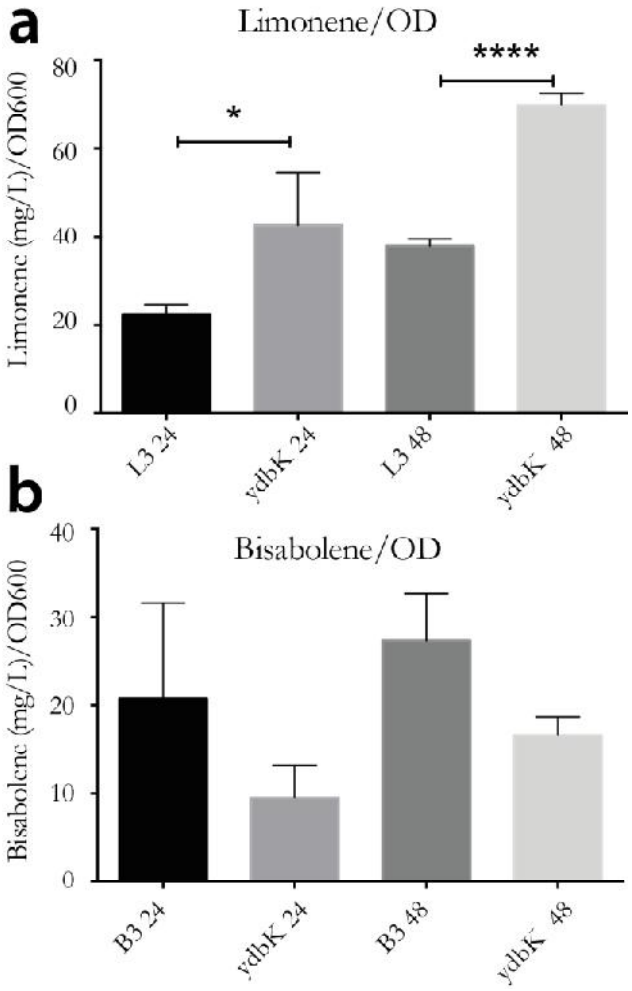


Figure S10. Growth and production for limonene and bisabolene in mutant (*Δ*ydbK) strains

Growth-normalized isopentenol titer (mg/L/OD600) is displayed for (a) strain L3 and L3 + *Δ*ydbK knockout and (b) B2 and B2 *Δ*ydbK knockout, at 24 and 48 hours. For L3, the knockout variant produces significantly more isopentenol than the highest producing strain, L3 (stars denote p-values: 24 hrs p = 0.0437 (**), 48 hrs p < 0.0001 (****), using an unpaired two-tail t-test).



References

Alonso-Gutierrez, Jorge, Rossana Chan, Tanveer S. Batth, Paul D. Adams, Jay D. Keasling, Christopher J. Petzold, and Taek Soon Lee. 2013. “Metabolic Engineering of Escherichia Coli for Limonene and Perillyl Alcohol Production.” *Metabolic Engineering* 19 (September): 33–41.

Becker, Scott A., and Bernhard O. Palsson. 2008. “Context-Specific Metabolic Networks Are Consistent with Experiments.” *PLoS Computational Biology* 4 (5): e1000082.

Chang, Antje, Maurice Scheer, Andreas Grote, Ida Schomburg, and Dietmar Schomburg. 2009. “BRENDA, AMENDA and FRENDA the Enzyme Information System: New Content and Tools in 2009.” *Nucleic Acids Research* 37 (suppl 1). Oxford Univ Press: D588–92.

Chubukov, Victor, Florence Mingardon, Wendy Schackwitz, Edward E. K. Baidoo, Jorge Alonso-

Gutierrez, Qijun Hu, Taek Soon Lee, Jay D. Keasling, and Aindrila Mukhopadhyay. 2015. “Acute Limonene Toxicity in Escherichia Coli Is Caused by Limonene Hydroperoxide and Alleviated by a Point Mutation in Alkyl Hydroperoxidase AhpC.” *Applied and Environmental Microbiology* 81 (14): 4690–96.

Datsenko, K. A., and B. L. Wanner. 2000. “One-Step Inactivation of Chromosomal Genes in Escherichia Coli K-12 Using PCR Products.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (12): 6640–45.

Feist, Adam M., and Bernhard Ø. Palsson. 2008. “The Growing Scope of Applications of Genome-Scale Metabolic Reconstructions Using Escherichia Coli.” *Nature Biotechnology* 26 (6). Nature Publishing Group: 659–67.

George, Kevin W., Amy Chen, Aakriti Jain, Tanveer S. Batth, Edward E. K. Baidoo, George Wang, Paul D. Adams, Christopher J. Petzold, Jay D. Keasling, and Taek Soon Lee. 2014. “Correlation Analysis of Targeted Proteins and Metabolites to Assess and Engineer Microbial Isopentenol Production.” *Biotechnology and Bioengineering* 111 (8): 1648–58.

Hanahan, D. 1983. “Studies on Transformation of Escherichia Coli with Plasmids.” *Journal of Molecular Biology* 166 (4): 557–80.

Keseler, Ingrid M., Julio Collado-Vides, Socorro Gama-Castro, John Ingraham, Suzanne Paley, Ian T. Paulsen, Martín Peralta-Gil, and Peter D. Karp. 2005. “EcoCyc: A Comprehensive Database Resource for Escherichia Coli.” *Nucleic Acids Research* 33 (suppl 1). Oxford Univ Press: D334–37.

Lewis, Nathan E., Gunnar Schramm, Aarash Bordbar, Jan Schellenberger, Michael P. Andersen, Jeffrey K. Cheng, Nilam Patel, et al. 2010. “Large-Scale in Silico Modeling of Metabolic Interactions between Cell Types in the Human Brain.” *Nature Biotechnology* 28 (12): 1279–85.

Mo, Monica L., Bernhard O. Palsson, and Markus J. Herrgård. 2009. “Connecting Extracellular Metabolomic Measurements to Intracellular Flux States in Yeast.” *BMC Systems Biology* 3 (March): 37.

Oberhardt, Matthew A., Bernhard Ø. Palsson, and Jason A. Papin. 2009. “Applications of Genome-Scale Metabolic Reconstructions.” *Molecular Systems Biology* 5 (1). Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1038/msb.2009.77/full.

Peralta-Yahya, Pamela P., Mario Ouellet, Rossana Chan, Aindrila Mukhopadhyay, Jay D. Keasling, and Taek Soon Lee. 2011. “Identification and Microbial Production of a Terpene-Based Advanced Biofuel.” *Nature Communications* 2 (September): 483.

Price, Nathan D., Jan Schellenberger, and Bernhard O. Palsson. 2004. “Uniform Sampling of Steady-State Flux Spaces: Means to Design Experiments and to Interpret Enzymopathies.” *Biophysical Journal* 87 (4): 2172–86.

Price, Nathan D., Ines Thiele, and Bernhard Ø. Palsson. 2006. “Candidate States of Helicobacter Pylori’s Genome-Scale Metabolic Network upon Application of ‘Loop Law’ Thermodynamic Constraints.” *Biophysical Journal* 90 (11): 3919–28.

Schomburg, Ida, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. 2004. “BRENDA, the Enzyme Database: Updates and Major New Developments.” *Nucleic Acids Research* 32 (suppl 1). Oxford Univ Press: D431–33.

Schuetz, Robert, Lars Kuepfer, and Uwe Sauer. 2007. “Systematic Evaluation of Objective Functions for Predicting Intracellular Fluxes in Escherichia Coli.” *Molecular Systems Biology* 3 (July): 119.

Shlomi, Tomer, Moran N. Cabili, Markus J. Herrgård, Bernhard Ø. Palsson, and Eytan Ruppin. 2008. “Network-Based Prediction of Human Tissue-Specific Metabolism.” *Nature Biotechnology* 26 (9): 1003–10.

Thiele, Ines, and Bernhard Ø. Palsson. 2010. “A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction.” *Nature Protocols* 5 (1). Nature Publishing Group: 93–121.

Thiele, Ines, Nathan D. Price, Thuy D. Vo, and Bernhard Ø. Palsson. 2005. “Candidate Metabolic Network States in Human Mitochondria. Impact of Diabetes, Ischemia, and Diet.” *The Journal of Biological Chemistry* 280 (12): 11683–95.