the plant journal

SEB Society for Experimental Biology

The Plant Journal (2016) **87,** 455–471 doi: 10.1111/tpj.13210

Identification of line-specific strategies for improving carotenoid production in synthetic maize through data-driven mathematical modeling

Jorge Comas[1,2,3], Rui Benfeitas[4,5], Ester Vilaprinyo[1,2], Albert Sorribas[1,2], Francesc Solsona[3], Gemma Farré[6], Judit Berman[6], Uxue Zorrilla[6], Teresa Capell[6], Gerhard Sandmann[7], Changfu Zhu[6], Paul Christou[6,8] and Rui Alves[1,2,*]

[1]*Departament de Ciencies Mèdiques Bàsiques, Universitat de Lleida, Lleida, Spain,*

[2]*Institut de Recerca Biomèdica de Lleida IRBLLeida, Edifici de Recerca Biomédica I, Av Rovira Roure 80, Lleida, Catalunya 25198, Spain,*

[3]*Computer Science Department and INSPIRES, University of Lleida, Jaume II 69, Lleida, Catalunya 25001, Spain,*

[4]*Center for Neuroscience and Cell Biology, University of Coimbra, Coimbra 3004-517, Portugal,*

[5]*Institute for Interdisciplinary Research, University of Coimbra, Coimbra 3030-789, Portugal,*

[6]*Department of Plant Production and Forestry Science, School of Agrifood and Forestry Science and Engineering (ETSEA), University of Lleida Agrotecnio Center, Avenida Alcalde Rovira Roure 191, Lleida 25198, Spain,*

[7]*Institute of Molecular Bioscience, J. W. Goethe University, Max von Laue Strasse 9, Frankfurt am Main D-60438, Germany, and*

[8]*ICREA, Institució Catalana de Recerca i Estudis Avancats, Passeig Lluís Companys, 23, 08010 Barcelona, Spain*

Received 5 December 2015; revised 25 April 2016; accepted 29 April 2016; published online 18 July 2016.
*For correspondence (e-mail ralves@cmb.udl.cat).

SUMMARY

Plant synthetic biology is still in its infancy. However, synthetic biology approaches have been used to manipulate and improve the nutritional and health value of staple food crops such as rice, potato and maize. With current technologies, production yields of the synthetic nutrients are a result of trial and error, and systematic rational strategies to optimize those yields are still lacking. Here, we present a workflow that combines gene expression and quantitative metabolomics with mathematical modeling to identify strategies for increasing production yields of nutritionally important carotenoids in the seed endosperm synthesized through alternative biosynthetic pathways in synthetic lines of white maize, which is normally devoid of carotenoids. Quantitative metabolomics and gene expression data are used to create and fit parameters of mathematical models that are specific to four independent maize lines. Sensitivity analysis and simulation of each model is used to predict which gene activities should be further engineered in order to increase production yields for carotenoid accumulation in each line. Some of these predictions (e.g. increasing *Zmlycb/Gllycb* will increase accumulated β-carotenes) are valid across the four maize lines and consistent with experimental observations in other systems. Other predictions are line specific. The workflow is adaptable to any other biological system for which appropriate quantitative information is available. Furthermore, we validate some of the predictions using experimental data from additional synthetic maize lines for which no models were developed.

Keywords: *Zea mays*, synthetic biology, systems biology, mathematical modeling, computational biology, metabolomics.

INTRODUCTION

Synthetic biology utilizes known molecular components and genes, modifying and/or combining them in new ways, with the aim of implementing different molecular circuits displaying novel functions and dynamic behavior that does not occur naturally (Balagaddé *et al.*, 2008; Bacchus *et al.*, 2013; Nielsen and Voigt, 2014; Way *et al.*, 2014). This discipline also provides the means to modify organisms and make them produce useful chemicals that are not present in their normal metabolism (e.g. DeLoache *et al.*, 2015).

Arguably, the most spectacular applications of synthetic biology for production purposes use microbes as the

substrate for organism modification. For example, *Saccharomyces cerevisiae* has been engineered to produce antimalarial drugs (Paddon *et al.*, 2013) or narcotics (DeLoache *et al.*, 2015), and *Escherichia coli* has been engineered to produce biofuels (Rahmana *et al.*, 2014). Other applications of synthetic biology that are likely to have a substantial impact in the short to medium term involve engineering new lines of nutritionally improved and widely used food staples. Examples include Golden Rice (Ye *et al.*, 2000), multivitamin corn (Zhu *et al.*, 2008) or other plants (Farré *et al.*, 2014) that constitute the basis of human diets across the globe.

There is ongoing debate on the regulatory and ethical aspects of organisms modified through synthetic biology (LaVan and Marmon, 2010; Adam *et al.*, 2011; Masip *et al.*, 2013). In addition, and given the complexity of genomes and gene regulation in plants, the basic scientific aspects of rational plant engineering are still less well developed than those of microbe engineering (Farré *et al.*, 2014). For example, understanding how pathways are systemically driven by and interact with the native metabolism after they have been engineered into their hosts is challenging, as is identifying the production bottlenecks in these pathways. If one can understand these aspects of new and/or engineered pathways, one can rationally further improve their production yields and the nutritional value of the staple crops. Mathematical modeling and analyses are important tools for achieving that understanding, because they can be used to predict how further modification of the pathways will affect pathway behavior (Atkinson *et al.*, 2003; Brophy and Voigt, 2014; Uzkudun *et al.*, 2015).

South African (SA) white maize is one of the staple crops that have been engineered to improve its nutritional value. This plant was engineered to produce carotenoid vitamins, which are almost absent in wild-type maize (Zhu *et al.*, 2008). Carotenoids are a broad group of organic tetraterpenoid pigments synthesized by plants, bacteria and fungi (Li *et al.*, 2010; Bai *et al.*, 2011; Berman *et al.*, 2014). The biosynthetic pathway responsible for their production is summarized in Figure 1. This pathway is described in more detail in the Background section of Appendix S1 in the Supporting Information. Animals obtain these carotenoids from their diets. In herbivores and omnivores (including humans), carotenoids act as metabolic precursors and antioxidants, some of which have specific health benefits. For example, pro-vitamin A carotenoids, the most important of which is β-carotene, act as precursors of vitamin A which is essential for developing and maintaining healthy vision in mammals (Zhu *et al.*, 2008; Bai *et al.*, 2011), among other health benefits (e.g. protection against various cancers; Basu and Imrhan, 2007).

Here, we focused on how the engineered biosynthesis of carotenoids is systemically driven at the molecular level in four synthetic lines derived from SA white maize (Zhu *et al.*, 2008). We also investigated if accumulation of carotenoids in the maize endosperm (the edible part of the

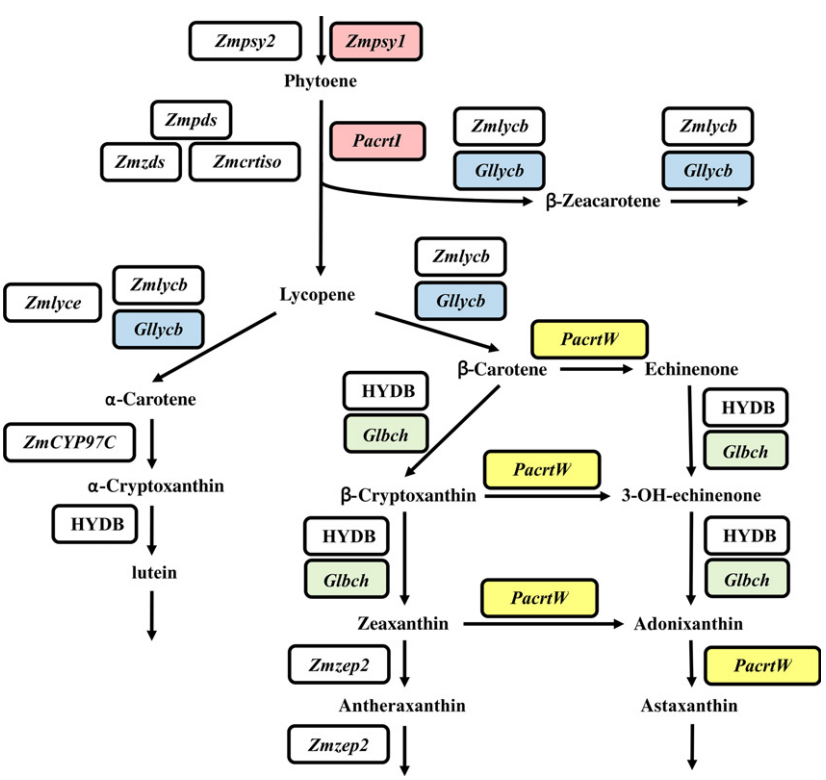

Figure 1. The general carotenoid biosynthetic pathway, starting with phytoene as its substrate.
Several intermediates are known to be synthesized between phytoene and lycopene. However, none of these intermediates is produced in sufficiently high amounts to be detected in our experiments. Because of this we omitted those steps in our representation. Lycopene can be used to produce either α- or β-carotenoids. β-Carotenoids can also be used to synthesize ketocarotenoids, the rightmost vertical branch in the figure. The genes that code for enzymes involved in catalyzing each of the steps are shown in boxes next to the relevant reaction. HYDB represents a β-carotene hydroxylase activity that can be catalyzed by enzymes encoded by genes *Zmbch1*, *Zmbch2*, *ZmCYP97A* or *ZmCYP97B*. White boxes encircle the genes that are expressed in the endosperm of the wild-type white maize. Red boxes encircle transgenes expressed in the endosperm of Lines 1–4. Blue boxes encircle transgenes expressed in Lines 2–4. Green boxes encircle transgenes expressed in Lines 3 and 4. Yellow boxes encircle transgenes expressed only in Line 4.

plant) may be increased further. To answer these questions, we employed a systems biology approach combining gene expression and metabolomics with data-driven mathematical modeling. The workflow is summarized in Figure 2. We used the experimental data to generate four models that were independently analyzed to understand the dynamics of carotenogenesis in the transgenic maize lines. As a final step in the analysis we identified which steps in the pathways should be targeted to more effectively control the production of individual and total carotenoids in transgenic maize. Finally, we validated some of these predictions using experimental data for four additional synthetic maize lines for which no models were developed.

RESULTS

Generating time series for gene expression and metabolite accumulation

Four lines of transgenic maize containing different combinations of carotenogenic genes were chosen for this study. The carotenoid biosynthesis pathways for the four lines are summarized in Figure S1 and the lines are further described in Zhu *et al.* (2008). The transcript levels of each of the 12 most relevant endogenous carotenogenic genes together with five carotenogenic transgenes were quantified using real-time quantitative (q)RT-PCR in the endosperm (the edible part of the maize kernel) of each maize line at 15, 20, 25 and 30 days after pollination (DAP). The amounts of RNA in the endosperm were below detection levels outside this time window because of the temporal expression of the promoters used to drive expression of the transgenes. The amounts of individual carotenoids produced in each line were measured at 15, 20, 25, 30, 40, 50 and 60 DAP.

The gene expression and metabolomics time series were interpolated to calculate the levels of metabolites and transcripts at time points between measurements (see Experimental Procedures). This is an approach that can be used whenever the variation between consecutive observed data points is sufficiently smooth and the error in experimental determination is low (Baud *et al.*, 1991; Buzzi-Ferraris and Manenti, 2010; Farré *et al.*, 2013).

Overall, this approach allowed us to obtain sufficiently populated time series that permitted the use of model optimization and fitting tools to create data-driven ordinary

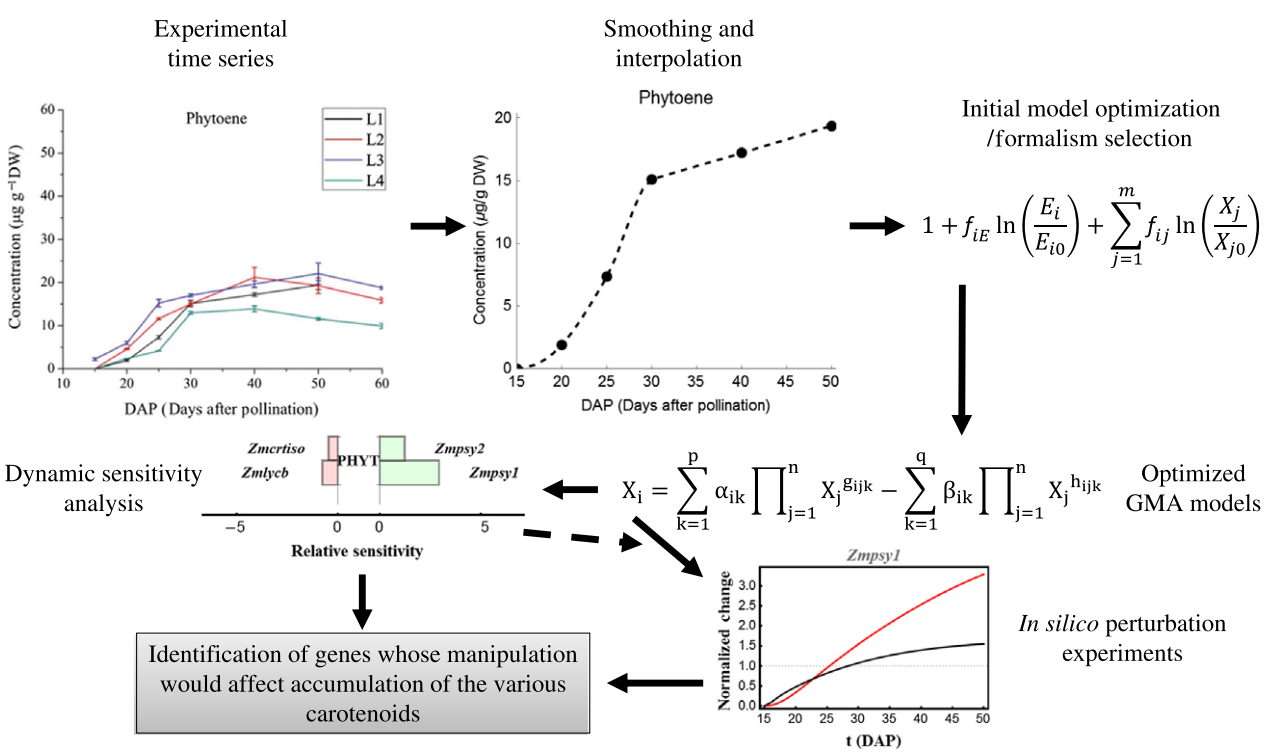

Figure 2. Workflow for data-driven modeling of carotenoid biosynthesis in modified maize lines.
For each line, experimental data were collected and used to interpolate dense time series of gene expression and metabolite concentrations. These time series were used to estimate the parameter values for an optimized mathematical model that best fits the experimental results for that line. A sensitivity analysis of the optimized model identified the genes whose overexpression was most likely to increase the accumulation of each metabolite. The expression of these genes was doubled and simulations were run to determine the effects of that doubling on metabolite concentration.

differential equation (ODE) models of carotenoid biosynthesis in each of the four maize lines.

Multilevel models of the four synthetic maize lines

We used the gene expression and metabolite level time series to build a model in which changes in gene expression drive changes in metabolite levels, but not vice versa. Because of how the transgenes were designed for the synthetic maize lines it is justified to assume that transgene expression is not driven by carotenoid levels to a significant extent. This is because the transgenes were inserted under the control of the endosperm-specific low-molecular weight glutenin or barley D-hordein promoters (Naqvi *et al.*, 2009; Farré *et al.*, 2013). It is very difficult to measure carotenoid enzymatic activity and protein levels because the proteins are located in the membrane and thus difficult to solubilize. However, there is evidence to show that the transcript level is directly correlated with the level of accumulated carotenoids in a number of flowers from various plant systems, as reviewed in Zhu *et al.* (2010). In the absence of further data specific to maize endosperm, this observation is consistent with assuming that the transcript level is positively correlated with enzyme activity; this is what we do in order to link the gene expression and metabolite levels. Additional work in *Medicago truncatula* seeds found that at least 50% of all genes have a strong positive correlation ($R^2 > 0.9$) between transcript levels and protein levels (Gallardo *et al.*, 2007). This number increases to 72% if we accept a positive correlation and an $R^2 > 0.5$, as can be calculated from Table S4 of Gallardo *et al.* (2007).

At 30 DAP gene expression levels go below detection limits. This means that we need to make a further assumption about how the enzyme activity changes between 30 and 60 DAP. Figure 2 in Méchin *et al.* (2007) shows that between 30 and 40 DAP, on average, enzyme activity in secondary metabolism in maize endosperm undergoes a linear decay of 1% per day. A similar dynamic behavior occurs in *M. trunculata* seeds (see, for example, Figure 1B of Gallardo *et al.*, 2007) Because all genes in the model code for metabolic enzymes from secondary metabolism, we assumed that transcript abundance undergo a 30% linear decay between 30 and 60 DAP.

To model the gene expression level we further assumed that the transcriptional rate (*TR*) of a given carotenogenic gene *G* is constant at any given developmental stage in the endosperm. This assumption is consistent with whole-transcriptome analysis that finds gene expression activity to be well correlated with the developmental stage of the maize endosperm (Chen *et al.*, 2014; Li *et al.*, 2014). Specifically the literature states that there is strong transcriptomic reprogramming in developing endosperm, which is mainly attributable to drastic changes in the early and late stages. This could explain why a piecewise approximation with one breakpoint is sufficient to model gene expression in our case (see Methods in Appendix S1). In addition we assume that the mRNA degradation rate follows first-order kinetics (Equation 1), as reported for higher plants (Green, 1993; Lambein, 2003). Initially we also created alternative models using other formalisms with more parameters (Alves *et al.*, 2008). However, a statistical analysis using the Bayesian information criterion (BIC) metric indicated that such models were not superior to the linear model described by Equation 1 (see Methods in Appendix S1):

$$\frac{dG}{dt} = TR - k \times G \quad (1)$$

Here *TR* represents the transcription rate for the gene and is assumed to be constant. The rate constant *k* is also constant, except where noted (see 'Model Building and Optimization' in Experimental Procedures). Given that we had no information about the regulation of expression for the various genes, we used Equation 1 to represent the overall gene expression dynamics of any given gene as a piecewise defined function (see Experimental Procedures). This approach allowed us to phenomenologically account for the various shifts in gene expression levels that are known to occur as endosperm goes through its various developmental stages in maize and other plants (Fraser *et al.*, 1994; Ghassemi-Golezani *et al.*, 2011).

To model the carotenoid accumulation level we used a generalized mass action (GMA) representation (Equation 2; Sorribas and Savageau, 1989). Again, we note that initially we also created alternative models using other formalisms with more parameters (Alves *et al.*, 2008). However, a statistical analysis using the BIC metric indicated that such models were no better than the power-law model described by Equation 2 (see Methods in Appendix S1 for details).

$$\frac{dX_i}{dt} = \sum_{k=1}^{p} \alpha_{ik} \prod_{j=1}^{n} X_j^{g_{ijk}} \sum_{k=1}^{q} \beta_{ik} \prod_{j=1}^{n} X_j^{h_{ijk}} \quad (2)$$

where X_i represents the metabolite of interest, X_j represents all individual metabolites or genes involved in the production or consumption of X_i, α_{ik} represents apparent production rate constants, β_{ik} represents apparent consumption rate constants, g_{ijk} represents apparent kinetic orders in production reactions and h_{ijk} represents apparent kinetic orders in consumption reactions. Apparent rate constants and apparent kinetic orders are the parameters that can be estimated from fitting the model to the experimental data.

Although this representation is simple it is still capable of capturing non-linearities in the dynamic behavior of the system being modeled. In addition, the number of parameters (α, β, *g* and *h*) to be estimated is typically lower than that for other non-linear models that can be created for the same system (Voit, 2013).

The full version of the carotenogenic pathway we used to anchor our model is depicted in Figure 1. This general pathway is then adapted to describe the dynamics of gene expression and metabolite levels in each maize line. The line-specific pathways are shown in Figure S1. Tables S1 and S2 indicate which genes in each line encode for enzymes that catalyze each of the reactions. We then created four independent mathematical models to describe the dynamic behavior of each pathway, using systems of ODEs in GMA format (see Experimental Procedures and the Methods in Appendix S1). Each model is then optimized to infer the parameters that best fit the experimental data, as described in Experimental Procedures.

At the gene expression level, the optimized models fit the experimental data very well (adjusted $R^2 \geq 0.8$). This indicates that our assumptions are consistent with the experimental data for transcript abundance. Good fits between modeling predictions and metabolite profiles were only possible after incorporating additional biological processes and reactions into the models (see Table 1).

Table 1 List of additional biological processes that were tested to improve the fitting of the models to experimental results

Modification (Type)	Line 1	Line 2	Line 3	Line 4
Inhibition of antheraxanthin by lutein (1)	×	×	–	×
Phytoene piecewise defined (2)	×	✓	×	×
β-Carotene piecewise defined (2)	×	✓	×	×
β-Cryptoxanthin piecewise defined (2)	×	✓	×	✓
Zeaxanthin piecewise defined (2)	✓	✓	×	✓
Antheraxanthin piecewise defined (2)	✓	✓	–	✓
3-OH-echinenone piecewise defined (2)	–	–	×	✓
Adonixanthin piecewise defined (2)	–	–	×	✓
Astaxanthin piecewise defined (2)	–	–	✓	×
β-Carotene extra degradation (3)	–	–	×	✓
β-Cryptoxanthin extra degradation (3)	–	–	✓	×
Echinenone extra degradation (3)	–	–	✓	✓
3-OH-echinenone extra degradation (3)	–	–	✓	×
Adonixanthin extra degradation (3)	–	–	✓	×

A dash (–) indicates modifications that were not attempted in the corresponding model because either the genes that code for the relevant reactions are not present in the lines or the relevant metabolites are below detection limits. A cross (×) indicates modifications that failed to improve the fitting. A check mark (✓) indicates modifications that improve model fitting to experimental results.

The references that support their plausible existence of the attempted modifications are the following: (i) lutein may be a substrate of zeaxanthin epoxidase (García-Plazaola *et al.*, 2007); (ii) carotenoid biosynthesis may undergo shifts in regulation throughout various developmental stages of the plant (Fraser *et al.*, 1994), and our baseline models may fail to capture said shifts; (iii) light exposure may also contribute to carotene degradation (Boon *et al.*, 2010).

These processes and reactions are biologically plausible, as indicated by the cited literature in Table 1. Their inclusion improved the way in which our models fitted the data. Thus, the modeling workflow helped to identify molecular processes that make a significant contribution to shaping the dynamics of carotenoid accumulation in a line-specific manner. These processes were previously unaccounted for in our understanding of the system.

The model's explanation of the metabolite experimental data in Line 1 (see Table 1) improves if one assumes that the parameters for the enzyme activities that produce and consume antheraxanthin are different before and after 22.5 DAP. For Line 2, in addition to antheraxanthin, we assumed a similar situation for phytoene, β-carotene, β-cryptoxanthin and zeaxanthin. For Line 3, we considered that the parameters for the rate of degradation of astaxanthin changed after 25 DAP. The time boundaries for the various piecewise equations come about as a consequence of abrupt changes in the slopes of the metabolic accumulation rate and are automatically determined by the optimization algorithm. The timing is consistent with a change in the developmental stage of maize endosperm that takes place between 20 and 25 DAP. Moreover, additional sink processes for adonixanthin, echinenone, 3-OH-echinenone and β-cryptoxanthin were considered. For Line 4, the parameters for the rate of production and degradation of adonixanthin, 3-OH-echinenone, antheraxanthin and zeaxanthin were allowed to change after 25 DAP. Additionally, we added an extra degradation term for echinenone, 3-OH-echinenone and β-cryptoxanthin.

Additional discussion of the modeling process, together with a visual description of the modeling workflow, is given in the Experimental Procedures. The detailed systems of ODEs for each maize line are given in the Results section of Appendix S1. Graphical representations of how well the models fit the experimental time series for metabolites and transcript levels in each maize line are shown in Figures S2–S18.

We note that the worst fitted metabolite was always the least abundant (α-cryptoxanthin in Line 1, β-cryptoxanthin in Line 2 and 3-OH-echinenone in Line 3). This was because every attempt we made to improve the fit to the dynamics of these lower-abundance metabolites introduced larger errors to the fitting of more abundant metabolites. As a consequence, larger global errors would be introduced in the model. Overall, the final models explained the experimental data for their respective maize lines with high adjusted R^2 values (0.93 for Line 1, 0.79 for Line 2, 0.91 for Line 3 and 0.92 for Line 4). The coefficient of determination, R^2, is a measure for quantifying the goodness of fit of a model. It is defined as the proportion of variability of the data that is explained by the model as a predictor. In linear models it is the popular r^2 (Spiess and Neumeyer, 2010).

Strategies for improving carotenoid production

Given how well the optimized models described the experimental data for each of the maize lines, we assume that they provide a good explanation for the carotenoid metabolism in each lines. Hence, we could use dynamic sensitivity analysis (see Experimental Procedures) to identify strategies for further modulation of gene expression that are expected to lead to improvements in the accumulation of the various metabolites. The dynamic sensitivity of a given metabolic concentration X_i to a specific parameter p_j estimates the change in concentration X_i at time t, if p_j changed its value at the beginning of the simulation. We use relative sensitivities in our analysis. Briefly, positive (negative) sensitivities with a value of $s(t)$ at time t indicate that a change of 1% in the parameter at the initial time will lead to a change of approximately plus (minus) $s(t)$ times in the corresponding metabolite at time t with respect to its value at time t in the original simulation.

In all maize line models, the production of the α-branch of carotenoids (α-cryptoxanthin and lutein) could be best improved by increasing the expression of the *Zmlyce* gene. Increasing the expression of either the *Zmlycb* or the *Gllycb* genes was predicted to improve the accumulation of β-branch carotenoids (β-carotene and β-cryptoxanthin). Accumulation of ketocarotenoids (echinenone, 3-OH-echinenone, adonixanthin and antheraxanthin) was predicted to improve by increasing the expression of the *PacrtW* and *Gllycb* genes. In addition, increasing the expression of *Zmpsy1* in Line 2 was predicted to lead to large increases in accumulated β-carotene. An increase was also predicted in Line 1. However, this increase was too small to be represented in Figure 3. In most cases, the values for the maximum sensitivities indicate that gene expression would need to be raised at least two-fold to achieve a two-fold accumulation of the relevant metabolites. The results of the sensitivity analysis (t = 60 DAP) are summarized in Figure S2 and given in detail in Figures S15–S18 and Tables S3–S10.

Testing non-linear effects of changing expression for single genes

In the previous subsection we presented a differential first-order sensitivity analysis. Such an analysis is accurate for predicting sensitivities to infinitesimal perturbations. We wanted to understand how well these predictions might hold up under realistic finite changes that could propagate non-linearly across the system. This led us to perform *in silico* experiments that implemented finite changes in the expression of relevant genes and measure the effect of those changes in the accumulation of the various carotenoids.

We implemented independent single-gene manipulation experiments in each maize line by using the mathematical model representing that line and individually increasing by two-fold the expression of *Zmpsy1*, *Zmlyce*, *Zmlycb*, *Gllycb* and *PacrtW*. These genes were chosen because

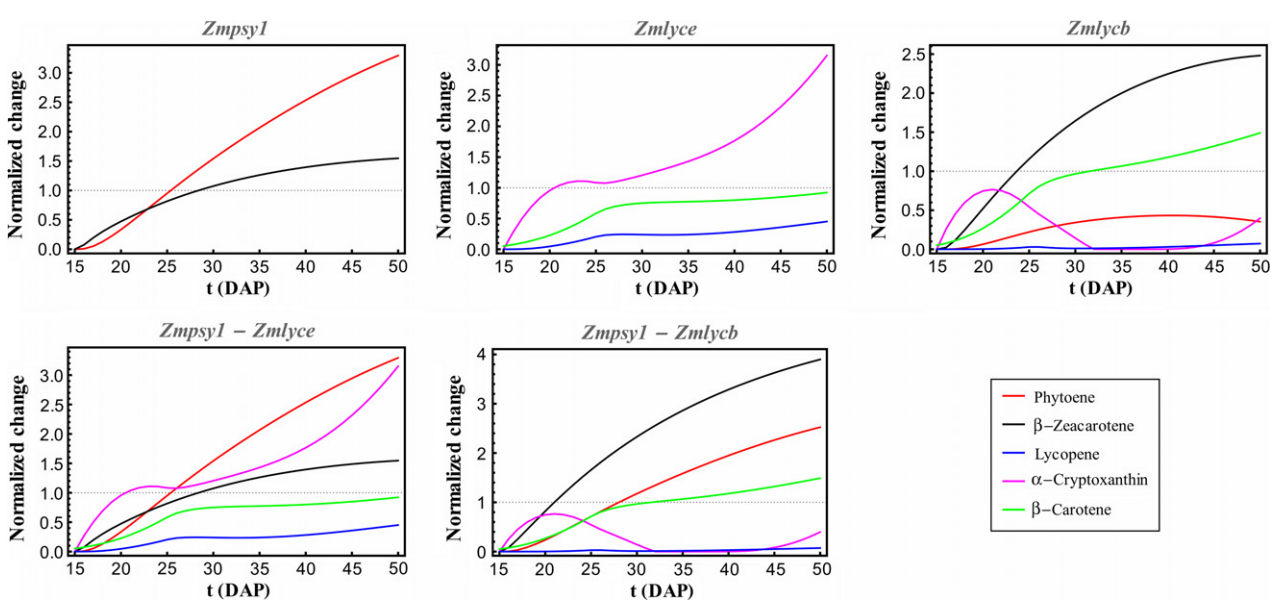

Figure 3. Effect of doubling basal gene expression in Line 1.
The effect of doubling the basal level of gene expression for *Zmpsy1*, *Zmlyce* and *Zmlycb* and the effect of simultaneously doubling the basal level of gene expression for *Zmpsy1* and *Zmlyce* (*Zmpsy1–Zmlyce*) and *Zmpsy1* and *Zmlycb* (*Zmpsy1–Zmlycb*). The *x*-axis in each plot represents time, while the *y*-axis represents the normalized accumulation of carotenoids. Normalization was done by dividing the amount of accumulated metabolite at each time point by the amount of that metabolite which accumulated at 50 days after pollination (DAP) in Line 1. If the curve goes above 1, the metabolite accumulates with respect to Line 1; if it stays below 1, the metabolite is less abundant after changing basal gene expression than it was in Line 1.

sensitivity analysis identified them as coding for the enzymes that are more likely to affect metabolic concentrations upon changing their expression. These experiments identify possible non-linear effects of changing gene expression in the maize lines that our sensitivity analysis might have missed.

The simulations for Line 1 showed that sensitivity analysis predicted quantitatively the finite changes in metabolite concentrations resulting from a two-fold increase in *Zmpsy1* expression (Figure 3). This was also true for *Zmlycb* and *Zmlyce*.

The simulations for Line 2 showed that sensitivity analysis predicted quantitatively the changes caused by a two-fold increase in the expression of either *Zmlycb* or *Gllycb* (Figure 4). Sensitivity analysis also correctly predicted the quantitative changes in metabolite accumulation caused by a two-fold increase in the expression of *Zmpsy1*, except for β-carotene. The accumulation of this metabolite was about three-fold less than predicted by the sensitivity analysis. Similarly, a two-fold increase in the expression of *Zmlyce* caused an accumulation of α-cryptoxanthin that was about three-fold less than expected from the sensitivity analysis. This identified the presence of non-linear effects that cause larger changes than those predicted by the differential sensitivity analysis when doubling the expression of *Zmpsy1* or *Zmlyce*.

The simulations for Lines 3 (Figure 5) and 4 (Figure 6) showed that sensitivity analysis quantitatively predicted most of the changes in metabolite accumulation caused by a two-fold increase in each of the tested genes. One exception is observed when increasing the expression of *Zmlycb* by two-fold in Line 3, which led to a significant decrease in β-carotene, while sensitivity analysis predicted that an increase of between one- and two-fold would occur. Another exception is observed when increasing the expression of *PacrtW* by two-fold in Line 4, which led to a significant decrease in adonixanthin, while sensitivity analysis predicted that only a very slight decrease would occur.

Testing the effects of simultaneous changes in the expression of two genes

To test the effect of simultaneously changing the expression of two genes in the accumulation of β-carotenoids, we performed several double-gene manipulation experiments.

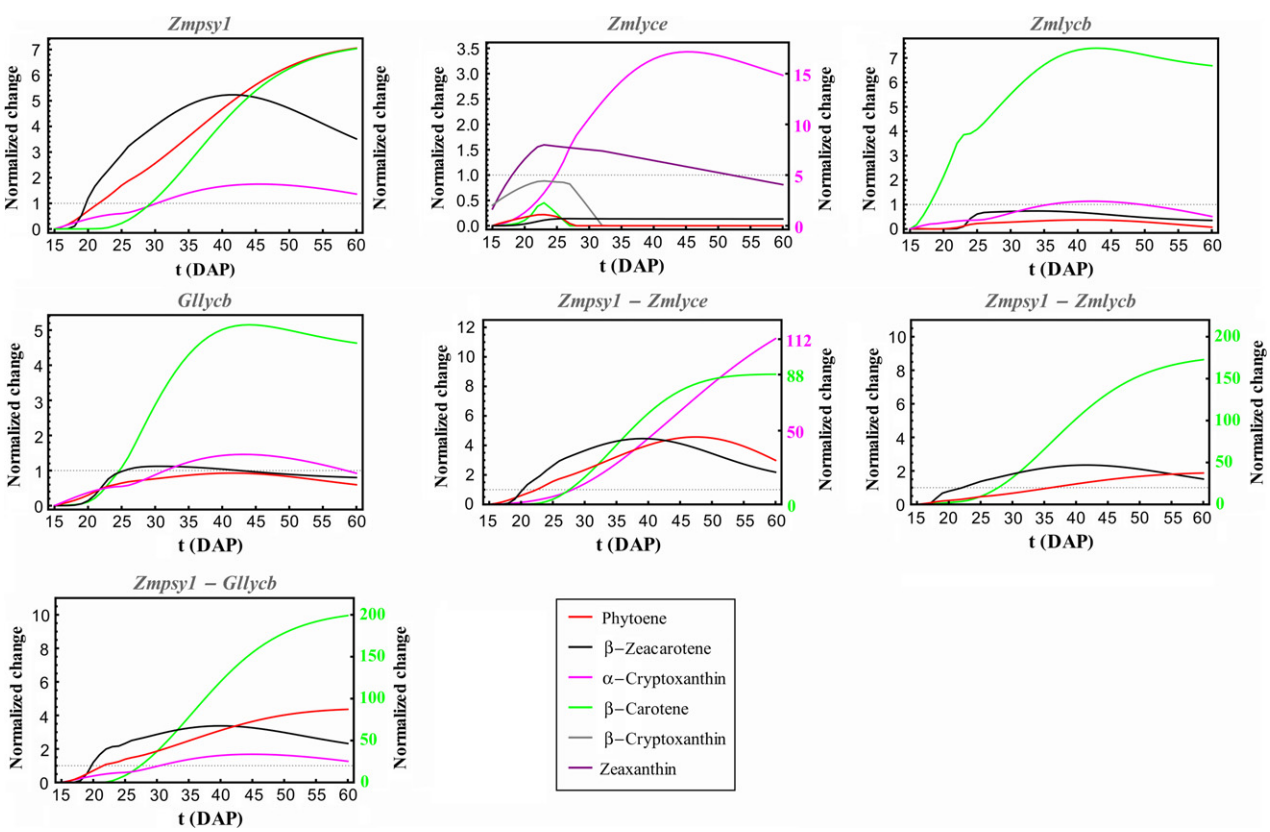

Figure 4. Effect of doubling basal gene expression in Line 2.
The effect of doubling the basal level of gene expression for *Zmpsy1*, *Zmlyce*, *Zmlycb* and *Gllycb* and the effect of simultaneously doubling the basal level of gene expression for *Zmpsy1* and *Zmlyce* (*Zmpsy1*–*Zmlyce*), *Zmpsy1* and *Zmlycb* (*Zmpsy1*–*Zmlycb*) and *Zmpsy1* and *Gllycb* (*Zmpsy1*–*Gllycb*). The *x*-axis in each plot represents time, while the *y*-axis represents the normalized accumulation of carotenoids. Normalization was done by dividing the amount of accumulated metabolite at each time point by the amount of that metabolite which accumulated at 60 days after pollination (DAP) in Line 2. If the curve goes above 1, the metabolite accumulates with respect to Line 2; if it stays below 1, the metabolite is less abundant after changing basal gene expression than it was in Line 2.

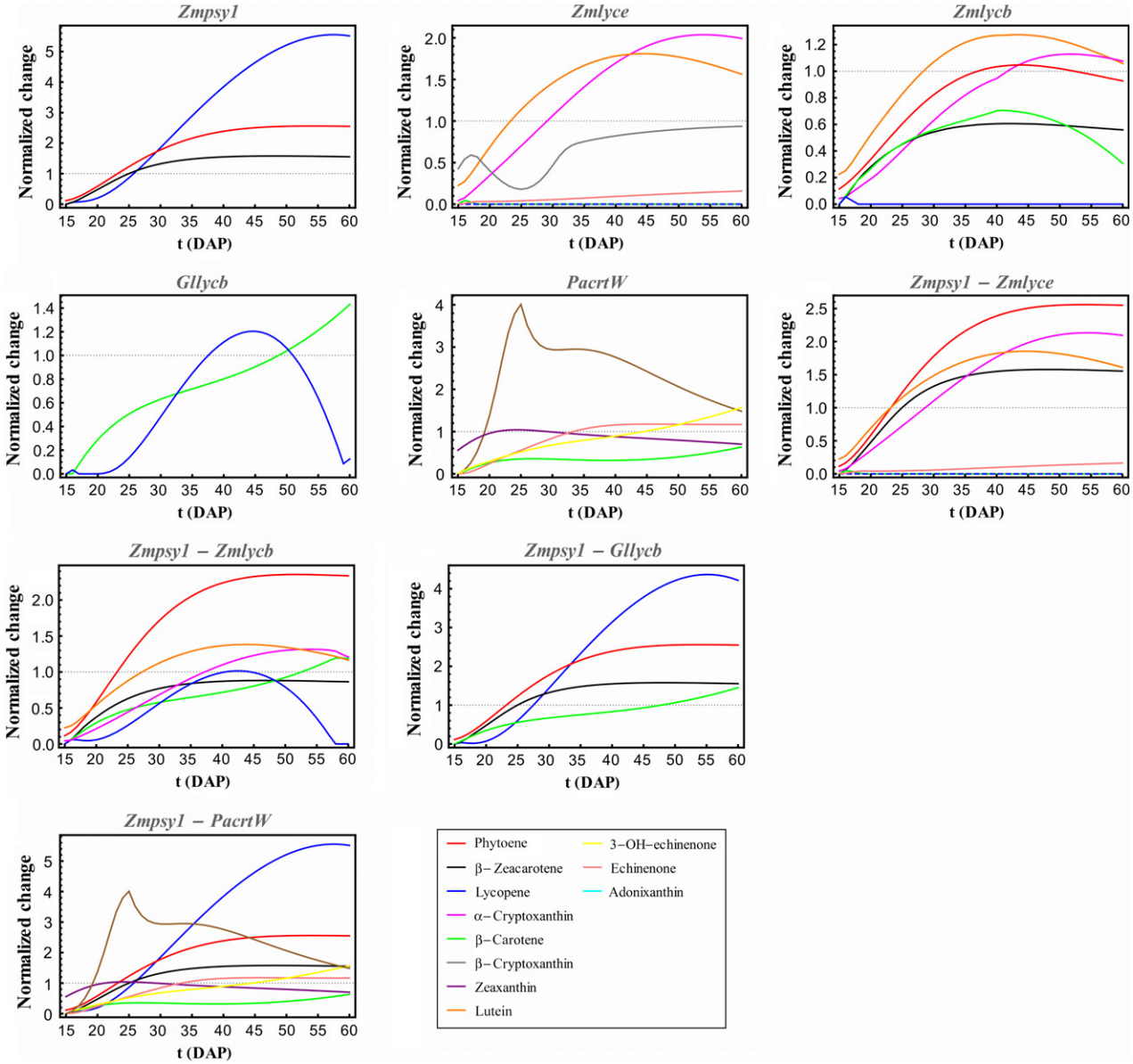

Figure 5. Effect of doubling basal gene expression in Line 3.
The effect of doubling the basal level of gene expression for *Zmpsy1*, *Zmlyce*, *Zmlycb*, *Gllycb* and *PacrtW* and the effect of simultaneously doubling the basal level of gene expression for *Zmpsy1* and *Zmlyce* (*Zmpsy1–Zmlyce*), *Zmpsy1* and *Zmlycb* (*Zmpsy1–Zmlycb*), *Zmpsy1* and *Gllycb* (*Zmpsy1–Gllycb*), and *Zmpsy1* and *PacrtW* (*Zmpsy1–PacrtW*). The *x*-axis in each plot represents time, while the *y*-axis represents the normalized accumulation of carotenoids. Normalization was done by dividing the amount of accumulated metabolite at each time point by the amount of that metabolite which accumulated at 60 days after pollination (DAP) in Line 3. If the curve goes above 1, the metabolite accumulates with respect to Line 3; if it stays below 1, the metabolite is less abundant after changing basal gene expression than it was in Line 3.

We independently increased by two-fold the expression of the gene pairs *Zmpsy1–Zmlyce*, *Zmpsy1–Zmlycb*, *Zmpsy1–Gllycb* and *Zmpsy1–PacrtW* in all lines where the relevant pair was present. The genes chosen for these experiments were those that have a stronger effect in changing metabolic accumulation in each line, when their expression was independently changed.

First, we independently doubled the expression of either *Zmpsy1–Zmlyce* or *Zmpsy1–Zmlycb* in Line 1 (Figure 3). We observed that these coordinated changes had effects that were approximately additive (in normalized or log-space). For example, accumulation of phytoene increased by three-fold in the *Zmpsy1* single-gene experiment, was not affected in the *Zmlyce* single-gene experiment and decreased by half in the *Zmlycb* single-gene experiment. This led to a three-fold increase in phytoene in the *Zmpsy1–Zmlyce* double-gene experiment and to a 2.5-fold increase in the *Zmpsy1–Zmlycb* double-gene experiment.

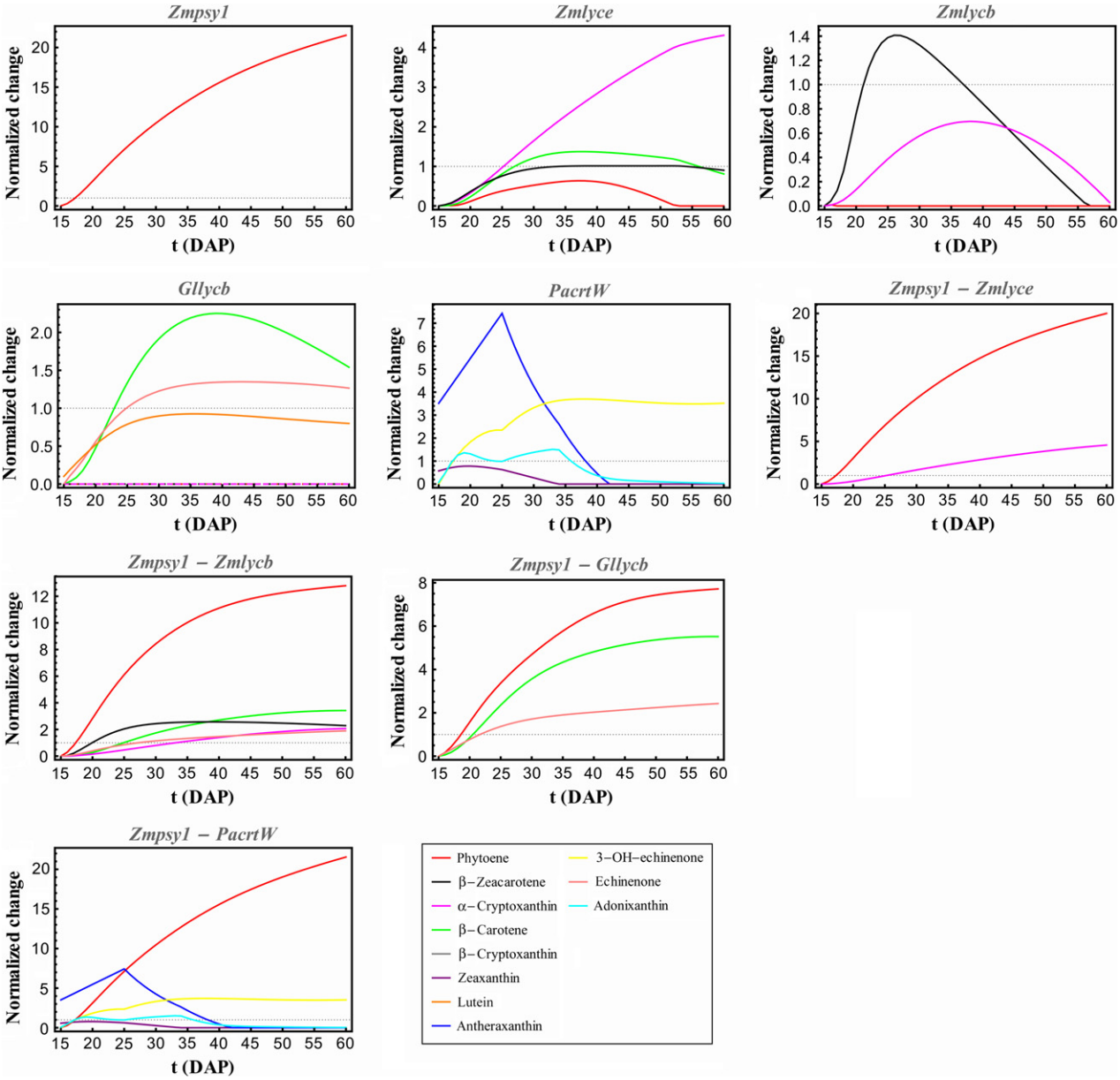

Figure 6. Effect of doubling basal gene expression in Line 4.
The effect of doubling the basal level of gene expression for *Zmpsy1*, *Zmlyce*, *Zmlycb*, *Gllycb* and *PacrtW* and the effect of simultaneously doubling the basal level of gene expression for *Zmpsy1* and *Zmlyce* (*Zmpsy1*–*Zmlyce*), *Zmpsy1* and *Zmlycb* (*Zmpsy1*–*Zmlycb*), *Zmpsy1* and *Gllycb* (*Zmpsy1*–*Gllycb*), and *Zmpsy1* and *PacrtW* (*Zmpsy1*–*PacrtW*). The *x*-axis in each plot represents time, while the *y*-axis represents the normalized accumulation of carotenoids. Normalization was done by dividing the amount of accumulated metabolite at each time point by the amount of that metabolite which accumulated at 60 days after pollination (DAP) in line 4. If the curve goes above 1, the metabolite accumulates with respect to Line 4; if it stays below 1, the metabolite is less abundant after changing basal gene expression than it was in Line 4.

In Line 2, most experiments also showed an approximately additive effect between the two genes of each pair (Figure 4). Some non-linearities were observed in the *Zmpsy1*–*Zmlyce* double-gene experiment. This experiment identified an increase in the amounts of accumulated α-cryptoxanthin and a decrease in the accumulation of β-carotene with respect to what one would expect from the independent effects of *Zmpsy1* and *Zmlyce*. The *Zmpsy1*–*Zmlycb* and *Zmpsy1*–*Gllycb* double-gene experiments also showed similar non-linearities regarding the production of β-carotene, resulting in the accumulation of unexpectedly high amounts of the metabolite.

Double-gene experiments in Lines 3 (Figure 5) and 4 (Figure 6) showed an approximately additive effect between the two genes of each pair (Figure 5).

Experimental validation of predictions

We hoped to validate two qualitative predictions resulting from our analysis. First, the prediction that an increase in Zmpsy1 transcript levels (which correlates to protein activity), would lead to increased levels of accumulated phytoene 60 DAP. Second, that an increase in Zmpsy1 transcript levels does not generally correlate to accumulated levels of β-carotene 60 DAP.

We therefore measured Zmpsy1 transcript levels in the endosperm of four additional synthetic maize lines (CARO2, KETO2, OR×CARO2 and OR×KETO2) 30 DAP: CARO2 expressed *Zmpsy1*, *Pacrtl* and *Gllycb*; KETO2 expressed *Zmpsy1*, *sCrbkt* and *sBrcrtZ*; OR×CARO2 expressed the same transgenes as CARO2 in addition to *AtOr*; OR×KETO2 expressed the same transgenes as KETO2 in addition to *AtOr*. Additionally, we also measured the accumulated levels of phytoene and β-carotene in those lines 60 DAP. The results are shown in Table 2.

To measure how changing the levels of *Zmpsy1* transcript affect end-point accumulation of phytoene at 60 DAP we compared line CARO2 with line OR×CARO2 and line KETO2 with line OR×KETO2. This pairing permits a more controlled comparison, because the two lines in each pairwise comparison only differ in one gene (*AtOR*). We calculated the finite relative sensitivity of phytoene to *Zmpsy1* transcript levels (*FS*(*phyt,psy1*), defined in the Experimental Procedures). If our prediction that increasing Zmpsy1 activity increases phytoene accumulation is qualitatively correct, then *FS*(*phyt,psy1*) should be larger than zero. Table 3 shows that this is the case.

Similarly, to measure how modified levels of *Zmpsy1* transcript affect end-point accumulation of β-carotene at 60 DAP we also calculated the finite relative sensitivity of β-carotene to *Zmpsy1* transcript levels (*FS*(*β-car,psy1*), defined in Experimental Procedures). Our predictions emphasize that increasing *Zmpsy1* transcript levels only leads to increases in β-carotene accumulation in specific situations (e.g. for Line 2) and not as a general rule. If these predictions are qualitatively correct, then *FS*(*β-car, psy1*) should vary greatly between positive and negative numbers, depending on the maize line. Table 3 shows that this is the case.

Table 2 Experimental measurements from lines CARO2, KETO2, OR×CARO2 and OR×KETO2

Line	*Zmpsy1* mRNA levels 30 DAP	β-Carotene levels 60 DAP	Phytoene levels 60 DAP
CARO2	1.6	8.2	117.5
KETO2	0.15	11.8	53.7
OR×CARO2	1	5.9	102.2
OR×KETO2	0.16	5.2	61.3

DAP, days after pollination.
mRNA transcript levels were normalized to the levels of actin mRNA. Carotenoid levels are measured in μg g^{-1} dry weight.

Table 3 Finite relative sensitivities of end-point accumulation of carotenoids to changes in *Zmpsy1* transcript levels

Comparison	*FS*(*phytonene, Zmpsy1*)	*FS*(*β-carotene, Zmpsy1*)
CARO2 versus OR×CARO2	0.3	0.8
KETO2 versus OR×KETO2	2.1	−8.4

DISCUSSION

Mathematical models are frequently used in synthetic biology to aid in the design and implementation of novel biological circuits with precise behavior (Benner and Sismour, 2005). Although less common, it is also possible to use such models to study a posteriori how to improve a novel biological circuit that has been introduced into an organism. Both methods rely on available quantitative information on the behavior of the parts used in the circuits. Having this information to hand is crucial for accurate mathematical modeling. Initiatives like BioBricks (Shetty *et al.*, 2008) or BIOFAB (Mutalik *et al.*, 2013) strive to create a repository of well-standardized biological parts that can be used for circuit design in synthetic biology. Quantitative information on biological parts is often limited to microbes, and knowledge of mechanisms and parameters is still absent for most plant and/or animal systems. In addition, it is often difficult to estimate *a priori* which interactions the synthetic pathway will have with the metabolism of the host organism. This leads to situations in which traditional model building is hindered by a lack of information about mechanisms, parameter values and identification of additional biological processes that influence the behavior of the synthetic circuit (Liu and Stewart, 2015). Consequently, using such models to efficiently identify minimal changes in the circuit that could improve its function becomes almost impossible. In this work we implemented a workflow that effectively sidesteps some of these challenges. The workflow is summarized in Figure 2 and combines experimental measurement of gene expression and metabolite concentrations (Farré *et al.*, 2013) with mathematical modeling.

First, we avoided the need for additional measurements to obtain quantitative information about the system. We did this by using mathematical interpolation to obtain a dense time series from the levels of gene expression and metabolites. This method is valid when the changes between experimentally measured points are smooth, as seems to be the case for our systems.

Second, we sidestepped the lack of knowledge about the mechanism of individual steps in the pathways. We did this by using the GMA representation, a mathematical formalism based on approximation theory (Voit and

Savageau, 1982; Voit, 2013), and allowing for the parameter optimization process to identify time points at which basal metabolic and gene expression activities could change, resulting in piecewise differential equations. This provided a heuristic way to capture changes in promoter activities that are well known to occur during the development of maize endosperm (Fraser *et al.*, 1994; Ghassemi-Golezani *et al.*, 2011).

Third, the fitting process revealed critical points in our understanding of how the pathway operates within the host and interacts with native cell metabolism, suggesting additional biological processes that must be considered if one is to quantitatively understand the dynamic behavior in the four maize lines. Those processes were sink reactions for some carotenoids that could result either from light-induced degradation that is known to occur (Boon *et al.*, 2010) or from non-specific utilization of those carotenoids by other aspects of endosperm metabolism.

Why are these processes not equally important in shaping the accumulation dynamics of carotenoids in all lines? The parameter values for the various models help us explain this. For example, consider β-carotene. In all lines, the rate parameter for production of β-carotene (α8) is of the order of magnitude of 10 DAP^{-1}. In Line 4, the order of magnitude of the rate parameter that determines usage of β-carotene to produce other carotenoids (α9) is of the order of magnitude 1 DAP^{-1}, which is also the order of magnitude of the rate parameter that governs degradation of β-carotene by other means (e.g. light). This means that the flux of β-carotene degradation through both channels (generic and synthesis of other carotenoids) is comparable in Line 4. In contrast, in Lines 1–3 the order of magnitude of the rate parameter that determines usage of β-carotene to produce other carotenoids (α9) is also of the order of magnitude of 10 DAP^{-1}. This means that adding an additional β-carotene degradation process, which we allowed for in the optimization process, will have little impact on the quality of model fitting and this little effect will not improve the fitting in a statistically significant way. Because of that, the rate constant for this process was not significantly different from zero and the process was not explicitly considered in the model. Therefore this process is negligible for Lines 1–3. Similar reasoning can be used to explain the other cases in Table 1.

The experimental measurements of gene expression and metabolite abundance in the four maize lines demonstrated the non-linearity in the dynamic behavior of the maize lines. Although modeling the non-linear nature of biochemical systems is not a trivial task, our models accurately capture these non-linearities in the four plant lines. That the models described the data well suggests that the application of sensitivity analysis to those models would predict how variations in transcriptional rates might affect final metabolite accumulation in the maize endosperm. Importantly, we identified those specific transcriptional rates in which changes would more effectively increase the accumulation of each specific carotenoid in the endosperm. This information permits rational planning of which genes to target for further engineering in order to improve metabolic yields in the pathways. We have experimentally validated some of the predictions by analyzing how changing levels of *Zmpsy1* affect the accumulation of phytoene and β-carotene at 60 DAP in four additional synthetic lines that are independent of those for which we created the models. We found that increasing levels of *Zmpsy1* were positively correlated with the accumulation of phytoene 60 DAP, as predicted. We also found that, as predicted, levels of *Zmpsy1* are not necessarily positively correlated with accumulated levels of β-carotene 60 DAP.

Moreover, we performed further *in silico* experiments to analyze the effect of such perturbations. Results from these experiments were quantitatively consistent with most of our predictions from sensitivity analysis. Discrepancies are a consequence of non-linear effects of changes in gene expression propagating through the system. The *in silico* experiments also identified various potentially useful synergistic effects. For example, simultaneous upregulating of the expression of *Zmpsy1* and *Gllycb* was predicted to lead to an accumulation of β-carotene superior to what one might expect from summing the effects of the individual gene perturbation experiments. A more accurate and systematic synergism analysis could be carried out using second-order modeling formalisms (Salvador, 2000). However, using such formalisms significantly increases the number of parameters that need to be optimized in the model, which would require increasing the size of the dataset used for model optimization.

There are several limitations in this work that should be taken into consideration. First, using an approximate formalism means that predictions derived from the model analyses have a range of validity that is model dependent. However, the validity of GMA models such as the ones we built tend to range over several orders of magnitude for the values of the concentrations (Sorribas and Savageau, 1989; Voit, 2013). Secondly, the limited number of experimental data necessitated interpolations (and to a lesser extent extrapolation) to estimate additional data points. As a result, if the changes in gene expression are not smooth between interpolators, these estimations and the parameter values of the models will be skewed. However, our data with synthetic maize suggests that changes in gene expression are smooth over time (Zhu *et al.*, 2008; Naqvi *et al.*, 2009, 2011; Li *et al.*, 2010; Bai *et al.*, 2011; Farré *et al.*, 2013). Thirdly, the lack of data on direct protein amounts and/or activity introduces an additional layer of uncertainty in the modeling, by having to assume that changes in gene expression are directly proportional with, or correlate to changes in protein levels. Published data suggest that this

assumption is reasonable for metabolic enzymes in plant systems, particularly in carotenoid metabolism (Fraser *et al.*, 1994). Nevertheless, the fact that the models explain the experimental data very well suggests that these limitations may not be critical for the systems we investigated in this work.

There is a final noteworthy limitation to our work. Plant synthetic biology still lacks the means to control the expression of genes as proposed here, making it impossible at this point to devise an experimental method to regulate gene expression in a controlled way and implement our *in silico* experiments other than by trial and error. Our approach could provide further motivation for the development of more quantitative methods for manipulation of gene expression in plants, for example by developing promoters with well-calibrated expression levels.

Overall, our analyses have revealed which genes should be further engineered to increase the accumulation of specific carotenoid metabolites. Our results showed that many of these genes are the same in the various maize lines, suggesting that there are common strategies for qualitatively increasing the yields of the four pathways.

Some of the strategies identified here correlate well with the genetic manipulations that an expert would use to attempt to improve accumulation of various carotenoids. For example, an expert in carotenoid biosynthesis would be likely to try to increase the activity of the first enzyme in each branch of the pathway in order to increase the accumulation of carotenoids from that branch. This strategy is also identified by our models. This partially validates the models and suggests that they could be trusted to identify at least some of the less obvious strategies that could improve carotenoid accumulation. For example, increasing *Zmpsy1* expression in Line 2 is predicted to increase the accumulation of β-carotene, a result that is consistent with other experimental observations in various plant systems (Shewmaker *et al.*, 1999; Lindgren *et al.*, 2003; Ravanello *et al.*, 2003; Diretto *et al.*, 2007; Maass *et al.*, 2009; Kim *et al.*, 2012; Luo *et al.*, 2013; da Silva Messias *et al.*, 2014).

The models can also be used to pinpoint the quantitative modulation of gene expression that one might require in order to reach specific carotenoid production targets. For example, our analysis reveals that if one has a specific quantitative goal for improving the yields the required changes in gene expression are specific for each pathway. Several examples can be used to illustrate this point. Increasing the expression of *Zmpsy1* is predicted to increase the accumulation of phytoene in all maize lines. However, depending on the specific line, doubling the basal level of *Zmpsy1* expression will cause phytoene accumulation to increase between 2 and 20 times. Increasing the expression of *Zmlyce* is predicted to increase the accumulation of α-carotenoids in all maize lines. However, depending on the specific line, doubling the basal level of *Zmlyce* expression will cause α-carotene accumulation to increase between 3 and 15 times. As a final example, increasing the expression of either *Zmlycb* or *Gllycb* is predicted to increase the accumulation of β-carotene in all maize lines that have the gene. However, depending on the specific line, doubling the basal level of *Gllycb* expression will cause accumulation of β-carotenes to increase between 1.5 and 5 times.

Our results also identified those lines whose further engineering is more likely to lead to higher yields for the various carotenoids. For example, if one wants to increase β-carotene accumulation by as much as possible, one should focus on further engineering Line 2. Doubling the basal gene expression levels of *Zmpsy1* and *Gllycb* is predicted to lead to a 200-fold increase in β-carotene accumulation in this line. If a single genetic manipulation is to be used, then doubling *Zmlycb* basal gene expression in this line is predicted to lead to a seven-fold increase in β-carotene.

Finally, the analysis also indicated that changing the expression of a gene may have pleiotropic effects, leading to the accumulation of some carotenoid(s) at the cost of depleting the metabolic pools of other pathway intermediates. For example, doubling the expression of *Zmlyce* in Line 2 is predicted to increase α-carotene by 15-fold while decreasing the accumulated amount of β-carotene by the same relative amount.

EXPERIMENTAL PROCEDURES

Transgenic plants, gene expression measurements and metabolomics data

The experimental data were generated from four transgenic plant lines carrying various combinations of carotenogenic transgenes, as described in Zhu *et al.* (2008) and Farré *et al.* (2016). Endosperm samples were taken from immature seeds at 15, 20, 25, 30, 40, 50 and 60 DAP, frozen in liquid nitrogen and stored at −80°C. Quantitative RT-PCR of the relevant genes and metabolomics measurements were performed as described in Farré *et al.* (2013). Three replicates of all measurements were made and averaged.

The experimental data to validate the predictions were generated from four other transgenic plant lines carrying various combinations of carotenogenic transgenes. Line CARO2 contains *Zmpsy1*, *Pacrtl* and *Gllycb*. Line KETO2 contains *Zmpsy1*, *sCrbkt* and *sBrcrtZ*. Line OR×CARO2 was obtained by crossing CARO2 with a maize line containing the *Arabidopsis thaliana* Orange gene (*AtOR*). Line OR×KETO2 was obtained by crossing KETO2 with a maize line containing *AtOR*. Endosperm samples were taken from immature seeds at 30 and 60 DAP, frozen in liquid nitrogen and stored at −80°C. Quantitative RT-PCR of *Zmpsy1* and metabolomics measurements of phytoene and β-carotene were performed as described in (Farré *et al.*, 2013).

Software

All calculations, including implementation and analysis of the mathematical models, were done using Mathematica (Wolfram, 2003). All notebooks are provided as supporting files (Notebook files S1–S4).

Obtaining data for model optimization

The transcripts of 12 endogenous carotenogenic genes, together with five carotenogenic transgenes, were measured. The accumulation of total and individual carotenoids was also measured. The measurements for the four transgenic maize lines were taken over a period of 2 months, providing molecular snapshots of carotenoid biosynthesis in maize endosperm at several developmental stages.

The temporal profiles of the transcripts only cover the period between 15 and 30 DAP, while the temporal profiles of carotenoids cover the period between 15 and 60 DAP. This is because RNA can only be isolated from immature endosperm tissue, as seed maturation leads to the degradation of RNA. In contrast, carotenoids can be extracted even if the seed is mature. As a result, we used linear extrapolation to estimate transcript abundance between 30 and 60 DAP in all lines. To do this, we assumed that transcript abundances undergo a 30% decay between 30 and 60 DAP, as suggested by published transcript measurements of endogenous secondary metabolism genes in maize endosperm (Méchin *et al.*, 2007).

In all cases, our time series dataset was too small to allow for a statistically significant optimization of the models for each line. To obtain additional (pseudo-)experimental data we interpolated the time series. This is an approach commonly used whenever the variation between consecutive observed data points is sufficiently smooth and the experimental error is suitably small (Baud *et al.*, 1991; Buzzi-Ferraris and Manenti, 2010; Farré *et al.*, 2013).

Several interpolation functions are available. We employed an Akima interpolation function, thus guaranteeing smooth curves that are continuous and have continuous first derivatives (Akima, 1974). These two aspects are important for model optimization (see below). The interpolated functions were then used to generate dense time series of extended experimental measurements that we employed for model optimization.

Model building and optimization

To model the gene expression layer we assumed that the transcriptional rate (*TR*) of a gene *G* at any given developmental stage is constant. We also assumed that mRNA degradation follows first-order kinetics (Equation 1). Both these assumptions have been validated for many genes in higher plants (Green, 1993; Lambein, 2003). Furthermore, we assumed the overall gene expression dynamics of any given gene may be approximated as a piecewise defined function. This is a phenomenological way of taking into account possible changes in gene expression occurring between the various developmental stages that maize endosperm undergoes after pollination (Ghassemi-Golezani *et al.*, 2011). In addition, it is well known that the metabolic activity of the maize endosperm decays when reaching its mature developmental stages (Méchin *et al.*, 2007). To account for this we assume that the degradation rate constant k depends linearly on time as follows: $k(t) = bt$, where b is a positive number. Overall, the rate equation that governs the gene expression dynamics for each gene is given by Equation 1.

To model the carotenoid accumulation layer we used the power-law formalism and the GMA representation (Equation 2). The properties of this mathematical and computational framework have been recently reviewed at length (Savageau, 1969a,b, 1971; Voit, 2013). This was an appropriate modeling framework under our experimental conditions, for four reasons. First, it allowed us to build models when no detailed information was available about the mechanisms of the processes being modeled. Second, the framework allowed for automated creation of a model from a conceptual diagram of the pathways in each line. Third, the mathematical representation is simple, yet non-linear, thus permitting the capture of some of the non-linearities in the dynamic behavior of the system being modeled. Finally, the number of parameters (αs, βs, gs and hs) to be estimated was lower than that for other non-linear models that can be created for the same system (Voit, 2013).

The GMA representation is canonical. This means that model construction, diagnosis and analysis follow strict rules (Voit, 2013). The starting point of model construction is identifying which variables are important and should be explicitly included in the model (Voit, 1991). Therefore, we included as variables all the carotenoids and transcripts involved in the carotenogenic pathway of the engineered maize lines whose corresponding abundance was measured previously (Farré *et al.*, 2013).

Subsequently, we enumerated the relationships among these variables. We did this by deciding whether or not variable X_j directly influences accumulation or clearance of variable X_i (Voit, 1991). We searched for such relationships in the literature and those we found are listed in Tables S1 and S2, next to the transgenes or endogenous genes that code for the corresponding enzymatic activity. A version of the full carotenogenic pathway in plants is depicted in Figure 1. This initial conceptual model was then individually adapted to each of the four maize lines as shown in Figure S1. The model for a specific line considered only those genes that are present in that line and the metabolites that are consistently above detection limits. This is because no variable for which quantitative data are absent may be included in the mathematical models.

Next, we formulated the independent GMA representations corresponding to each of the conceptual diagrams in Figure S1. That is, the instantaneous change $\dot{X}_i$ of each metabolite/dependent variable X_i was defined as a combination of the rates of the different reactions acting on X_i, either producing or degrading it. Each term is a product of power-law functions containing every relevant variable that directly affects the process, as described in Equation 2. We used the GMA representation to write the system of ODEs for each maize line. These ODE systems allowed for numerical simulation and analysis of the temporal dynamics of the pathways, and facilitated the prediction of pathway responses to changes in any of the model parameters (Voit, 1991).

The next step is to decide on the method for model optimization that will assign adequate numerical values to the parameters in the GMA models. This is arguably one of the most difficult steps in the analysis of biological systems (Voit, 2013). We used the slope-substitution method (Voit and Savageau, 1982; Voit *et al.*, 2009; Lee *et al.*, 2011). This method requires a numerical estimation of the rates of change for each metabolite at a sufficiently large number of time points. These estimations are given by the numerical first derivative of the Akima interpolation functions. Here, we sampled the time series at 1-day intervals. For each time, the left-hand side of the equations is replaced with numerical rates for that time point. As a result the ODE system becomes a system of $n \times m$ algebraic equations, where n is the number of equations in the original ODE system and m is the number of time points we sampled (Voit, 2013). This procedure is statistically valid (Brunel, 2008) and has an important advantage: it avoids all numerical integrations of the differential equations, which significantly speeds up the process (Voit, 2013). We note that the power-law formalism imposes constraints on the possible values of the parameters. Apparent rate constants are always larger than zero, while positive kinetic orders are known to have values that are lower than small integers, typically five (Sorribas and Savageau, 1989). Here we allowed six to be the upper bound for

those kinetic orders. The constrained optimizations were performed by the NonlinearModelFit function of Mathematica, using an interior point method to minimize the least squares of the residuals.

Next, we estimated the optimal parameter values of the transcript time series. Due to the way in which the transgenes were inserted in the four maize lines, their expression is probably not regulated by either the amount of carotenoid or the expression levels of the other genes. This enabled us to estimate the kinetic parameters for each gene expression independently, using Equation 1. In some cases we did this with two piecewise functions as described by Machina *et al.* (2010), because Equation 1 does not always capture the non-monotonic behaviors of some of the transcript time series. These piecewise equations provide an implicit way to account for any major changes in basal gene expression throughout maize endosperm development.

Subsequently, we estimated the parameter values of the GMA models for the metabolites, through multi-objective non-linear optimization of the set of algebraic equations obtained using the slope-substitution method. Whenever a single GMA equation could not possibly capture the non-monotonic behaviors of the time series for the relevant metabolite, we allowed that equation to be described by a piecewise function with one breakpoint as described by Machina *et al.* (2010) and as mentioned above for gene expression.

Assessment of model quality

In order to assess the quality of the optimized parameters we computed their 95% confidence intervals (see Tables S11–S14). To do this we used a jackknife approach, adapted from Alper and Gelb (1990). We randomly extracted 10% of the points in the data sets used in the optimization. We locked the value of all parameters but one at their best estimated value. We then optimized the remaining parameter and repeated this step up to 100 times, each time with a new set of data. In this way we were able to compute the confidence interval of all the estimated parameters without any assumption regarding their distribution. We note that confidence intervals computed for parameters whose value is close to the boundary imposed by our constrained optimization are only indicative and may be significantly biased. We remind the reader that the constraints imposed on parameter values are that $0 \leq$ kinetic orders ≤ 6 and $0 \leq$ rate constants.

Sensitivity analysis

We performed a sensitivity analysis on our optimized models. By definition, sensitivity measures how much a feature changes if one of the parameters in the system is varied by a certain amount (Voit, 2013). This analysis allowed us to understand how the variables of the system depended on its parameters, therefore providing information about potentially useful and relevant regulatory targets.

Sensitivity analysis also separates situations where small inaccuracies in parameter values are almost irrelevant from those where noise or error in the estimates can lead to almost unpredictable results. Consequently, this analysis also allowed us to identify components of the model that may be problematic due to unusually high sensitivity values (Savageau, 1975; Voit, 2013).

Relative steady-state parameter sensitivities are defined as 'the relative change in a system component (X) that is caused by a relative change in a parameter value (p)', (Voit, 1991):

$$\bar{S}(X,p) = \frac{\partial X/X}{\partial p/p} = \frac{\partial \log X}{\partial \log p}. \tag{3}$$

The Methods section in Appendix S1 shows how we can extend this definition out of steady state for a dynamic system with multiple variables and parameters. That extension is used to compute the relative dynamic sensitivity of each dependent variable to the estimated rate constant parameters in our models. Those sensitivities are presented in Tables S3–S10.

Given that we were interested in how changing the expression of specific genes would affect the accumulation of the various carotenoid metabolites, we analyzed the sensitivities of each metabolite to changes in the estimated transcriptional rate of each gene. These sensitivities allowed us to predict how to further engineer the transgenic maize lines by revealing which genes should be over- or underexpressed, in order to increase accumulation of specific carotenoids.

In order to validate some of the predictions of the sensitivity analysis we calculated a finite relative sensitivity of end-point accumulation of a metabolite M to changes in the activity of a gene G using Equation 4:

$$\begin{aligned} FS(M,G) &= \frac{\Delta \log(M)}{\Delta \log(G)} \\ &= \frac{[G]_{\text{line with highest activity of } G}([M]_{\text{line with highest activity of } G} - [M]_{\text{line with lowest activity of } G})_{\text{final}}}{[M]_{\text{line with highest activity of } G}([G]_{\text{line with highest activity of } G} - [M]_{\text{line with lowest activity of } G})} \end{aligned} \tag{4}$$

Here, $[G]$ and $[M]$ represent the levels of G transcript and M accumulation, respectively. The subscript 'final' indicates that we consider the latest DAP for which metabolite measurements are available in the relevant line (50 DAP in Line 1, 60 DAP in Lines 2–4). *FS* provides a finite estimation of the relative change in M with respect to a relative change in G. As is the case with differential sensitivities, $FS(M,G) > 0$ (<0) means that an increase in the activity of G will lead to an increase (decrease) in M.

ACKNOWLEDGEMENTS

This work was partially funded by grants BFU2010-17704 from the Spanish MINECO and from small grants CMB and TR255 from the University of Lleida to RA, TIN2011-28689-C02-02 and TIN2014-53234-C2-2-R to FS, BIO2011-23324, BIO02011-22525 and PIM2010PKB-0074 for MINECO to PC and CFZ, a European Research Council IDEAS Advanced Grant Program (BIOFORCE) (to PC), ERC-2013-PoC 619161 (to PC) and RecerCaixa (to PC). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Line-specific carotenoid biosynthetic pathways.

Figure S2. Relative sensitivities of metabolites to changes in the basal transcription rate of each gene.

Figure S3. Experimental data, Akima interpolation and simulation results for Line 1.

Figure S4. Standardized residuals of the optimized model for Line 1.

Figure S5. Experimental data, Akima interpolation and simulation results for Line 2.

Figure S6. Standardized residuals of the optimized model for Line 2.

Figure S7. Experimental data, Akima interpolation and simulation results for Line 3.

Figure S8. Standardized residuals of the optimized model for Line 3.

Figure S9. Experimental data, Akima interpolation and simulation results for Line 4.

Figure S10. Standardized residuals of the optimized model for Line 4.

Figure S11. Analysis of goodness of fit for the gene expression of Line 1.

Figure S12. Analysis of goodness of fit for the gene expression of Line 2.

Figure S13. Analysis of goodness of fit for the gene expression of Line 3.

Figure S14. Analysis of goodness of fit for the gene expression of Line 4.

Figure S15. Dynamic relative sensitivities of the various metabolites in Line 1 to changes in the rate constants for gene expression.

Figure S16. Dynamic relative sensitivities of the various metabolites in Line 2 to changes in the rate constants for gene expression.

Figure S17. Dynamic relative sensitivities of the various metabolites in Line 3 to changes in the rate constants for gene expression.

Figure S18. Dynamic relative sensitivities of the various metabolites in Line 4 to changes in the rate constants for gene expression.

Table S1. List of transgenes used to create the four synthetic maize lines.

Table S2. List of endogenous genes active in the endosperm and relevant for carotenoid biosynthesis.

Table S3. Relative sensitivity of each metabolite to changes in the apparent rate constant of each reaction in Line 1, 50 days after pollination.

Table S4. Relative sensitivity of each metabolite to changes in the apparent rate constant of each reaction in Line 2, 60 days after pollination.

Table S5. Relative sensitivity of each metabolite to changes in the apparent rate constant of each reaction in Line 3, 60 days after pollination.

Table S6. Relative sensitivity of each metabolite to changes in the apparent rate constant of each reaction in Line 4, 60 days after pollination.

Table S7. Relative sensitivity of each metabolite to changes in the rate constant for transcription of each gene considered in Line 1, 50 days after pollination.

Table S8. Relative sensitivity of each metabolite to changes in the rate constant for transcription of each gene considered in Line 2, 60 days after pollination.

Table S9. Relative sensitivity of each metabolite to changes in the rate constant for transcription of each gene considered in Line 3, 60 days after pollination.

Table S10. Relative sensitivity of each metabolite to changes in the rate constant for transcription of each gene considered in Line 4, 60 days after pollination.

Table S11. Parameter values for the mathematical model for Line 1.

Table S12. Parameter values for the mathematical model for Line 2.

Table S13. Parameter values for the mathematical model for Line 3.

Table S14. Parameter values for the mathematical model for Line 4.

Appendix S1. Supporting background, methods, results and references.

Notebook file S1. Mathematica notebook containing all the code for the simulation and analysis of Line 1.

Notebook file S2. Mathematica notebook containing all the code for the simulation and analysis of Line 2.

Notebook file S3. Mathematica notebook containing all the code for the simulation and analysis of Line 3.

Notebook file S4. Mathematica notebook containing all the code for the simulation and analysis of Line 4.

REFERENCES

Adam, L., Kozar, M., Letort, G., Mirat, O., Srivastava, A., Stewart, T., Wilson, M.L. and Peccoud, J. (2011) Strengths and limitations of the federal guidance on synthetic DNA. *Nat. Biotechnol.* **29**, 208–210.

Akima, H. (1974) A method of bivariate interpolation and smooth surface fitting based on local procedures. *Commun. ACM*, **17**, 18–20.

Alper, J.S. and Gelb, R.I. (1990) Standard errors and confidence intervals in nonlinear regression: comparison of Monte Carlo and parametric statistics. *J. Phys. Chem.* **94**, 4747–4751.

Alves, R., Vilaprinyo, E., Hernádez-Bermejo, B. and Sorribas, A. (2008) Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways. *Biotechnol. Genet. Eng. Rev.* **25**, 1–40.

Atkinson, M.R., Savageau, M.A., Myers, J.T. and Ninfa, A.J. (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli. *Cell*, **113**, 597–607.

Bacchus, W., Aubel, D. and Fussenegger, M. (2013) Biomedically relevant circuit-design strategies in mammalian synthetic biology. *Mol. Syst. Biol.* **9**, 691.

Bai, C., Twyman, R.M., Farré, G., Sanahuja, G., Christou, P., Capell, T. and Zhu, C. (2011) A golden era—pro-vitamin A enhancement in diverse crops. *Vitr. Cell. Dev. Biol. - Plant*, **47**, 205–221.

Balagaddé, F.K., Song, H., Ozaki, J., Collins, C.H., Barnet, M., Arnold, F.H., Quake, S.R. and You, L. (2008) A synthetic Escherichia coli predator-prey ecosystem. *Mol. Syst. Biol.* **4**, 187.

Basu, A. and Imrhan, V. (2007) Tomatoes versus lycopene in oxidative stress and carcinogenesis: conclusions from clinical trials. *Eur. J. Clin. Nutr.* **61**, 295–303.

Baud, M., Mercier, M. and Chatelain, F. (1991) Transforming signals into quantitative values and mathematical treatment of data. *Scand. J. Clin. Lab. Investig. Suppl.* **205**, 120–130.

Benner, S.A. and Sismour, A.M. (2005) Synthetic biology. *Nat. Rev. Genet.* **6**, 533–543.

Berman, J., Zorrilla-López, U., Farré, G., Zhu, C., Sandmann, G., Twyman, R.M., Capell, T. and Christou, P. (2014) Nutritionally important carotenoids as consumer products. *Phytochem. Rev.* **14**, 727–743.

Boon, C.S., McClements, D.J., Weiss, J. and Decker, E.A. (2010) Factors influencing the chemical stability of carotenoids in foods. *Crit. Rev. Food Sci. Nutr.* **50**, 515–532.

Brophy, J.A.N. and Voigt, C.A. (2014) Principles of genetic circuit design. *Nat. Methods*, **11**, 508–520.

Brunel, N.J.-B. (2008) Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Stat.* **2**, 1242–1267.

Buzzi-Ferraris, G. and Manenti, F. (2010) *Interpolation and Regression Models for the Chemical Engineer* 1st edn, Weinheim: Wiley-VCH.

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A. and Lai, J. (2014) Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol.* **166**, 252–264.

DeLoache, W.C., Russ, Z.N., Narcross, L., Gonzales, A.M., Martin, V.J.J. and Dueber, J.E. (2015) An enzyme-coupled biosensor enables (S)-reticuline production in yeast from glucose. *Nat. Chem. Biol.* **11**, 465–471.

Diretto, G., Al-Babili, S., Tavazza, R., Papacchioli, V., Beyer, P. and Giuliano, G. (2007) Metabolic engineering of potato carotenoid content through tuber-specific overexpression of a bacterial mini-pathway. *PLoS ONE*, **2**, e350.

Farré, G., Maiam Rivera, S., Alves, R. *et al.* (2013) Targeted transcriptomic and metabolic profiling reveals temporal bottlenecks in the maize carotenoid pathway that may be addressed by multigene engineering. *Plant J.* **75**, 441–455.

Farré, G., Blancquaert, D., Capell, T., Straeten, D. Van Der, Christou, P. and Zhu, C. (2014) Engineering Complex Metabolic Pathways in Plants. *Annu. Rev. Plant Biol.* **65**, 187–223.

Farré, G., Perez-Fons, L., Decourcelle, M. *et al.* (2016) Metabolic engineering of astaxanthin biosynthesis in maize endosperm and characterization of a prototype high oil hybrid. *Transgenic Res.* 1–13. doi:10.1007/s11248-016-9943-7

Fraser, P.D., Truesdale, M.R., Bird, C.R., Schuch, W. and Bramley, P.M. (1994) Carotenoid Biosynthesis during tomato fruit development (Evidence for Tissue-Specific Gene Expression). *Plant Physiol.* **105**, 405–413.

Gallardo, K., Firnhaber, C., Zuber, H., Héricher, D., Belghazi, M., Henry, C., Küster, H. and Thompson, R. (2007) A combined proteome and transcriptome analysis of developing Medicago truncatula seeds: evidence for metabolic specialization of maternal and filial tissues. *Mol. Cell Proteomics*, **6**, 2165–2179.

García-Plazaola, J.I., Matsubara, S. and Osmond, C.B. (2007) The lutein epoxide cycle in higher plants: its relationships to other xanthophyll cycles and possible functions. *Funct. Plant Biol.* **34**, 759.

Ghassemi-Golezani, K., Tajbakhsh, Z. and Raey, Y. (2011) Seed development and quality in maize cultivars. *Not. Bot. Horti Agrobot. Cluj-Napoca*, **39**, 178–182.

Green, P.J. (1993) Control of mRNA stability in higher plants. *Plant Physiol.* **102**, 1065–1070.

Kim, M.-J., Kim, J.K., Kim, H.J. *et al.* (2012) Genetic modification of the soybean to enhance the β-carotene content through seed-specific expression. *PLoS ONE*, **7**, e48287.

Lambein, I. (2003) Decay kinetics of autogenously regulated CGS1 mRNA that codes for cystathionine -synthase in Arabidopsis thaliana. *Plant Cell Physiol.* **44**, 893–900.

LaVan, D.A. and Marmon, L.M. (2010) Safe and effective synthetic biology. *Nat. Biotechnol.* **28**, 1010–1012.

Lee, Y., Chen, P.-W. and Voit, E.O. (2011) Analysis of operating principles with S-system models. *Math. Biosci.* **231**, 49–60.

Li, Q., Farre, G., Naqvi, S. *et al.* (2010) Cloning and functional characterization of the maize carotenoid isomerase and β-carotene hydroxylase genes and their regulation during endosperm maturation. *Transgenic Res.* **19**, 1053–1068.

Li, G., Wang, D., Yang, R. *et al.* (2014) Temporal patterns of gene expression in developing maize endosperm identified through transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **111**, 7582–7587.

Lindgren, L.O., Stålberg, K.G. and Höglund, A.-S. (2003) Seed-specific overexpression of an endogenous Arabidopsis phytoene synthase gene results in delayed germination and increased levels of carotenoids, chlorophyll, and abscisic acid. *Plant Physiol.* **132**, 779–785.

Liu, W. and Stewart, C.N. (2015) Plant synthetic biology. *Trends Plant Sci.* **20**, 309–317.

Luo, Z., Zhang, J., Li, J., Yang, C., Wang, T., Ouyang, B., Li, H., Giovannoni, J. and Ye, Z. (2013) A STAY-GREEN protein SlSGR1 regulates lycopene and β-carotene accumulation by interacting directly with SlPSY1 during ripening processes in tomato. *New Phytol.* **198**, 442–452.

Maass, D., Arango, J., Wüst, F., Beyer, P. and Welsch, R. (2009) Carotenoid crystal formation in Arabidopsis and carrot roots caused by increased phytoene synthase protein levels. *PLoS ONE*, **4**, e6373.

Machina, A., Ponosov, A. and Voit, E.O. (2010) Automated piecewise power-law modeling of biological systems. *J. Biotechnol.* **149**, 154–165.

Masip, G., Sabalza, M., Pérez-Massot, E., Banakar, R., Cebrian, D., Twyman, R.M., Capell, T., Albajes, R. and Christou, P. (2013) Paradoxical EU agricultural policies on genetically engineered crops. *Trends Plant Sci.* **18**, 312–324.

Méchin, V., Thévenot, C., Guilloux, M.Le, Prioul, J.-L. and Damerval, C. (2007) Developmental analysis of maize endosperm proteome suggests a pivotal role for pyruvate orthophosphate dikinase. *Plant Physiol.* **143**, 1203–1219.

Mutalik, V.K., Guimaraes, J.C., Cambray, G. *et al.* (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods*.

Naqvi, S., Zhu, C., Farre, G. *et al.* (2009) Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three distinct metabolic pathways. *Proc. Natl. Acad. Sci. USA*, **106**, 7762–7767.

Naqvi, S., Zhu, C., Farre, G., Sandmann, G., Capell, T. and Christou, P. (2011) Synergistic metabolism in hybrid corn indicates bottlenecks in the carotenoid pathway and leads to the accumulation of extraordinary levels of the nutritionally important carotenoid zeaxanthin. *Plant Biotechnol. J.* **9**, 384–393.

Nielsen, A.A.K. and Voigt, C.A. (2014) Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks. *Mol. Syst. Biol.* **10**, 763.

Paddon, C.J., Westfall, P.J., Pitera, D.J. *et al.* (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, **496**, 528–532.

Rahmana, Z., Sung, B.H., Yi, J.-Y., Bui, L.M., Lee, J.H. and Kim, S.C. (2014) Enhanced production of n-alkanes in Escherichia coli by spatial organization of biosynthetic pathway enzymes. *J. Biotechnol.* **192 Pt A**, 187–191.

Ravanello, M.P., Ke, D., Alvarez, J., Huang, B. and Shewmaker, C.K. (2003) Coordinate expression of multiple bacterial carotenoid genes in canola leading to altered carotenoid production. *Metab. Eng.* **5**, 255–263.

Salvador, A. (2000) Synergism analysis of biochemical systems. I. Conceptual Framework. *Math. Biosci.* **163**, 105–129.

Savageau, M.A. (1969a) Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.* **25**, 365–369.

Savageau, M.A. (1969b) Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. *J. Theor. Biol.* **25**, 370–379.

Savageau, M.A. (1971) Concepts relating the behavior of biochemical systems to their underlying molecular properties. *Arch. Biochem. Biophys.* **145**, 612–621.

Savageau, M.A. (1975) Significance of autogenously regulated and constitutive synthesis of regulatory proteins in repressible biosynthetic systems. *Nature*, **258**, 208–214.

Shetty, R.P., Endy, D. and Knight, T.F. (2008) Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.* **2**, 5.

Shewmaker, C., Sheehy, J., Daley, M., Colburn, S. and Ke, D. (1999) Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects. *Plant J.* **20**, 401–412X.

da Silva Messias, R., Galli, V., Silva Anjos, E., Dos, S.D. and Rombaldi, C.V. (2014) Carotenoid biosynthetic and catabolic pathways: gene expression and carotenoid content in grains of maize landraces. *Nutrients*, **6**, 546–563.

Sorribas, A. and Savageau, M.A. (1989) A comparison of variant theories of intact biochemical systems. II. Flux-oriented and metabolic control theories. *Math. Biosci.* **94**, 195–238.

Spiess, A.-N. and Neumeyer, N. (2010) An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacol.* **10**, 6.

Uzkudun, M., Marcon, L. and Sharpe, J. (2015) Data-driven modelling of a gene regulatory network for cell fate decisions in the growing limb bud. *Mol. Syst. Biol.* **11**, 815.

Voit, E.O. ed. (1991) *Canonical Nonlinear Modeling. S-System Approach to Understanding Complexity*, New York: Van Nostrand Reinhold.

Voit, E.O. (2013) Biochemical systems theory: a review. *ISRN Biomathematics*, Article ID 897658, doi:10.1155/2013/897658.

Voit, E.O. and Savageau, M.A. (1982) Power-law approach to modeling biological systems. *J. Ferment. Technol.* **60**, 233–241.

Voit, E.O., Goel, G., Chou, I.-C. and Fonseca, L.L. (2009) Estimation of metabolic pathway systems from different data sources. *IET Syst. Biol.* **3**, 513–522.

Way, J.C., Collins, J.J., Keasling, J.D. and Silver, P.A. (2014) Integrating biological redesign: where synthetic biology came from and where it needs to go. *Cell*, **157**, 151–161.

Wolfram, S. (2003) *The Mathematica Book*, 5th edn. Champaign: Wolfram Media Inc.

Ye, X., Al-Babili, S., Klöti, A., Zhang, J., Lucca, P., Beyer, P. and Potrykus, I. (2000) Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science*, **287**, 303–305.

Zhu, C., Naqvi, S., Breitenbach, J., Sandmann, G., Christou, P. and Capell, T. (2008) Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in maize. *Proc. Natl. Acad. Sci. USA*, **105**, 18232–18237.

Zhu, C., Bai, C., Sanahuja, G., Yuan, D., Farré, G., Naqvi, S., Shi, L., Capell, T. and Christou, P. (2010) The regulation of carotenoid pigmentation in flowers. *Arch. Biochem. Biophys.* **504**, 132–141.

