Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering

Jorge Alonso-Gutierrez^{1,2}, Eun-Mi Kim^{1,2}, Tanveer S. Batth^{1,2}, Nathan Cho^{1,2,3}, Qijun Hu^{1,2,4},
Leanne Jade G. Chan^{1,2}, Christopher J. Petzold^{1,2}, Nathan J. Hillson^{1,2}, Paul D. Adams^{1,2,4},
Jay D. Keasling^{1,2,4,5}, Hector Garcia-Martin^{1,2,*}, Taek Soon Lee^{1,2,*}

¹Joint BioEnergy Institute, 5885 Hollis Street, Emeryville, CA 94608, USA.

²Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

³Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA 90095, USA

⁴Department of Bioengineering, University of California, Berkeley, CA 94720, USA.

⁵Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA.

*Corresponding authors:

Taek Soon Lee, Tel.: +1 510 495 2469, Fax: +1 510 495 2437, E-mail: tslee@lbl.gov

Hector Garcia Martin, Tel.: +1 510 495 2462, Fax: +1 510-486-4252, E-mail: hgmartin@lbl.gov

Key words

Targeted proteomics; principal component analysis (PCA); metabolic engineering; mevalonate pathway; heterologous pathway; *Escherichia coli*

Abstract

Targeted proteomics is a convenient method determining enzyme expression levels, but a quantitative analysis of these proteomic data has not been fully explored yet. Here, we present and demonstrate a computational tool (Principal Component Analysis of Proteomics, PCAP) that uses quantitative targeted proteomics data to guide metabolic engineering and achieve higher production of target molecules from heterologous pathways. The method is based on the application of principal component analysis to a collection of proteomics and target molecule production data to pinpoint specific enzymes that need to have their expression level adjusted to maximize production. We illustrated the method on the heterologous mevalonate pathway in *E. coli* that produces a wide range of isoprenoids and requires balanced pathway gene expression for high yields and titers. PCAP-guided engineering resulted in over a 40% improvement in the production of two valuable terpenes. PCAP could potentially be productively applied to other heterologous pathways as well.

Introduction

Microbial systems are renewable alternatives to chemical synthesis for the production of desired chemicals (Keasling, 2010; Lee et al., 2012; Woolston et al., 2013). By combining genes from different organisms into a metabolic pathway and expressing it in a host organism, it is possible to produce a wide range of valuable molecules from biofuels (Atsumi et al., 2008; Bokinsky et al., 2011; Dellomonaco et al., 2011; Goh et al., 2012; Peralta-Yahya et al., 2012) to commodity chemicals and pharmaceuticals (Ajikumar et al., 2010; Juminaga et al., 2012; Keasling, 2010; Martin et al., 2003; Stephanopoulos, 2012). Improving the titer of the desired metabolite is not a trivial task, and frequently it is dependent on a delicate balance of the expression level of the genes in the heterologous pathway (Ajikumar et al., 2010; Dahl et al., 2013; Holtz and Keasling, 2010; Yadav et al., 2013; Zhang et al., 2012).

Among previous attempts to coordinate gene expression in heterologous pathways to date (Ajikumar et al., 2010; Anthony et al., 2009; Dueber et al., 2009; Moon et al., 2010; Pitera et al., 2007), only a few have used ‘omics’ data, and even in those studies, they used ‘omics’ data for just a small number of samples and in a qualitative fashion (i.e. pathway bottlenecks were detected either through monitoring of the accumulation of specific metabolites (Pitera et al., 2007) or by pinpointing enzymes with anomalously low expression level (Redding-Johanson et al., 2011)). However, small quantitative deviations in pathway expression from the optimal window can lead to huge differences in production (Ajikumar et al., 2010). Therefore a tool able to rationalize how these quantitative differences translate in such disparate production levels would be highly desired.

In other areas of engineering, mathematical models and tools are routinely used to drive the engineering process. Unfortunately, the lack of accurate predictive capabilities in biological systems has limited the use of models in metabolic engineering (Di Ventura et al., 2006) which is more often guided by intuition than a quantitative understanding of pathway function. The ability to predict and test the ideal pathway balance using mathematical models based on analytical data would represent an important milestone in metabolic engineering.

Only recently, a more rational approach has been applied to address the expression balance and direct further engineering efforts based on transcriptomic (or q-PCR) data (Ajikumar et al., 2010). In this approach, the metabolic pathway is partitioned into pre-defined modules and then a systematic multivariate search is used to identify expression conditions that optimally balance the pathway modules so as to maximize the production of the target (Ajikumar et al., 2010; Xu et al., 2013). Although this method enables significant improvements in the titers of the desired products, it may be limited in use since it requires the researchers to pre-define the pathway modules, based on previous knowledge. While the

transcript level is, at best, a surrogate for the level of the encoded protein (Maier et al., 2009), targeted proteomics approaches, via selected-reaction monitoring (SRM) mass spectrometry coupled to liquid-chromatography separation (LC-MS and LC-MS/MS), provide accurate measurements of many specific proteins in a sample (Redding-Johanson et al., 2011). Since enzyme expression is typically the only parameter of the metabolic reaction directly controlled through genetic engineering tools, proteomics arises as the most informative data type to initially base our models for pathway balancing.

Here, we present an objective and simple mathematical tool based on Principal Component Analysis (PCA) of proteomics data able to identify specific enzymes or groups of enzymes putatively limiting production in a pathway. Although pathway performance is the result of complex interactions of different enzymes and metabolites, a direct proportionality between the rate of reaction and the enzyme concentration has been proposed to be of primary importance in the determination of flux changes, whereas the complexity of enzyme-metabolite interactions has been considered secondary (Rossell et al., 2011). Therefore, we assume here that the analysis of a pathway's relative enzyme levels would be sufficient to determine the optimal balance for a specific producing system.

PCA is a well-established, mathematical tool commonly used for data analysis and modeling in many fields (Abdi and Williams, 2010; Alter et al., 2000; Cangelosi and Goriely, 2007; Crescenzi and Giuliani, 2001; Holter et al., 2000; Misra et al., 2002; Pearson, 1901). PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables (in this case, relative enzyme concentrations) into a set of values of linearly uncorrelated variables called principal components (PC) that reveal the internal structure of the data in a way that best explains their variance.

In the present study, we used targeted proteomics to explore how the heterologous expression of the nine mevalonate pathway enzymes in *E. coli* affects the total production of

the monoterpene limonene. While gene sequences were maintained constant, the expression of the enzymes in different gene clusters was generated through variation of promoter strength, different plasmid copy numbers, and under different induction timings and levels (Figure 1). This protocol generated a diverse set of proteomics data for these nine genes and product levels. PCA of proteomics data (PCAP) for all these conditions was used to provide a natural basis for understanding the final product titers in terms of the proteomics data. We limited the PCA to the proteomics data since it is protein expression (and not production) that can be directly manipulated through standard metabolic engineering approaches, such as the ones used here. PCAP suggested that a combination of both an overexpression of the bottom portion of the pathway and a balanced expression of the remaining pathway enzymes are key to improve production. In the following sections we show evidence of how engineering based on PCAP results helped to improve production of two valuable terpenes ~40% beyond already high titers.

2. Materials and Methods

All chemicals, solvents and medium components were purchased and used without modification from Sigma-Aldrich (St. Louis, MO), Fisher Scientific (Pittsburgh, PA), or VWR (West Chester, PA) unless otherwise noted. *E. coli* strains DH10B (Invitrogen, Carlsbad, CA) and DH1 (ATCC) were used for plasmid construction and production experiments, respectively. For targeted proteomics experiments, mass spectrometric-grade trypsin was obtained from Sigma-Aldrich and prepared according to manufacturer's instructions.

2.1. Strain construction

E. coli DH10B was used as the host for cloning, and *E. coli* DH1 was used as the host

for limonene and bisabolene production. All plasmids and strains used in this study are described in Table 1. The original isoprenoid precursor-producing system consists of two plasmids (pMevT and pMBI) (Martin et al., 2003) and was subjected to different modifications to increase limonene production (Alonso-Gutierrez et al., 2013). In the present study, the best set of genes from the previous limonene production work was arranged in different ways to explore the optimal expression profile. Three different *E. coli* DH1 strains were used (Figure 1, Table 1): strain 2p (previously described as 2p-SA-f-tr (Alonso-Gutierrez et al., 2013)), strain B (previously described as 1pC (Alonso-Gutierrez et al., 2013)) and strain A (constructed for the present study). The different strains contain the same best set of genes for the production of limonene but organized in clusters and plasmids with different promoter strengths and copy numbers. In order to enable rapid cloning and assembly of genes, we employed the BglBrick cloning strategy (Anderson et al., 2010). All plasmids denoted pBb were prepared using the BglBrick standard collection of expression vectors (Lee et al., 2011).

Detailed plasmid descriptions (*i.e.*, annotated sequence files available in FASTA, Genbank, and SBOL XML/RDF formats and physical location) are available in the public instance of the JBEI Public Registry (<https://public-registry.jbei.org> (Ham et al., 2012); Table 1)

2.2. Growth conditions and terpene analysis

E. coli strains harboring limonene or bisabolene production plasmids were grown in EZ-Rich defined medium (Teknova, Hollister, CA) supplemented with 1% glucose and appropriate antibiotics (Table 1). Working concentrations were 30 and 100 µg/mL for chloramphenicol and ampicillin, respectively. Starter cultures were grown overnight at 37°C,

and all samples were shaken in rotary shakers at 200 rpm (Kühner, Basel, Switzerland). Production cultures were inoculated at an optical density (OD_{600nm}) of 0.1, and allowed to continue growth at 30°C, as the lower temperature improved terpene production. Each of the three different strains was induced at three different OD_{600nm} : 2.5 (H); 1.2 (M) or 0.2 (L); and at three different isopropyl β -D-1-thiogalactopyranoside (IPTG) concentrations: 500 μ M (h); 100 μ M (m); 25 μ M (l)); and were sampled every 24 hours up to 72 hours for both terpene production and OD_{600nm} , while protein measurements were only taken at 24 hours during late exponential phase. OD_{600nm} measurements were conducted on a UV–Vis spectrophotometer (Beckman, Fullerton, CA) operating at 600 nm. For production assays, a dodecane layer (10% by volume) was added to the culture upon induction to trap the terpene, and 10 μ L of the organic phase was sampled and analyzed using gas-chromatography mass-spectrometry (GC/MS) as reported previously (Alonso-Gutierrez et al., 2013; Peralta-Yahya et al., 2011).

2.3.2. Targeted proteomics

Peptide samples were analyzed on a LC/MS/MS system consisting of an Agilent 1100 HPLC coupled to an AB Sciex (Foster City, CA) 5500Q-Trap MS or on an Agilent 1260 HPLC coupled to an Agilent 6460QQQ MS. Peptide samples were loaded and separated on a Sigma Ascentis Express Peptide C18 (5 cm length x 2.1 mm ID, 2.7 μ m particle size) column via an Agilent micro-auto well plate sampler (for loading) and pump operating in normal flow at 400 μ L/min. A separation gradient for the 1100/5500Q-Trap system were as follows: Buffer A (2% acetonitrile, 0.1% formic acid) and Buffer B (98% acetonitrile, 0.1% formic acid) was applied as follows: 5% B for 0.5 min, followed by an increase to 40% B in 9.5 minutes followed by a ramp to 90% B in 30 seconds where it was held for 3.5 minutes. Buffer B was then decreased to 5% B in 30 seconds where it was held for 5.5 minutes to allow for re-equilibration. For the 1260/6460QQQ system, a 13-minute method with the

following gradient was utilized: 95% Buffer A (0.1% formic acid), 5% Buffer B (98% acetonitrile, 0.1 % formic acid). Buffer B was increased to 35% over 5.70 min, followed by an increase to 90% B in 0.3 min, where it was held for 2 min. Buffer B was then decreased to 5% in 0.5 min where it was held for 4.5 min to re-equilibrate the column. Data on the 5500Q-Trap MS were acquired in positive-ion mode with Analyst™ 1.5 operating in scheduled MRM mode with a scheduling window of one minute and target scan time of 0.6 seconds with source parameters of: ion spray voltage = 5000 V, 650°C source temperature, nebulizing gas (GS1 = 45), heating gas (GS2 = 45) and curtain gas (CUR = 25). Resolution for Q1 and Q3 were set to unit resolution (0.7 Da). For the 6460QQQ, the peptides were ionized by an Agilent Jet Stream ESI source operating in positive-ion mode with the following source parameters: Gas Temperature = 350°C, Gas Flow = 10 L/min, Nebulizer Pressure = 45 psi, Sheath Gas Temperature = 400°C, Sheath Gas Flow = 11 L/min, Nozzle Voltage = 1000 V, Chamber Voltage = 5000 V. The data were acquired using Agilent MassHunter, version B.05.00. and Q1 and Q3 resolution set to unit.

2.3.3. Proteomic quantitative analysis

Skyline version 1.3 software (MacCoss Lab, University of Washington)(MacLean et al., 2010) was used to determine the peak area for all SRM transitions. Two transitions were selected for every peptide, two peptides per protein, peak areas were summed or one peptide validated from a synthetic standard for each protein to give a total peak area. Total peak area for each protein was normalized to the total peak area for the internal standard, BSA, to give a normalized total peak area for each protein, similar to a method reported earlier (Silva et al., 2006). Each normalized total peak area was averaged for three replicate analyses, and the standard deviation was calculated.

2.4. Principal component analysis

We used Principal Components Analysis (Wall et al., 2003) to find which linear combinations of protein levels would explain most of the variability of the data for the different assays. The protein data were stored in a matrix $\mathbf{X}$ with elements X_{ij} corresponding to the i th protein ($i=1\dots m$) and j th scenario ($j=1\dots n$). The rows of $\mathbf{X}$ were ‘centered’ by subtracting the mean of the row from each row, producing the matrix $\underline{\mathbf{X}}$, where:

$$\underline{X}_{ij} = X_{ij} - \sum_{j=1}^n X_{ij} / n$$

The rows of $\underline{\mathbf{X}}$ are then the centered values of the protein levels for each scenario:

$$\underline{X} = \begin{bmatrix} \dots & \vec{g}_1 & \dots \\ \dots & \vec{g}_2 & \dots \\ \dots & \vdots & \dots \\ \dots & \vec{g}_m & \dots \end{bmatrix}$$

The product of $\underline{\mathbf{X}}$ by its transpose $\underline{\mathbf{X}}^T$ is then the covariance matrix of the protein levels:

$$(\underline{X}\underline{X}^T)_{ij} = \vec{g}_i \cdot \vec{g}_j = \sum_{k=1}^n \underline{X}_{ik} \underline{X}_{jk}$$

If this covariance matrix were diagonal (e.g., non-zero elements only on the diagonal) that would mean that the levels of the any two of the proteins in the mevalonate pathway (MEV) are uncorrelated. The values in the diagonal would then be proportional to the variance over the scenarios for each protein. Large values would indicate proteins that change considerably over the different scenarios and are more likely to influence production. Smaller values of the diagonal would correspond to protein levels that do not vary much over scenarios, and are less likely to influence production.

In order to reduce our data set to this easily diagnosable situation, we obtained a diagonal covariance matrix by using a linear combination of protein levels instead of the original protein levels. Through a Singular Value Decomposition (SVD) one can obtain a diagonal matrix S and matrixes U and V where the columns of U and V are orthonormal bases such that:

$$\underline{X} = U S V^T$$

by using the matlab (The MathWorks, Inc., Natick, MA) function: $[U, S, V] = \text{svd}(X)$.

The covariance matrix then looks like:

$$\underline{X} \underline{X}^T = U S V^T (U S V^T)^T = U S V^T V S^T U^T = U S^2 U^T$$

And by multiplying by U^T and U in the front and in the back, one can diagonalize the covariance matrix:

$$U^T \underline{X} \underline{X}^T U = U^T U S^2 U^T U = S^2$$

This is the same as the covariance matrix for an experiment in which the profiles of protein levels are linear combinations of the original data (really just a projection into a new basis since U represents an orthonormal base):

$$U^T \underline{X} \underline{X}^T U = U^T \vec{g}_i \vec{g}_j^T U = (U^T \vec{g}_i) \cdot (U^T \vec{g}_j)$$

The first two components of the projected data ($\text{proj} = U' * X$, in matlab code) are plotted as principal components 1 (PC1) and 2 (PC2) in Figure 3a. The relative variances for each PC are plotted in Figure 3b: $\text{variance} = 100 * \text{diag}(S.^2) / \text{sum}(\text{diag}(S.^2))$, in matlab code. The first two PC's amount to 92% of the variance. Finally the normalized contribution of each protein for each of the first two principal components can be seen in Figure 3c.

3. Results and Discussion

3.1. Proteomics profile generation and qualitative analysis

In a previous work, a nine-enzyme mevalonate pathway was introduced into *E. coli* to

produce the monoterpene limonene from 1% glucose (Alonso-Gutierrez et al., 2013). Genes were selected from different organisms, subjected to different modifications (e.g. codon optimization, truncation, etc) and organized into three clusters expressed on either one or two plasmids (strains B and 2p, respectively (Figure 1, Table 1)). These engineering efforts led to a ~435 mg/L of limonene production, representing a 225-fold production improvement from the original system (Alonso-Gutierrez et al., 2013). However, the rationale used to choose the relative protein expression levels and desired targets to engineer was heuristic and relied heavily on previous successful attempts for other terpene production systems (Martin et al., 2003; Peralta-Yahya et al., 2011; Redding-Johanson et al., 2011). We hypothesized that a more rational approach based on exploring the different proteomic profiles for those nine enzymes with its corresponding terpene level would help us to find a better expression profile and therefore to improve the titers.

We redesigned three different strains for limonene biosynthesis (Figure 1, Table 1); strain 2p expresses the nine pathway proteins using two plasmids, and strains A and B are single plasmid-harboring versions for limonene production. In these strains, we defined gene clusters (CLs) based on the order of the enzymes in the pathway reactions (Figure 1). CL1 includes genes for mevalonate formation, and CL2 includes genes for IPP and DMAPP formation. CL3 includes GPP synthase and limonene synthase which are the last two enzymes in the pathway. Each of these strains has different genetic context for limonene production (i.e. different expression control for the MEV coding genes in the clusters (Figure 1, Table 1)), and all three strains were induced at three different cell densities (OD_{600nm}) and at three different IPTG levels as described in Materials and Methods. Thus, a total of 27 different production scenarios were generated and analyzed (Figure 2).

The highest producers among the 27 different scenarios were 2p-Mm and B-Ml (i.e. Strains 2p and B induced at medium OD_{600nm} (1.2 (M)) with either medium (100 μ M (m)) or

low (25 μ M (l)) IPTG levels, respectively). The most conspicuous feature from Figure 2 is that relatively low overall levels of pathway enzymes are needed to attain the highest limonene production. The maximum production level was nearly identical for both highest producing strains; B-Ml produced 435 ± 18 mg/L and 2p-Mm produced 435 ± 23 mg/L of limonene after 72 hours of cultivation. However, overall pathway enzyme levels for B-Ml were much lower than those of 2p-Mm (particularly the level of limonene synthase) indicating that balanced enzyme expression is more productive than the simple overexpression of a single gene.

The next interesting observation from Figure 2 is that strain B is highly sensitive to the inducer level (i.e., compare B-Ll, B-Lm and B-Lh). These results support our hypothesis that comparatively high titers of isoprenoid product can be obtained by fine-tuning of the expression of MEV genes rather than overexpressing them. Another surprising result is the low production by strain A regardless what the induction conditions are. Strain A was one of our ‘rationally-designed’ attempts to improve the production by achieving a better consumption of acetyl-CoA using high copy plasmid (Alonso-Gutierrez et al., 2013). However, the overexpression of pathway enzymes in strain A led to very low production levels compared to 2p and B. Moreover, strain A reaches its highest production only when induced at a higher OD, in contrast to strains 2p and B where maximum production levels are only reached when induced at medium cell densities, and this indicates that A strain might accumulate toxic intermediates at an early stage as we will discuss later. In any case, this is a good indication that our three systems are covering a good range of different expression conditions making the PCA more informative and thus providing robust conclusions.

Most of all, the targeted proteomics data shows no clear trends to indicate any single crucial enzyme to be controlled or an expression profile to target in order to improve production. Although strains with low levels of enzymes in a balanced context yield higher

terpene titers than strains with generally high levels of enzymes as can be seen in Figure 2 and has also been reported previously as a heuristic principle (Ajikumar et al., 2010; Dueber et al., 2009; Pitera et al., 2007), there is no obvious next engineering step to improve production. For a given set of genes, we hypothesized the existence of an optimal proteomics profile (i.e. relative concentration of pathway enzymes) that leads to a maximum production and that such optimal profile can be found by exploring correlations of proteomics and production data. Therefore, a mathematical analysis (through PCA) of the protein expression profiles from these 27 scenarios was used to quantitatively pinpoint which enzymes to target in order to balance the pathway and maximize its limonene production in this multi-variant system.

3.2. Principal component analysis (PCA) of proteomic data

In our model, we assume that the relevant variables to describe our system are the protein concentrations of each of the 9 pathway enzymes for limonene biosynthesis. These concentrations can be used to define a vector in a nine-dimensional space (i.e., one axis per enzyme), which describes the state of the system. The production of limonene is a function of this nine-dimensional vector. The optimal vector (i.e., protein expression profile that maximizes production) could, in principle, be found through an exhaustive search of this phase space by experimentally varying the expression level of each gene in the pathway. However, the amount of experiments needed for this approach is not feasible to conduct at this time. Therefore, we decided to consider this multivariate dataset of 27 scenarios as coordinates in the nine-dimensional data phase space (with titers as function values) and use PCA to visualize a lower-dimensional picture, a projection of this high dimensional space from its most informative viewpoint in order to intuitively understand how production varies depending on the position in this space.

When the set of 27 different production scenarios were analyzed using PCA, the variation in production as a result of the associated proteomics became self-explanatory. By using only the first few, most informative principal components (PCs), the dimensionality of the transformed data was reduced to only two. The first two components of the PCA (Figure 3b) explained 92% of the data variance and PC1 by itself accounted for almost 60% of the total. PCs are a simplification of the phase space obtained from all the available data (protein levels for the 9 enzymes) and their interpretation is based on the percentage of contribution of each of the variables and user criteria guided by scientific knowledge. Figure 3 clearly shows that both maximum production scenarios, B-MI and 2p-Mm, are completely separated in the PCA space but still bear similar production levels (i.e., ~435 mg/L). The mechanistic reasons why each scenario reaches that maximum value are therefore different and can be explained by taking a closer look at the relative contributions of each gene to each principal component. This difference would have been easily overlooked without the PCA. From the protein absolute values (Figure 2) it is clear that B-MI displays much lower overall protein levels than 2p-Mm. In the PCA phase space in Figure 3, B-MI is situated near the center of the plot, showing a balanced contribution of both principal components. However, when we look at the other two strains (i.e., 2p and A) we can see a clear separation of their associated trends: 2p-Mm maximum is biased towards the top of the PC2 and the very left of PC1 but its production is not affected compared to B-MI indicating that a higher contribution of PC1 is not detrimental for production. In contrast, all of the A-derived scenarios are situated on the bottom of the PCA chart and have a production titer of ~90 mg/L (4 times lower than 2p or B systems). Irrespective of the contribution of the PC1, the PC2 contribution towards the bottom seems to determine low production levels. Therefore, it is clear that PC1 and PC2 have opposite effects on production; while PC1 contribution towards left is helpful and harmless, PC2 is clearly detrimental towards the bottom.

The proteomics data and its contribution to each PC (Figures 2 and 3c, respectively), indicate that the relative level of limonene synthase (LS) is mainly associated with PC1. In fact, this hypothesis is confirmed by the mathematical analysis (Figure 3c), where more than 40% of PC1 is due to LS levels. Since PC1 is mostly contributed by LS, we choose to rename it as “LS” to help understanding which engineering direction should be taken to improve production. PC2, on the other hand, is not dominated by a single gene expression and is therefore not easily labeled in this reductionist manner. Using the information derived from the strain A proteomic levels (Figure 2), we can infer the components of PC2: an overexpression of the top portion (especially HMGS and HMGR) seems to be the main factor that sets the A producing system apart from B and 2p. In fact, the unbalanced expression of these enzymes is known to lead to HMG-CoA accumulation and toxicity (Pitera et al., 2007). This is in agreement with the consistently low production levels from strain A and the fact that maximum productions are only reached when induced at high OD levels. PC2 therefore seems mostly related to the protein levels of the top portion enzymes and indeed, the mathematical analysis (Figure 3c) supports that both HMGS and HMGR enzymes are main contributors of PC2. Since the definition in this case is not as clear as PC1 and none of any specific enzymes are clearly defining PC2, it seems reasonable to rename this component as “TOP” (or “Top portion”).

In light of this analysis, we hypothesize that high production requires a relatively balanced pathway enzyme levels with a higher LS expression with respect to the rest of the pathway, and especially to the top portion. Indeed, every data point in the PCA distribution with the redefined components (Figure 3a) supports such a hypothesis. For example, the B-MI scenario (i.e., strain B induced at medium OD and low IPTG concentration) results in low LS levels but also lower levels of the rest of the pathway enzymes, maintaining a good balance of both and therefore leading to the highest limonene level. However, small

deviations from this delicate balance lead to dramatic drops in limonene production in most cases. Production scenarios derived from 2p strain follow an analogous trend as those from strain B, where deviations from a similar balance also lead to drops in titer. However, the LS concentrations of 2p strain are consistently high and, therefore, top portion protein expression can be increased to higher levels than in the case of strain B without encountering deleterious effects for production. This supports the idea that strain 2p is a more robust production system and explains why most of the scenarios derived from this two-plasmid system are among the top producers irrespective of the induction condition. In radical contrast, strain A consistently generates only low limonene titers due to the clear unbalance between “TOP” and “LS” expressions. The high expression of top portion enzymes HMGS and HMGR in relation to the rest of the pathway and especially to LS situates all the strain A scenarios on the bottom of the PC2 where only low production levels can be attained.

3.3. PCAP-directed metabolic engineering leads to balanced expression and higher titer

Following our working hypothesis and based on the data distribution, one would expect higher production levels for scenarios on the upper-left part of the PCA phase space in Figure 3a. The easiest way to create such a scenario would be to overexpress the genes contributing to the specific components in the desired direction. However, having higher enzyme level does not necessarily translate into more activity or higher production. Many proteins are easily saturated in the cytoplasm at relatively low concentrations and overexpressing them would only lead to increments of its insoluble and inactive fraction (Francis and Page, 2010). Fortunately, the B system uses very low expression levels to reach its maximum production and therefore there should have capacity for a relative overexpression of any of the enzymes. However, the tight balance of the protein expression in B strain should not be perturbed. Furthermore, in order to avoid detrimental effects derived

from changes in PC2, we would like to minimize the number of modifications. As previously noted, relative increases in LS level are beneficial or at least non-detrimental for production. Therefore, based on the PCA distribution, the simplest way to shift the balance to the left of PC1 without modifying PC2 would be to increase LS levels in strain B.

We implemented this change and prepared a new strain by co-transforming a second plasmid containing additional copy of LS together with plasmid B (strain BL in Figure 4a). As already discussed, the PCA space is not very smooth and the B strain is highly sensitive to protein levels, so it is difficult to predict *a priori* what the new system (BL) will do along different expression levels. To explore the gradient and find the optimum induction condition for the new strain (BL), the cells were induced at three different IPTG levels (i.e. 25 μ M (l), 100 μ M (m) or 500 μ M (h)) at a medium OD_{600nm} (OD ~ 1.2). The optimal induction condition for production was achieved at over 100 μ M IPTG (BL-Mm scenario; Figure 4b), and as expected, the simple overexpression of LS in a second plasmid improved limonene titer up to ~605 mg/L, representing a 40% improvement over the best previous level (Figure 4b). In contrast with the strain B, which reaches its maximum titer when induced at 25 μ M IPTG (i.e., B-Ml scenario) and a small increase in induction level leads to significant changes in production, the production in strain BL was less affected by the inducer concentration than that of strain B. This difference of optimal induction conditions between the single- (i.e., strain B) and the two-plasmid systems (i.e., BL and 2p strains) is not entirely surprising. It has been observed that the amount of LacI (repressor) synthesized from the second, high-copy plasmid in BL and 2p strains is quite high and therefore tighter regulation is achieved. As a result, more inducer is required for this two-plasmid system to attain a similar enzyme levels as in strain B (Alonso-Gutierrez et al., 2013).

To further test our working hypothesis, we ran the PCAP method again including proteomic results from the three scenarios derived from the BL strain. Indeed, the position of

the maximum producers is in accordance with the PCs definition and our hypothesis (Figure 4c). BL-Mm and BL-Mh scenarios are positioned towards the top left of the PCA space. The relative increase of LS, in the context of the low levels provided by plasmid B for the remaining pathway enzymes, moved the BL-Mm and BL-Mh scenarios towards the left of the PCA diagram as expected but also upwards, since the relative contribution of PC2 decreased. BL-Ml, which was induced at 25 μ M IPTG, did not have enough inducer level to overcome the high levels of LacI and thus led to overall low protein and limonene levels. The PCA may not offer any clear explanation how these changes in production profile occur, and the interpretation is left to the researcher. However, this experiment strongly supports our hypothesis on limonene production: the key for balancing the pathway enzymes expression and therefore increasing limonene production from our current production platform is to maintain the top portion expression level and to overexpress the limonene synthase from a high-copy plasmid.

3.4. Extension of the inferred balance strategy to a different MEV system

The strategy suggested from PCAP to balance the expression of MEV enzymes to produce limonene was tested for a different pathway endpoint, bisabolene, which uses the same basic MEV architecture (Figure 5). The isoprenoid building blocks, isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP), as well as most of the metabolic intermediates are universal for the production of any terpene, and therefore the benefits derived from the balanced expression of this system seem likely to apply to any other terpenoid biosynthesis. Recently, microbial production of bisabolene, whose fully hydrogenated form (bisabolane) is considered a D2 diesel alternative, was demonstrated by expressing the codon-optimized version of the bisabolene synthase gene from *Abies grandis* (AgBIS) (Peralta-Yahya et al., 2011) (Figure 5a). A similar set of the MEV gene clusters,

which has been used for limonene biosynthesis, was used in *E. coli* to provide the AgBIS with precursors. Since bisabolene is a sesquiterpene (C₁₅), two enzymes (FPP synthase instead of GPP synthase, and bisabolene synthase instead of limonene synthase) are different from the limonene producing pathway, but the overall cluster structure is similar (Figure 5). As is frequently observed in most terpene synthases, AgBIS is known to be a very slow enzyme and does not express as a soluble protein very well (Özaydın et al., 2013; Peralta-Yahya et al., 2011). Therefore, we hypothesized that applying the same balancing strategy based on the relative overexpression of the terpene synthase in a balanced context of the rest of the pathway genes would increase bisabolene production as well. Analogous to strain construction for limonene production, we started with the strain reported to produce the highest bisabolene titers (strain PY (Table 1) in the present study (Peralta-Yahya et al., 2011)). The plasmid pBbA5c-CL1'-T1-CL2', providing pathway genes to transform acetyl-CoA into FPP, has a similar gene context to pBbA5c-CL1-T1-CL2 of strain 2p for limonene production, and therefore provides a balanced expression of the top portion similar to the limonene system (Table 1, Figures 1b and 5b). As a mimic of B strain in limonene system, strain C was prepared by combining pBbA5c-CL1'-T1-CL2' and pTrc-AgBIS into a single plasmid (pBbA5c-CL1'-T61-CL2'-T21-AgBIS) (Figure 5b, Table 1). Following the same strategy that leads to BL strain construction, this single plasmid was co-transformed into *E. coli* with a high-copy plasmid containing an extra copy of bisabolene synthase to generate strain CB (Figure 5b, Table 1). Induction of this modified bisabolene producer (strain CB) at OD_{600nm} of 1.2 with 100 µM IPTG rendered about 1150 mg/L of bisabolene (Figure 5c).

This bisabolene level is the highest reported so far and a 2-fold increase from the published two-plasmid system (strain PY) (Peralta-Yahya et al., 2011), which produces about 550 mg/L bisabolene under the conditions of this study. This result, which showed a similar outcome to limonene production engineering, suggests that this approach (i.e., balancing the

pathway enzyme levels and simultaneously overexpressing the terpene synthase) is potentially of general application to improve any terpenoid production system based on the same set of pathway genes. It is very interesting that overexpression of terpene synthase by itself does not always improve the production, but it needs a simultaneously decreased level of HMGS and IspA (or PC2) to balance the pathway for higher titer as shown in the scenarios CB-500 μ M, C-500 μ M or C-100 μ M (Figure 6). Furthermore, PCAP of these strains confirms again the trend and hypothesis derived from the limonene system, possibly because both systems share the precursors, redox and ATP requirements. In this case, the PC1 associated with the bisabolene synthase (BIS) level by itself explained 98% of the variance in production (Figure 6). This increase in production may support the idea that the insights obtained from PCAP can produce biological knowledge that could be used to devise general metabolic engineering strategies.

3.5. PCAP is a robust tool predicting unbalances from protein expression

To further confirm that simply overexpressing the terpene synthase is not sufficient and that unbalances of the top portion must be addressed first, we generated strain BH, which harbors plasmid B and a second plasmid overexpressing HMGS and HMGR, and strain AL, which carries plasmid A and a second plasmid overexpressing LS (Table 1). We analyzed protein and limonene levels for strains BH and AL together with strains A, B and BL as a reference (Supplementary Figure S2). Overexpression of HMGS and HMGR in strain BH lead to a dramatic decrease in production and a shift from strain B, in the PCAP space, towards the right end of PC1 (associated with higher HMGS/HMGR levels; Supplementary Figure S2). Low limonene titers in strain BH are most probably due to toxicity from HMG-CoA accumulation as previously reported (Pitera et al., 2007); low OD_{600nm} values of this strain support this hypothesis (data not shown). In contrast, strain BL shows a shift towards

the top of PC2 (associated with LS levels) attaining higher production than strain B, as already observed. However, overexpression of LS is not always the answer to improve production. Strain AL did not recover production by increasing the LS levels relative to those of strain A. As expected, due to overexpression of LS, the proteomic results for strain AL placed it at a higher end on the PC2 (associated with higher LS levels) relative to strain A (Supplementary Figure S2). However, a drastic reduction in production rather than any improvement was observed for strain AL. The already unbalanced expression in strain A makes this production system not suitable for improvement. All these observations further confirm our working hypothesis that not only higher expression of the terpene synthase, but also a balanced and low expression of the rest of the pathway, is necessary for improvement.

4. Conclusion

We have shown an example of how mathematical tools such as PCA based on analytical data can direct metabolic engineering of heterologous pathways and achieve significant production improvements in “non-intuitive” manners. In our systems, the terpene synthase already appears to be the most abundantly expressed protein in the pathway, and a qualitative analysis of the proteomics data would not have picked it for further overexpression. Counterintuitively, however, PCAP and production data suggested that both an overexpression of the terpene synthase combined with a balanced expression of the rest of the enzymes are keys to improve production. Extra copies of the terpene synthase while balancing expression of the other pathway enzymes led to a 40% increase in limonene production as suggested by the PCAP. Furthermore, inspired by the same strategy suggested by the PCAP for limonene, an unprecedented 1150 mg/L of the valuable biofuel bisabolene was also attained. This amount represents nearly a doubling of the already high titers previously reported (Peralta-Yahya et al., 2011).

While the PCAP method presented in this particular case has pointed to a single protein (terpene synthase) as an engineering target, this method can, in principle, be applied for the system whose relevant principal component is significantly contributed by more than one enzyme. In this case, the method can pinpoint the enzymes whose expression must be altered to increase production and the actual genetic manipulations to achieve these expression level changes are more involved. Recent efforts to predictably produced desired amounts of protein through precise and reliable gene expression control elements (Mutalik et al., 2013) will aid both predictions from PCAP and the subsequent engineering steps suggested by the analysis. Advances in proteomics analysis have enabled PCAP to draw interesting conclusions from a handful set of data. In the near future, we envision biological engineering as a discipline directed by increasingly complex and more predictive metabolic models based on larger sets of omics data. High-throughput sampling and analysis of transcriptomics, metabolomics and proteomics (Batth et al., 2014) provides a wealth of data suitable for PCA or modeling. Models based on such detailed information will be able to predict changes in fluxes and suggest engineering strategies to optimize production of valuable chemicals, which are not apparent by simple observation of the omics data.

Disclosure statement

Jay D. Keasling has financial interests in Amyris, LS9, and Lygos. Nathan J. Hillson has financial interests in TeselaGen Biotechnology, Inc.

Acknowledgments

The authors thank Pamela Peralta-Yahya and Han Min Woo at JBEI for providing plasmids pTrc-LS and pBbE1a-HMGR-HMGS, respectively. This work was part of the DOE Joint BioEnergy Institute (<http://www.jbei.org>) supported by the U.S. Department of Energy,

Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. J.A-G. thanks “Fundación Ramón Areces” for his postdoctoral fellowship. Q.H. is sponsored by Student Achievement Guided by Experience (SAGE) Program at UC Berkeley.

Figure legends

Figure 1. MEV pathway genes and expression control on three different clusters for limonene production in *E. coli* (a) Nine different genes are necessary to produce the monoterpene limonene from glucose and were organized in three different clusters (CL) (colored-boxes), described in Table 1 (b) Three different limonene-producing strains were constructed using different copy numbers and promoter strengths driving the expression of the different gene clusters. Relative expression level is represented by the size of the colored-boxes.

Figure 2. Specific limonene production and protein levels of 27 different expression scenarios (3 strains x 3 induction OD's x 3 IPTG concentrations) ordered from the lowest to the highest producer. Far left column on the x-axis represents limonene production while the remaining columns represent protein level for each of the nine genes in Figure 1. Labels on the y-axis indicate the scenario: i.e. the strain identity (2p, B or A), the OD_{600nm} level at induction (H, M or L for High, Medium or Low, respectively) and the IPTG induction level (h, m or l for high, medium or low, respectively). For example, B-Ml (the best scenario for production of limonene) refers to strain B induced at medium OD_{600nm} with low IPTG induction level. A 2-D plot of the same data can be found in Supplementary Figure S1.

Figure 3. Principal component analysis (PCA) of data from Figure 2. (a) Principal components (PC) 1 and 2, explaining 92% of the variance (see (b)), are plotted. The size of

the cross indicates the amount of production and the position on the PCA space is a function of the relative protein levels for each of the 9 enzymes in the pathway (b) Contribution of each PC to the total variance. (c) Contribution of each enzyme to the most informative components (i.e. PC1 and PC2)

Figure 4. (a) The new expression system (BL), based on the PCA results of Figure 3, expresses higher terpene synthase levels on the context of a balanced and moderate expression of the rest of the pathway enzymes (b) Improvement in production of BL-Mm compared with the previous best (B-MI) (c) PCAP results of the original data and the new three scenarios. New BL scenarios are in red color. The PCAP result supports the hypothesis suggested by the original PCAP result

Figure 5. (a) MEV genes for bisabolene production (b) Strains for bisabolene production (c) Production of bisabolene from three strains at three IPTG levels. New strains (C and CB strains) based on the optimization of PY strain (Peralta-Yahya et al., 2011) using PCAP of the limonene production in Figure 4 doubled the production.

Figure 6. PCAP analysis of different bisabolene producing systems (a) Specific bisabolene production and proteomics of 9 different expression scenarios (Figure 5; 3 strains x 3 IPTG) ordered from the lowest to the highest producer. Far left column on the x-axis represents bisabolene production while the remaining columns represent protein expression for each of the nine genes in the pathway. (b) Principal component analysis (PCA) of data. Principal component (PC) 1 explains by itself 97% of the variance (see (c), and it is mainly explained by AgBIS levels (60%, see (d)). The size of the cross indicates the amount of production and the position on the PCA space is a function of the relative protein levels for each of the 9 enzymes in the pathway (c) Contribution of each PC to the total variance. (d) Contribution of each enzyme to the most informative components (i.e. PC1 and PC2)

Supplementary Material;

Supplementary Figure S1. Two-dimensional version of Figure 1. Same description apply

Supplementary Figure S2. PCAP analysis of strains BH and AL in comparison to A, B and BL

References

- Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2, 433–459. doi:10.1002/wics.101
- Ajikumar, P.K., Xiao, W.-H., Tyo, K.E.J., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T.H., Pfeifer, B., Stephanopoulos, G., 2010. Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. Science 330, 70–4. doi:10.1126/science.1191652
- Alonso-Gutierrez, J., Chan, R., Batth, T.S., Adams, P.D., Keasling, J.D., Petzold, C.J., Lee, T.S., 2013. Metabolic Engineering of *Escherichia coli* for Limonene and Perillyl Alcohol Production. Metab. Eng. 19, 33–41. doi:10.1016/j.ymben.2013.05.004
- Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. U. S. A. 97, 10101–6.
- 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. doi:10.1186/1754-1611-4-1 [pii] 10.1186/1754-1611-4-1
- Anthony, J.R., Anthony, L.C., Nowroozi, F., Kwon, G., Newman, J.D., Keasling, J.D., 2009. Optimization of the mevalonate-based isoprenoid biosynthetic pathway in *Escherichia coli* for production of the anti-malarial drug precursor amorpha-4,11-diene. Metab. Eng. 11, 13–19.
- Atsumi, S., Hanai, T., Liao, J.C., 2008. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature 451, 86–U13. doi:10.1038/Nature06450
- 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. *in press*. doi:10.1016/j.ymben.2014.08.004
- Bokinsky, G., Peralta-Yahya, P.P., George, A., Holmes, B.M., Steen, E.J., Dietrich, J., Lee, T.S., Tullman-Ercek, D., Voigt, C.A., Simmons, B.A., Keasling, J.D., 2011. Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 108, 19949–19954.

- Burke, C., Croteau, R., 2002. Geranyl diphosphate synthase from *Abies grandis*: cDNA isolation, functional expression, and characterization. *Arch. Biochem. Biophys.* 405, 130–136.
- Cangelosi, R., Goriely, A., 2007. Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* 2, 2. doi:10.1186/1745-6150-2-2
- Crescenzi, M., Giuliani, A., 2001. The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. *FEBS Lett.* 507, 114–8.
- Dahl, R.H., Zhang, F., Alonso-Gutierrez, J., Baidoo, E., Batth, T.S., Redding-Johanson, A.M., Petzold, C.J., Mukhopadhyay, A., Lee, T.S., Adams, P.D., Keasling, J.D., 2013. Engineering dynamic pathway regulation using stress-response promoters. *Nat. Biotechnol.* 1–10. doi:10.1038/nbt.2689
- Dellomonaco, C., Clomburg, J.M., Miller, E.N., Gonzalez, R., 2011. Engineered reversal of the β -oxidation cycle for the synthesis of fuels and chemicals. *Nature* 476, 355–9. doi:10.1038/nature10333
- Di Ventura, B., Lemerle, C., Michalodimitrakis, K., Serrano, L., 2006. From in vivo to in silico biology and back. *Nature* 443, 527–33. doi:10.1038/nature05127
- Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A. V, Prather, K.L.J., Keasling, J.D., 2009. Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.* 27, 753–9. doi:10.1038/nbt.1557
- Francis, D.M., Page, R., 2010. Strategies to optimize protein expression in *E. coli*. *Curr. Protoc. Protein Sci.* Chapter 5, Unit 5.24.1–29. doi:10.1002/0471140864.ps0524s61
- Goh, E.B., Baidoo, E.E.K., Keasling, J.D., Beller, H.R., 2012. Engineering of Bacterial Methyl Ketone Synthesis for Biofuels. *Appl. Environ. Microbiol.* 78, 70–80.
- Ham, T.S., Dmytriv, Z., Plahar, H., Chen, J., Hillson, N.J., Keasling, J.D., 2012. Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Res.* 40, e141. doi:10.1093/nar/gks531
- Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., Fedoroff, N. V, 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. U. S. A.* 97, 8409–14. doi:10.1073/pnas.150242097
- Holtz, W.J., Keasling, J.D., 2010. Engineering static and dynamic control of synthetic pathways. *Cell* 140, 19–23. doi:10.1016/j.cell.2009.12.029
- Hyatt, D.C., Youn, B., Zhao, Y., Santhamma, B., Coates, R.M., Croteau, R.B., Kang, C., 2007. Structure of limonene synthase, a simple model for terpenoid cyclase catalysis. *Proc. Natl. Acad. Sci. U. S. A.* 104, 5360–5. doi:10.1073/pnas.0700915104

- 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.
- Keasling, J.D., 2010. Manufacturing molecules through metabolic engineering. *Science* 330, 1355–8. doi:10.1126/science.1193990
- Lee, J.W., Na, D., Park, J.M., Lee, J., Choi, S., Lee, S.Y., 2012. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.* 8, 536–46. doi:10.1038/nchembio.970
- Lee, T.S., Krupa, R.A., Zhang, F., Hajimorad, M., Holtz, W.J., Prasad, N., Lee, S.K., Keasling, J.D., 2011. BglBrick vectors and datasheets: A synthetic biology platform for gene expression. *J. Biol. Eng.* 5, 12. doi:10.1186/1754-1611-5-12
- Maier, T., Güell, M., Serrano, L., 2009. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–73. doi:10.1016/j.febslet.2009.10.036
- Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., Keasling, J.D., 2003. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.* 21, 796–802. doi:10.1038/Nbt833
- Misra, J., Schmitt, W., Hwang, D., Hsiao, L.-L., Gullans, S., Stephanopoulos, G., Stephanopoulos, G., 2002. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res.* 12, 1112–20. doi:10.1101/gr.225302
- Moon, T.S., Dueber, J.E., Shiue, E., Prather, K.L.J., 2010. Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *E. coli*. *Metab. Eng.* 12, 298–305. doi:10.1016/j.ymben.2010.01.003
- Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.-A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P., Endy, D., 2013. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* 10, 354–60. doi:10.1038/nmeth.2404
- Özaydın, B., Burd, H., Lee, T.S., Keasling, J.D., 2013. Carotenoid-based phenotypic screen of the yeast deletion collection reveals new genes with roles in isoprenoid production. *Metab. Eng.* 15, 174–83. doi:10.1016/j.ymben.2012.07.010
- Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 559–572.
- Peralta-Yahya, P.P., Ouellet, M., Chan, R., Mukhopadhyay, A., Keasling, J.D., Lee, T.S., 2011. Identification and microbial production of a terpene-based advanced biofuel. *Nat. Commun.* 2, 483. doi:10.1038/ncomms1494
- Peralta-Yahya, P.P., Zhang, F., del Cardayre, S.B., Keasling, J.D., 2012. Microbial engineering for the production of advanced biofuels. *Nature* 488, 320–8. doi:10.1038/nature11478

- Pitera, D.J., Paddon, C.J., Newman, J.D., Keasling, J.D., 2007. Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. *Metab. Eng.* 9, 193–207. doi:10.1016/j.ymben.2006.11.002
- Redding-Johanson, A.M., Batth, T.S., Chan, R., Krupa, R., Szmidt, H.L., Adams, P.D., Keasling, J.D., Lee, T.S., Mukhopadhyay, A., Petzold, C.J., 2011. Targeted proteomics for metabolic pathway optimization: Application to terpene production. *Metab. Eng.* 13, 194–203.
- Rossell, S., Solem, C., Jensen, P.R., Heijnen, J.J., 2011. Towards a quantitative prediction of the fluxome from the proteome. *Metab. Eng.* 13, 253–262. doi:10.1016/j.ymben.2011.01.010
- Silva, J.C., Gorenstein, M. V, Li, G.-Z., Vissers, J.P.C., Geromanos, S.J., 2006. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* 5, 144–56. doi:10.1074/mcp.M500230-MCP200
- Stephanopoulos, G., 2012. Synthetic biology and metabolic engineering. *ACS Synth. Biol.* 1, 514–25. doi:10.1021/sb300094q
- Tsuruta, H., Paddon, C.J., Eng, D., Lenihan, J.R., Horning, T., Anthony, L.C., Regentin, R., Keasling, J.D., Renninger, N.S., Newman, J.D., 2009. High-level production of amorphadiene, a precursor of the antimalarial agent artemisinin, in *Escherichia coli*. *PLoS One* 4, e4489. doi:10.1371/journal.pone.0004489
- Wall, M.E., Rechtsteiner, A., Rocha, L.M., 2003. Chapter 5 Singular value decomposition and principal component analysis, in: Berrar, D.P., Dubitzky, W., Granzow, M. (Eds.), *A Practical Approach to Microarray Data Analysis*. Kluwer, Norwell, MA, pp. 91–109.
- Woolston, B.M., Edgar, S., Stephanopoulos, G., 2013. Metabolic engineering: past and future. *Annu. Rev. Chem. Biomol. Eng.* 4, 259–88. doi:10.1146/annurev-chembioeng-061312-103312
- Xu, P., Gu, Q., Wang, W., Wong, L., Bower, A.G.W., Collins, C.H., Koffas, M. a G., 2013. Modular optimization of multi-gene pathways for fatty acids production in *E. coli*. *Nat. Commun.* 4, 1409. doi:10.1038/ncomms2425
- Yadav, V.G., De Mey, M., Lim, C.G., Ajikumar, K., Stephanopoulos, G., 2013. The Future of Metabolic Engineering and Synthetic Biology: Towards a Systematic Practice 14, 233–241. doi:10.1016/j.ymben.2012.02.001
- Zhang, F., Carothers, J.M., Keasling, J.D., 2012. Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat. Biotechnol.* 1–7. doi:10.1038/nbt.2149

Table 1. Description of *E. coli* base strains, plasmids and gene clusters used in this study.

<i>E. coli</i> strain	Description	Reference(s)
DH1	F- endA1 recA1 gyrA96 thi-1 glnV44 relA1 hsdR17(rK- mK+) λ-	Hanahan, 1983
DH10B	F- endA1 recA1 galE15 galK16 nupG rpsL ΔlacX74 Φ80lacZΔM15 araD139 Δ(ara,leu)7697 mcrA Δ(mrr-hsdRMS-mcrBC) λ ⁻	Casabian & Cohen, 1980
2p	DH1 harboring plasmids pBbA5c-CL1 ^a -T1 ^b -CL2 ^c + pTrc ^d -CL3 ^e	(Alonso-Gutierrez, et al., 2013)
B	DH1 harboring plasmid pBbA5c-CL1-T1-CL2-T21 ^f -CL3	(Alonso-Gutierrez, et al., 2013)
A	DH1 harboring plasmid pBbE1a-CL3-CL1-T1-CL2	(Alonso-Gutierrez, et al., 2013)
AL	DH1 harboring plasmid pBbE1a-CL3-CL1-T1-CL2 + pBbB1c-LS	This study
BL	DH1 harboring plasmids pBbA5c-CL1-T1-CL2-T21 ^f -CL3 + pTrc-LS	This study
BH	DH1 harboring plasmids pBbA5c-CL1-T1-CL2-T21 ^f -CL3 + pBbE1a-HMGR-HMGS	This study
PY	DH1 harboring plasmids pBbA5c-CL1 ^g -T1-CL2 ^h + pTrc-AgBIS ⁱ	(Peralta-Yahya et al., 2011)
C	DH1 harboring plasmid pBbA5c-CL1 ^j -T61 ^j -CL2 ^j -T21-AgBIS	This study
CB	DH1 harboring plasmids pBbA5c-CL1 ^j -T61 ^j -CL2 ^j -T21-AgBIS + pTrc-AgBIS ⁱ	This study

Plasmid name	Plasmid reference	Description (Origin of replication, Antibiotic marker, Promoters and Clusters)	Reference(s)
pBbA5c-CL1 ^a -T1 ^b -CL2 ^c	JPUB_002462	p15A, Cm ^r , PlacUV5, CL1, T1, CL2	(Alonso-Gutierrez, et al., 2013)
pTrc ^d -CL3 ^e	JPUB_002464	colE1, Amp ^r , Ptrc, CL3	(Alonso-Gutierrez, et al., 2013)
pBbA5c-CL1-T1-CL2-T21 ^f -CL3	JPUB_002470	p15A, Cm ^r , PlacUV5, CL1, T1, CL2, T1002, Ptrc, CL3	(Alonso-Gutierrez, et al., 2013)
pBbE1a-CL3-CL1-T1-CL2	JPUB_002471	colE1, Amp ^r , Ptrc, MTSA, T1, MBI-f, T1002, Ptrc, trGPPS, LS	This study
pTrc-LS	JPUB_002473	colE1, Amp ^r , Ptrc, LS	This study
pBbA5c-CL1 ^g -T1-CL2 ^h	JPUB_002460	p15A, Cm ^r , PlacUV5, CL1 ^j , T1, CL2 ^j	(Peralta-Yahya et al., 2011)
pTrc-AgBIS ⁱ	JPUB_002466	colE1, Amp ^r , Ptrc, AgBIS	This study
pBbA5c-CL1 ^j -T61 ^j -CL2 ^j -T21-AgBIS	JPUB_002475	p15A, Cm ^r , PlacUV5, CL1 ^j , T1, CL2 ^j , T1002, Ptrc, AgBIS	This study
pBbE1a-HMGR-HMGS	JPUB_002468	colE1, Amp ^r , Ptrc, HMGR and HMGS from <i>S. aureus</i>	This study
pBbB1c-LS	JPUB_002478	pBBR1, Cm ^r , Ptrc, LS	This study

^a CL1: CLUSTER 1 contains genes for the conversion of acetyl CoA to mevalonate: acetoacetyl-CoA synthase from *E. coli* (AtoB), HMG-CoA synthase from *Staphylococcus aureus* (HMGS), and HMG-CoA reductase from *Staphylococcus aureus* (HMGR) (Tsuruta et al., 2009) codon optimized for *E. coli*

^b T1: double terminator followed by Ptrc promoter

^c CL2: CLUSTER 2 contains genes for the conversion of mevalonate to isoprenoid precursors (IPP and DMAPP): Mevalonate Kinase (MK), PhosphoMevalonate Kinase (PMK), Mevalonate PyrophosphoMevalonate Decarboxylase from *S. cerevisiae* (PMD) and isopentenyl diphosphate isomerase from *E. coli* (idi) (Martin et al., 2003). These genes have been fixed for expression in *E. coli* as previously described (Redding-Johanson et al., 2011)

^d pTrc: Plasmid backbone pTRC99

^e CL3: CLUSTER 3 contains truncated sequence of Geranyl PyroPhosphate Synthase from *Abies grandis* (Burke and Croteau, 2002) codon optimized (co)^c sequence for *E. coli* (tr-GPPS(co)) and LS: Limonene synthase from *Mentha spicata* (truncated version without plastidial targeting sequence (Hyatt et al., 2007) and codon optimized for *E. coli*)

^f T21: Terminator T1002 followed by Ptrc promoter

^g CL1^j: CLUSTER 1^j contains genes for the conversion of acetyl CoA to mevalonate: acetoacetyl-CoA synthase from *E. coli* (AtoB), HMG-CoA synthase from *Saccharomyces cerevisiae* (HMGS), and HMG-CoA reductase from *Saccharomyces cerevisiae* (HMGR) (Tsuruta et al., 2009) codon optimized for *E. coli*

^h CL2^j: CLUSTER 2^j contains genes for the conversion of mevalonate to isoprenoid precursors (IPP and DMAPP): Mevalonate Kinase (MK), PhosphoMevalonate Kinase (PMK), Mevalonate PyrophosphoMevalonate Decarboxylase from *S. cerevisiae* (PMD) and isopentenyl diphosphate isomerase (idi) and farnesyl diphosphate synthase (ispA) from *E. coli* (idi) (Martin et al., 2003). These genes have been fixed for expression in *E. coli* as previously described (Redding-Johanson et al., 2011)

ⁱ AgBIS: Bisabolene synthase coding sequence from *Abies grandis* codon-optimized for *E. coli*

^j T61: Terminator T1006 followed by Ptrc promoter

Figure

Fig 1

Fig 2

Fig 3

4. bio.ii

Fig 5

60
F