

# Predicting C<sub>4</sub> Photosynthesis Evolution: Modular, Individually Adaptive Steps on a Mount Fuji Fitness Landscape

David Heckmann,<sup>1</sup> Stefanie Schulze,<sup>2</sup> Alisandra Denton,<sup>3</sup> Udo Gowik,<sup>2</sup> Peter Westhoff,<sup>2,4</sup> Andreas P.M. Weber,<sup>3,4</sup> and Martin J. Lercher<sup>1,4,\*</sup>

<sup>1</sup>Institute for Computer Science

<sup>2</sup>Institute for Plant Molecular and Developmental Biology

<sup>3</sup>Institute for Plant Biochemistry

Heinrich Heine University, 40225 Düsseldorf, Germany

<sup>4</sup>Cluster of Excellence on Plant Sciences (CEPLAS)

\*Correspondence: lercher@cs.uni-duesseldorf.de

<http://dx.doi.org/10.1016/j.cell.2013.04.058>

## SUMMARY

An ultimate goal of evolutionary biology is the prediction and experimental verification of adaptive trajectories on macroevolutionary timescales. This aim has rarely been achieved for complex biological systems, as models usually lack clear correlates of organismal fitness. Here, we simulate the fitness landscape connecting two carbon fixation systems: C<sub>3</sub> photosynthesis, used by most plant species, and the C<sub>4</sub> system, which is more efficient at ambient CO<sub>2</sub> levels and elevated temperatures and which repeatedly evolved from C<sub>3</sub>. Despite extensive sign epistasis, C<sub>4</sub> photosynthesis is evolutionarily accessible through individually adaptive steps from any intermediate state. Simulations show that biochemical subtraits evolve in modules; the order and constitution of modules confirm and extend previous hypotheses based on species comparisons. Plant-species-designated C<sub>3</sub>-C<sub>4</sub> intermediates lie on predicted evolutionary trajectories, indicating that they indeed represent transitory states. Contrary to expectations, we find no slowdown of adaptation and no diminishing fitness gains along evolutionary trajectories.

## INTRODUCTION

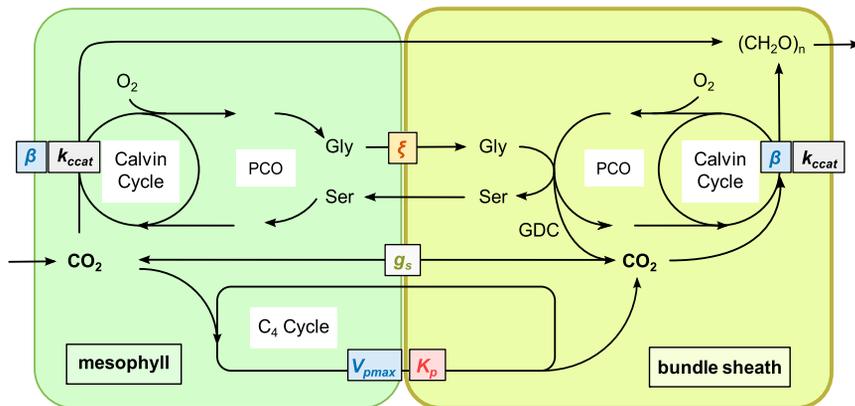
To predict the evolution of biological systems, it is necessary to embed a systems-level model for the calculation of fitness into an evolutionary framework (Papp et al., 2011). However, explicit theories to predict strong correlates of fitness exist for very few complex model systems (Papp et al., 2011; Stern and Orgogozo, 2008). A major example is the stoichiometric metabolic network models of microbial species, which have been used to predict bacterial adaptation to nutrient conditions in laboratory experiments (Fong and Palsson, 2004; Hindré et al., 2012; Ibarra et al., 2002). On a macroevolutionary timescale, related methods

have been applied to predict the outcome and temporal order of reductive genome evolution in endosymbiotic bacteria (Pál et al., 2006; Yizhak et al., 2011). These studies on microbial evolution have employed metabolic yield of biomass production as a correlate of fitness, an approach that cannot be transferred directly to multicellular organisms.

However, it is likely that the efficiency with which limiting resources are converted into biomass precursors is under strong selection across all domains of life. For multicellular eukaryotes, this trait may be most easily studied in plants, which use energy provided by solar radiation to build sugars from water and CO<sub>2</sub>. To fix carbon from CO<sub>2</sub>, plants use the enzyme RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygenase). RuBisCO has a biologically relevant affinity for O<sub>2</sub>, resulting in a toxic product that must be recycled in the energy-consuming metabolic repair pathway known as photorespiration (Maurino and Peterhansel, 2010). The decarboxylation of glycine—a key metabolite within this pathway—by the glycine decarboxylase complex (GDC) releases CO<sub>2</sub>. About 30 million years ago, photorespiration increased to critical levels in many terrestrial ecosystems due to the depletion of atmospheric CO<sub>2</sub>. To circumvent this problem, C<sub>4</sub> photosynthesis evolved to concentrate CO<sub>2</sub> around RuBisCO in specific cell types (Edwards et al., 2010; Sage et al., 2012).

CO<sub>2</sub> first enters mesophyll (M) cells, where most RuBisCO is located in C<sub>3</sub> plants. In contrast, C<sub>4</sub> plants have shifted RuBisCO to neighboring bundle sheath (BS) cells. In the M of C<sub>4</sub> plants, PEPC (phosphoenolpyruvate carboxylase, which does not react with oxygen) catalyzes the primary fixation of CO<sub>2</sub> as bicarbonate. The resulting C<sub>4</sub> acids enter the BS and are decarboxylated, releasing CO<sub>2</sub> in proximity to RuBisCO. BS cells are surrounded by thick cell walls, believed to reduce CO<sub>2</sub> leakage (Kiirats et al., 2002). Such an energy-dependent biochemical CO<sub>2</sub>-concentrating pump is the defining feature of C<sub>4</sub> plants; species differ in the decarboxylating enzyme employed and in the metabolites shuttled between cell types (Drincovich et al., 2011; Furbank, 2011; Pick et al., 2011).

Despite the complexity of C<sub>4</sub> photosynthesis, this trait constitutes a striking example of convergent evolution: it has evolved



**Figure 1. Overview of C<sub>3</sub>-C<sub>4</sub> Biochemistry, Modeled as Two Interacting Cell Types**

CO<sub>2</sub> enters the M and is either fixed by RuBisCO in the M or shuttled to the BS through the C<sub>4</sub> cycle and fixed by RuBisCO there. The resulting C<sub>3</sub> acids are fed into the Calvin cycle. Deleterious fixation of O<sub>2</sub> by RuBisCO leads to photorespiration (PCO). Model parameters are  $\beta$ , the fraction of RuBisCO active sites in the M;  $k_{cat}$ , the maximal turnover rate of RuBisCO;  $\xi$ , the fraction of M derived glycine decarboxylated by GDC in the BS (note that for  $\xi < 1$ , decarboxylation of glycine also takes place in the M);  $V_{pmax}$ , the activity of the C<sub>4</sub> cycle;  $K_p$ , the Michaelis-Menten constant of PEPC for bicarbonate; and  $g_s$ , the BS conductance for gases. See also Figure S2 and Table S2.

independently in more than 60 angiosperm lineages from the ancestral C<sub>3</sub> photosynthesis (Sage et al., 2011). The leaf anatomy typical for C<sub>4</sub> plants—close vein spacing and prominent BS cells, designated “Kranz” anatomy—is also adaptive for C<sub>3</sub> species in environments associated with C<sub>4</sub> evolution (Brodribb et al., 2010). A rudimentary Kranz anatomy was thus likely already present in the C<sub>3</sub> ancestors of C<sub>4</sub> species (Sage et al., 2012), forming a “potentiating” anatomical state (Christin et al., 2011, 2013). Furthermore, all enzymes required for C<sub>4</sub> photosynthesis have orthologs in C<sub>3</sub> species, where they perform unrelated functions. In the evolution of C<sub>4</sub> biochemistry, these enzymes required concerted changes in their cell-type-specific gene expression as well as adjustment of their kinetic properties (Aubry et al., 2011; Gowik and Westhoff, 2011; Sage, 2004).

Some plant species have biochemistry that is intermediate between C<sub>3</sub> and C<sub>4</sub> (Edwards and Ku, 1987). These species possess a rudimentary Kranz anatomy and divide RuBisCO between M and BS cells. Often, however, photorespiratory glycine decarboxylation by GDC is largely shifted to the BS (see Figure 1), resulting in a moderate increase in the CO<sub>2</sub> concentration in BS cells (Sage et al., 2012).

C<sub>4</sub> plants make up 3% of today’s vascular plant species but account for ~25% of terrestrial photosynthesis (Edwards et al., 2010; Sage et al., 2012). How C<sub>4</sub> photosynthesis evolved and why it evolved with such repeatability, are two fundamental questions in plant biology (Sage et al., 2012). Low atmospheric CO<sub>2</sub>/O<sub>2</sub> ratio, heat, aridity, and high light are discussed as important factors promoting C<sub>4</sub> evolution, explaining the abundance of C<sub>4</sub> plants in tropical and subtropical environments (Edwards et al., 2010; Ehleringer et al., 1991). However, C<sub>4</sub> metabolism also allows higher biomass production rates in temperate regions (Beale and Long, 1995). The resulting accelerated growth makes engineering of the C<sub>4</sub> trait into major crops a promising route toward meeting the growing demands on food production (Hibberd et al., 2008). Rational strategies to approach this challenge require a detailed understanding of not only the C<sub>4</sub> state but also the fitness landscape connecting it with the ancestral C<sub>3</sub> biochemistry.

Here, we map the biochemical fitness landscape on which evolution from C<sub>3</sub> to C<sub>4</sub> photosynthesis occurs. Inserting the fitness estimates into a population genetic framework, we then explore the probability distribution of evolutionary trajectories

leading from C<sub>3</sub> to C<sub>4</sub> systems. We thereby predict biochemical evolution in a multicellular eukaryote on macroevolutionary time-scales (Hindré et al., 2012; Papp et al., 2011). Our results show that C<sub>4</sub> evolution is repeatable and predictable in its details. Importantly, experimentally determined parameter sets for C<sub>3</sub>-C<sub>4</sub> intermediates fall well within the clustered distribution of predicted evolutionary trajectories. This agreement not only validates the model but also further provides important insights into the evolutionary nature of these species as transitory states in the evolution toward full C<sub>4</sub> photosynthesis.

## RESULTS

### A Biochemical Model for C<sub>3</sub>-C<sub>4</sub> Evolution

RuBisCO is the most abundant protein on earth, responsible for up to 30% of nitrogen investment and 50% of total protein investment in plants (Ellis, 1979). C<sub>4</sub> plants typically contain lower amounts of RuBisCO per leaf area than C<sub>3</sub> plants (Ghannoum et al., 2011), explaining their lower nitrogen requirements (Brown, 1978). Reduced RuBisCO production is facilitated by higher CO<sub>2</sub> assimilation per RuBisCO protein, allowing C<sub>4</sub> plants to channel protein investment into other processes. In addition, C<sub>4</sub> plants do not need to open their stomata as much as C<sub>3</sub> plants to ensure sufficient internal CO<sub>2</sub> partial pressure, and they thus lose less water in hot and arid environments (Ghannoum et al., 2011). We assume that the overall fitness gain associated with C<sub>4</sub> photosynthesis is proportional to the amount of CO<sub>2</sub> that can be fixed using a given quantity of RuBisCO per leaf area ( $A_c$ ).

To predict the steady-state enzyme-limited net CO<sub>2</sub> assimilation rate,  $A_c$ , from phenotypic parameters, we modified a mechanistic biochemical model developed by von Caemmerer (2000) to describe C<sub>3</sub>-C<sub>4</sub> intermediates (Figure 1 and Experimental Procedures; see also Peisker, 1986). The underlying von Caemmerer model is itself based on models describing gas exchange in C<sub>3</sub> and in C<sub>4</sub> plants (Berry and Farquhar, 1978; Farquhar et al., 1980; von Caemmerer, 1989, 2000); these models have been used and validated in a variety of contexts (Yin and Struik, 2009). An extensive discussion of the model’s generality and the choice of parameters can be found in the von Caemmerer book (2000).

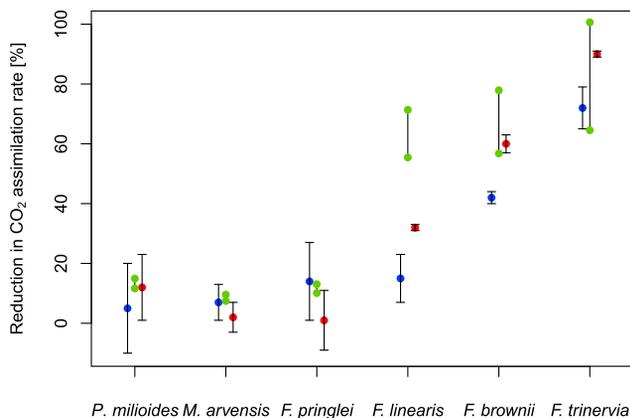
C<sub>3</sub> and C<sub>4</sub> metabolisms represent limiting cases of the model, and representative parameter ranges were derived from C<sub>3</sub> and

C<sub>4</sub> species (Experimental Procedures). Evolution is modeled via changes in the following parameters:  $\beta$ , the fraction of RuBisCO active sites in the M, which ranges from ~95% in C<sub>3</sub> to 0% in some C<sub>4</sub> plants (where all RuBisCO is shifted to the BS);  $k_{ccat}$ , the maximal turnover rate of RuBisCO, which is lower in C<sub>3</sub> plants due to a trade-off with CO<sub>2</sub> specificity (Savir et al., 2010);  $\xi$ , the fraction of glycine derived from unwanted fixation of O<sub>2</sub> in M cells that is decarboxylated by GDC in the BS, ranging from 0 in C<sub>3</sub> to 1 in many C<sub>3</sub>-C<sub>4</sub> intermediates (i.e., activity of the photorespiratory CO<sub>2</sub> pump);  $V_{pmax}$ , quantifying the activity of the C<sub>4</sub> cycle (i.e., the PEPc-dependent CO<sub>2</sub> pump);  $K_p$ , the Michaelis-Menten constant of PEPc (the core protein of the C<sub>4</sub> cycle) for bicarbonate; and  $g_s$ , the BS gas conductance (which quantifies the combined effects of cell geometry and cell wall properties).

Other kinetic parameters for RuBisCO were shown to be strongly linked to  $k_{ccat}$  (Savir et al., 2010) and are modeled accordingly (Extended Experimental Procedures and Figure S1 available online). The model describes the core steps of carbon fixation in communicating M and BS cells (Figure 1). CO<sub>2</sub> and O<sub>2</sub> enter M cells, with diffusion into and out of BS cells ( $g_s$ ). CO<sub>2</sub> can be fixed in both cell types at rates characterized by the allocation ( $\beta$ ) and kinetics ( $k_{ccat}$ ) of RuBisCO. Alternatively, CO<sub>2</sub> may initially be fixed into a C<sub>4</sub> acid through the action of the C<sub>4</sub> cycle in M cells, characterized by the activity ( $V_{pmax}$ ) and the kinetics ( $K_p$ ) of its rate-limiting enzyme, PEPc. The C<sub>4</sub> acids then diffuse into the BS cells, where they are decarboxylated to free CO<sub>2</sub>. We assume PEPc to be rate limiting (von Caemmerer, 2000), and thus neither this part of the C<sub>4</sub> cycle nor the recycling of the CO<sub>2</sub> carrier to the M is modeled explicitly. Finally, due to downregulation of GDC in the M, a fraction of the glycine resulting from the fixation of O<sub>2</sub> in the M is decarboxylated by GDC in BS cells ( $\xi$ ).

The C<sub>3</sub> ancestors of C<sub>4</sub> species likely possessed a potentiating anatomy, characterized by decreased vein spacing and increased BS size (Christin et al., 2011, 2013). These anatomical features enable efficient diffusion of photorespiratory and C<sub>4</sub> cycle metabolites between compartments. C<sub>3</sub> plants that are closely related to C<sub>4</sub> species were further shown to exhibit a specific localization of chloroplasts and mitochondria in the BS cells. This “proto-Kranz” anatomy (Muhaidat et al., 2011) may be necessary for the establishment of a photorespiratory CO<sub>2</sub> pump by allowing the loss of GDC activity in the M to be compensated by the BS (Sage et al., 2012). Accordingly, our model starts from a C<sub>3</sub> state with proto-Kranz anatomy. This morphology can evolve further toward full C<sub>4</sub> Kranz anatomy (McKown and Dengler, 2007) via two main processes: (1) a reduction in the relative number of M cells and (2) an increase of BS cell size. Both processes influence our model exclusively by changing the proportion of RuBisCO allocated to BS cells instead of M cells (i.e., by decreasing  $\beta$ ).

All parameters were normalized to total leaf area. At environmental conditions relevant for the evolution of C<sub>4</sub> photosynthesis and the constant RuBisCO concentration assumed in the model, C<sub>3</sub> and C<sub>4</sub> parameterizations lead to  $A_c$  values of 15.5 and 83.8  $\mu\text{mol m}^{-2} \text{s}^{-1}$ , respectively. These hypothetical  $A_c$  values are assumed to reflect fitness gains during C<sub>4</sub> evolution, even if these fitness gains are in fact partially realized by the channeling of resources from RuBisCO production into other processes.



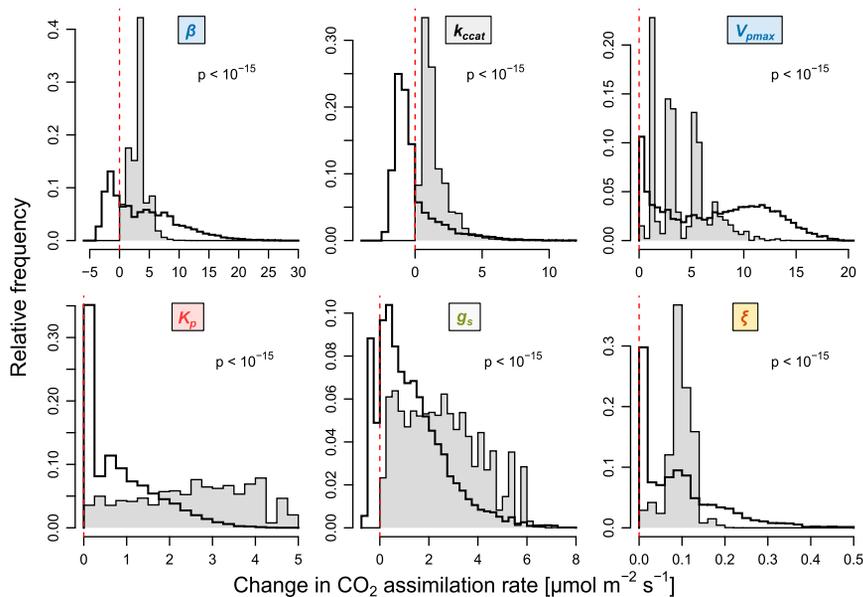
**Figure 2. The Model Predicts the Reduction in Carbon Fixation Rate when the C<sub>4</sub> Cycle Is Reduced by Inhibiting PEPc**

Blue and red dots show  $A_c$  reduction at 1 mM and 4 mM DCDP, respectively, with error bars indicating SD (Brown et al., 1991). Green dots show the range of predicted  $A_c$  reduction at 80%–100% inhibition of the C<sub>4</sub> cycle. See Extended Experimental Procedures for details.

C<sub>4</sub> species have been categorized into three subtypes, depending on the predominant decarboxylating enzyme (NAD malic enzyme, NAD-ME; NADP malic enzyme, NADP-ME; or phosphoenolpyruvate carboxykinase, PEPCK) (Hatch et al., 1975). Our model is compatible with the stoichiometry of all three of these pathways under excess light. This agrees with experimental observations, which show that fitness-relevant traits are independent of C<sub>4</sub> subtype (Ehleringer and Pearcy, 1983; Ghanoum et al., 2001).

One major reason for the generality of our modeling approach is that carbon fixation is largely decoupled from other parts of plant metabolism. When light and nitrogen are available in excess, we thus expect that biomass production is strictly proportional to the carbon fixation rate,  $A_c$ . To confirm this, we coupled our C<sub>3</sub>/C<sub>4</sub> model to a full plant metabolic network (Dal’Molin et al., 2010). The full model can be modified to reflect the different subtypes of C<sub>4</sub> metabolism (NAD-ME, NADP-ME, PEPCK). We sampled the parameter space of our C<sub>3</sub>/C<sub>4</sub> model, using the predicted metabolite fluxes to constrain flux-balance analyses (FBA) of the full model (Oberhardt et al., 2009). For each of the three C<sub>4</sub> subtypes, we demonstrated that biomass production is indeed directly proportional to  $A_c$  (Figure S2; Pearson’s  $R^2 > 0.999$ ). These results support the robustness of our model to differences in the metabolism of different plant lineages.

As long as RuBisCO is active in both M and BS ( $0 < \beta < 1$ ), our model predicts that CO<sub>2</sub> assimilation increases with decreasing M GDC expression (i.e., decreasing  $\xi$ ). This prediction is consistent with experimental data from crosses between C<sub>3</sub>-C<sub>4</sub> intermediate *Moricandia* and C<sub>3</sub> *Brassica* (Hylton et al., 1988). Furthermore, the model predicts the quantitative influence of experimentally suppressed C<sub>4</sub> cycles in phylogenetically diverse C<sub>3</sub>-C<sub>4</sub> intermediates and C<sub>4</sub> plants (Brown et al., 1991) (Figure 2). A discrepancy between model and experiments is observed only for *F. linearis*. In this species, PEPc activity appears to be a sub-optimal predictor for C<sub>4</sub> cycle activity, likely because of insufficient activity of PPK (pyruvate, Pi dikinase) (Ku et al., 1983).



**Figure 3. Realized Fitness Gains Are More Narrowly Distributed Than Potential Fitness Gains**

White bars show potential fitness gains when one parameter is changed towards the C<sub>4</sub> value. Gray bars show fitness gains realized in the evolutionary simulations. Negative values (to the left of the dashed red lines) indicate fitness reductions. Fitness is approximated by CO<sub>2</sub> assimilation rate. Although potential fitness gains vary widely, realized fitness gains are comparable between parameters. The distributions of potential and realized fitness gains are significantly different ( $p < 10^{-15}$  for each parameter, median tests). See also Figure S4.

All 20 involve an interaction between  $\beta$  and  $k_{\text{cat}}$  at intermediate activity of the C<sub>4</sub> cycle ( $V_{p\text{max}}$ ). At these points, changes toward C<sub>4</sub> of  $\beta$  or  $k_{\text{cat}}$  individually increase fitness. However, the C<sub>4</sub> cycle is not sufficiently active to compensate for

the associated reduction in M photosynthetic efficiency when both parameters change simultaneously.

Changes of the model parameters are ultimately caused by DNA mutations of protein coding or regulatory regions, and hence occur in discrete steps. Although each model parameter is known to show genetic variation, we currently lack a detailed understanding of the genotype-phenotype relationships. We thus divided each parameter range into six equidistant phenotypic states, with C<sub>3</sub> and C<sub>4</sub> states as endpoints. Choosing different discretizations did not change the observed patterns (Figure S3), except for  $\xi$  (see Discussion).

### Despite Extensive Epistasis, the C<sub>4</sub> State Is Accessible from Every Point in the Fitness Landscape

The phenotypic parameters that distinguish C<sub>3</sub> from C<sub>4</sub> metabolism span a six-dimensional fitness landscape. Due to functional dependencies between the parameters, this landscape shows strong epistasis: fitness effects of changes in one parameter vary widely depending on the values of other parameters (Figure 3). Parameters differ in their potential influence on fitness. Whereas any individual increase in  $\xi$  raises  $A_c$  by at most  $0.5 \mu\text{mol m}^{-2} \text{s}^{-1}$  (and never decreases fitness), a single increase in  $\beta$  can boost  $A_c$  by as much as  $27 \mu\text{mol m}^{-2} \text{s}^{-1}$  or diminish  $A_c$  by as much as  $3.7 \mu\text{mol m}^{-2} \text{s}^{-1}$ .

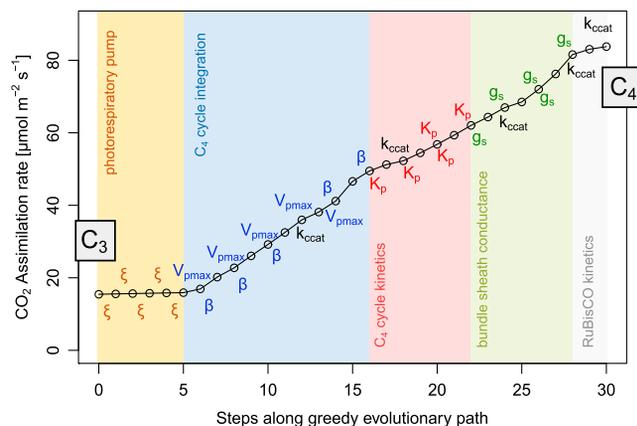
For half of the parameters ( $\beta$ ,  $k_{\text{cat}}$ ,  $g_s$ ), the same parameter change toward C<sub>4</sub> can both increase and decrease fitness, depending on the background provided by the remaining parameter values. This type of interaction has been termed sign epistasis (Weinreich et al., 2005) and affects 5.5% of the discretized fitness landscape (25,145 out of 486,000 pairwise combinations of parameter changes). Sign epistasis can be further classified as reciprocal if changing either of two parameters modifies fitness in one direction, while subsequently adding the second change modifies fitness in the opposite direction (Poelwijk et al., 2011). Reciprocal sign epistasis is a necessary (though not sufficient) condition for the existence of multiple fitness maxima (Poelwijk et al., 2011). The discrete C<sub>3</sub>/C<sub>4</sub> fitness landscape contains only 20 points with reciprocal sign epistasis.

the associated reduction in M photosynthetic efficiency when both parameters change simultaneously.

Maximal fitness is achieved when all parameters reach their C<sub>4</sub> values. Despite strong and often sign-changing epistasis, there is always at least one parameter change (median four changes) toward the C<sub>4</sub> state that increases fitness (Figure S4). Thus, the global fitness optimum is evolutionary accessible (Weinreich et al., 2005) from every position in the landscape. It immediately follows that there are no local maxima, giving the biochemical fitness landscape an exceedingly simple, smooth, “Mount (Mt.) Fuji-like” structure.

### Modular Evolution of a Complex Trait

To evolve from C<sub>3</sub> to C<sub>4</sub> metabolism, our model requires 30 individual mutational changes (five steps in each of the six parameters). Parameters change with unequal probabilities. For example, the mutational target for inactivation of M GDC (increasing  $\xi$ ) is large (Sage, 2004). Active GDC is a multienzyme system consisting of four distinct subunits, and downregulation of any of these will result in reduced GDC activity (Engel et al., 2007). Furthermore, M expression of each subunit is likely regulated by several transcription factor binding sites, each with several nucleotides important for binding. Random mutations at any of these sites are likely to downregulate M GDC expression. This inactivation is sufficient to establish a photorespiratory CO<sub>2</sub> pump, as we assume a low diffusional distance between M and BS cells, as well as a specific subcellular distribution of organelles in the BS (proto-Kranz anatomy). Due to this photorespiratory pump, any RuBisCO present in the BS will operate under increased CO<sub>2</sub> pressure, thereby increasing organismal fitness. Conversely, reduced GDC activity in BS cells would lead to decreased CO<sub>2</sub> pressure in the BS and hence would reduce organismal fitness. Thus, while random mutations may be equally likely to diminish GDC activity in M and in BS cells, only reductions in M activity are likely to be fixed in a population.



**Figure 4. Fitness Changes along the “Greedy” Path through the Fitness Landscape from C<sub>3</sub> to C<sub>4</sub>**

This trajectory always chooses the most likely parameter change, combining mutation and fixation probabilities. The label centered above or below each edge indicates the mutation connecting two states. Evolution along the greedy path is modular (colored areas), except for the RuBisCO turnover rate  $k_{ccat}$ . CO<sub>2</sub> assimilation rate is used as a proxy for fitness. See also Figures S3 and S5.

In contrast to the large mutational target for the reduction of M GDC expression, other parameter changes involve increases in tissue-specific gene expression or changes in enzyme kinetics, which require specific mutations, restricted to only a few potential target nucleotides. Specifically, mutations that increase C<sub>4</sub> cycle activity appear much less likely, as different enzymes need to be upregulated in BS and in M cells, respectively. In the absence of precise estimates, we used plausible relative mutational probabilities for the model parameters (Extended Experimental Procedures). The general evolutionary patterns were found to be robust over a wide range of mutational probabilities and discretizations (Figure S3B).

Once a mutation that changes a model parameter occurs, its probability of fixation in the evolving plant population is determined by the associated change in fitness. Our simulations assume a “strong selection, weak mutation” regime, such that beneficial mutations are fixed in the population before the next mutation occurs (Gillespie, 1983). We estimated the fixation probability using a population genetic model first derived by Kimura (1957), assuming a constant population size of 100,000 individuals.

Each sequence of evolutionary changes linking the C<sub>3</sub> to the C<sub>4</sub> state defines an adaptive trajectory (or path) through the biochemical fitness landscape. The probability of individual steps is estimated as a combination of mutation and fixation probabilities. Figure 4 shows fitness changes associated with a unique “greedy” path, which always realizes the most likely parameter change. Here, changes for all but one of the six parameters are strictly clustered in modules (Figure 4). First, photorespiration is shifted to the BS ( $\xi \uparrow$ ). Next, the C<sub>4</sub> cycle is established ( $V_{pmax} \uparrow$ ), while RuBisCO is simultaneously shifted to the BS ( $\beta \downarrow$ ). Then, the Michaelis-Menten constant of PEPC is adjusted ( $K_p \downarrow$ ). Finally, gas diffusion is reduced ( $g_s \downarrow$ ) in order to avoid leakage of CO<sub>2</sub> from the BS. The only parameter whose changes are not modular in this scenario is the maximal turnover

rate of RuBisCO ( $k_{ccat} \uparrow$ ), which is continuously adjusted along the greedy evolutionary trajectory, reflecting a shifting optimum due to the different CO<sub>2</sub> concentrations in M and BS.

Evolution is not deterministic, and the greedy path shown in Figure 4 represents only one of more than  $10^{19}$  possible sequences of changes from C<sub>3</sub> to C<sub>4</sub>. To more realistically characterize the evolution of C<sub>4</sub> biochemistry, we thus performed Monte Carlo simulations. At each step, we chose one parameter at random, weighted by the relative mutational probabilities. Using the biochemical model (Figure 1), we calculated the fitness change associated with adjusting the chosen parameter one step toward C<sub>4</sub>. The change was accepted with a corresponding probability, derived from the population genetics model.

Despite the strong influence of chance, our Monte Carlo simulations support the same qualitative succession of modular changes in C<sub>4</sub> evolution (Figures S3A and S5). As observed in the greedy path,  $k_{ccat}$  is the only parameter that is continuously adjusted along the evolutionary trajectory, whereas  $\xi$ ,  $V_{pmax}$  combined with  $\beta$ ,  $K_p$ , and  $g_s$  tend to cluster with themselves ( $p < 10^{-15}$  for dispersion higher than random of  $k_{ccat}$  and for modularity of  $\xi$ ,  $V_{pmax}$  combined with  $\beta$ ,  $K_p$ , and  $g_s$ ; median tests for the distance between changes in the same parameter compared to random model).

### Changes Early and Late in Adaptation Lead to Similar Fitness Increases

Strikingly, the greedy path through the fitness landscape (Figure 4) shows an almost linear fitness increase toward the C<sub>4</sub> state, with each evolutionary step resulting in a similar fitness increase. The only exceptions are the early establishment of a photorespiratory pump ( $\xi$ ), the initial establishment of the C<sub>4</sub> cycle ( $V_{pmax}$ ), and the two last adjustments of  $k_{ccat}$ . Thus, realized fitness gains along the greedy evolutionary path are very similar among the different parameters. This finding is in stark contrast to the broad distribution of potential fitness changes across the landscape (Figure 3).

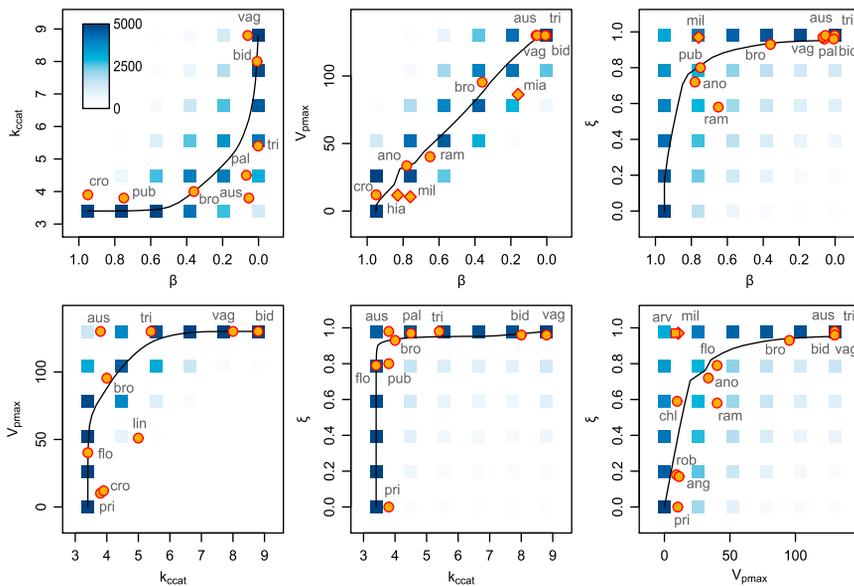
Again, the stochastic evolutionary simulations support the result for the greedy path. Figure 3 shows that the distributions of realized fitness changes are much narrower than those of possible fitness changes. Furthermore, the median of realized fitness gains is similar across parameters, and lies around  $2 \mu\text{mol m}^{-2} \text{s}^{-1}$  for all parameters except  $\xi$ . Accordingly, the time needed until the next parameter change is fixed in the population remains similar along evolutionary trajectories (Figure S6).

### Repeatability of Evolution

The observed modularity and the narrow distributions of realized fitness gains demonstrate that the order of evolutionary changes toward C<sub>4</sub> is not arbitrary. Thus, evolution of this biochemical system is expected to repeat itself qualitatively in different species. Simulated evolutionary trajectories indeed cluster narrowly around a “mean path” ( $p < 10^{-15}$ ; Figures 5 and S7).

### Experimental Data from C<sub>3</sub>-C<sub>4</sub> Intermediates Validate the Model

Our model of C<sub>4</sub> evolution is based on a number of simplifying assumptions and uses rough estimates of relative mutational



**Figure 5. Projections of Trajectories through the Six-Dimensional Fitness Landscape Predicted by the Combined Biochemical and Stochastic Populations Genetics Model**

Density of blue dots is proportional to the number of times a given parameter combination was crossed by a simulated trajectory. Black lines show the mean path of the set of trajectories. Orange dots are the *Flaveria* data described in the text, except for  $V_{pmax}$ , which was capped at  $130 \mu\text{mol m}^{-2} \text{s}^{-1}$ . Abbreviations of species names: ang, *F. angustifolia*; ano, *F. anomala*; aus, *F. australasica*; bid, *F. bidentis*; bro, *F. brownii*; chl, *F. chloraefolia*; cro, *F. cronquistii*; flo, *F. floridana*; lin, *F. linearis*; pal, *F. palmeri*; pri, *F. pringlei*; pub, *F. pubescens*; ram, *F. ramosissima*; rob, *F. robusta*; tri, *F. trinervia*; vag, *F. vaginata*. Diamonds correspond to *Panicum* species: mil, *P. milioides*; hia, *P. hians*; mia, *P. miliaceum*. The square corresponds to *Moricandia arvensis*. See also Figures S6 and S7.

probabilities and population size. To assess its ability to quantitatively describe the evolution of real plants, we compared the model predictions to experimental data from the genera *Flaveria*, *Moricandia*, and *Panicum*. The experimental parameter sets for four plants and one plant correspond to the  $C_3$  and  $C_4$  endpoints, respectively. In addition, our data set included 15 species that have measured biochemical parameters intermediate between  $C_3$  and  $C_4$  (Figure 5); some of these species were previously classified as either  $C_3$  or  $C_4$  based on other criteria (McKown et al., 2005). Each of the intermediate species constitutes a separate point on evolutionary trajectories that started at  $C_3$  biochemistry.

We collected experimental estimates of the biochemical model parameters for each of the 20 species from the literature, and we extended this data set by experimentally determining  $V_{pmax}$  and  $\xi$  for several *Flaveria* species (Experimental Procedures). With few exceptions, the experimentally determined parameter sets indeed lie very close to the predicted mean path through the fitness landscape (Figure 5). The model predicts experimental parameter combinations much better than a null model assuming a random order of evolutionary changes (Figure 6;  $p < 10^{-15}$ , median test).

## DISCUSSION

The evolution of  $C_4$  photosynthesis represents a rare opportunity to predict the functional evolution of a complex system: a closed six-parameter model calculates a phenotypic variable ( $A_c$ ) of high relevance to fitness. The comparison to experimental data from diverse  $C_3$ - $C_4$  intermediates confirms the model's ability to quantitatively predict biochemical evolution over a timescale of several million years (Sage et al., 2012). While the majority of the data describe the genus *Flaveria*, the model also correctly predicts data from two phylogenetically distant genera (Figure 5). Comparisons to additional  $C_3$ - $C_4$  intermediates are currently limited by the availability of species-specific protocols for the separation of BS and M cells.

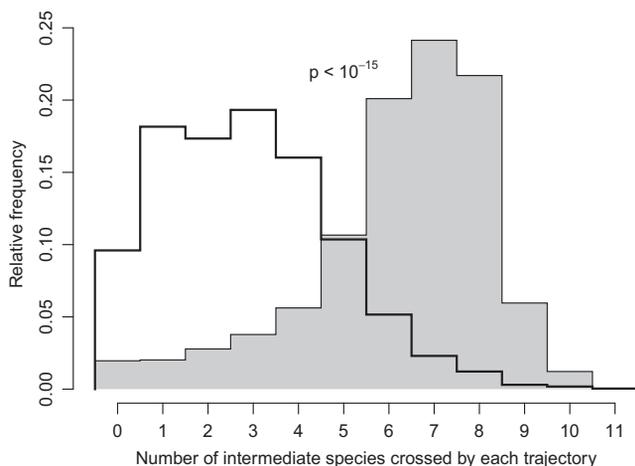
A hypothesis for the evolutionary succession of biochemical and morphological changes in the evolution of the  $C_4$  syndrome was previously derived from phylogenetically informed analyses of  $C_3$ - $C_4$  intermediates (Sage et al., 2012). This hypothesis assumes modular biochemical changes, starting with a shift of photorespiration to the BS, followed by the establishment of a  $C_4$  cycle in conjunction with a shift of RuBisCO to the BS, and finally an optimization stage in which parameters are fine-tuned. Our simulations support this scenario, narrowing it further by indicating that upregulation of the  $C_4$  cycle usually precedes a shift of RuBisCO to the BS (Figure S3) even after previous establishment of a photorespiratory pump.

As expected due to the stochastic nature of evolution, the simulations indicate that modules are not strict and that the order of events may vary between independently evolving species. In particular, we find that the initial establishment of a photorespiratory pump (or  $C_2$  cycle) is typical of evolutionary trajectories toward  $C_4$  photosynthesis but may not be mandatory, as suggested previously (Sage et al., 2012).

## Model Assumptions

While our model tracks changes in a phenotypic biochemical space, evolution is ultimately based on genomic mutations. We used qualitative reasoning when choosing relative mutational probabilities and the distribution of discrete steps linking  $C_3$  and  $C_4$  states. The sensitivity analysis (Figure S3B) demonstrates that other parameterizations lead to qualitatively very similar results. The only exception is the early establishment of a photorespiratory pump ( $\xi$ ), which occurs with high probability only when the large mutational target for deactivation of the M GDC is taken into account.

The full  $C_4$  cycle requires expression shifts in at least four separate enzymes. At each point in evolution, one of the enzymes that constitute the  $C_4$  cycle will be rate limiting, making it the next target for fitness-enhancing upregulation. Distinct implementations of the  $C_4$  cycle were shown to overlap in a



**Figure 6. Distribution of the Number of Different C<sub>3</sub>-C<sub>4</sub> Intermediate Species Whose Experimental Parameter Combinations Are Crossed by Each Single Predicted Trajectory**

The combined biochemical and population genetics model (gray) fits the experimental data much better than a random model that ignores fitness effects (white) ( $p < 10^{-15}$ , median test). The parameter sets for *F. robusta*, *F. pringlei*, *F. cronquistii*, *F. angustifolia*, and *F. vaginata* are located at the C<sub>3</sub> or C<sub>4</sub> endpoints and hence crossed by every trajectory; they were excluded from this analysis.

single species (Furbank, 2011; Pick et al., 2011), potentially increasing the size of the mutational target. Our model uses the central enzyme PEPC to represent the complete pathway, accounting for the complexity of the C<sub>4</sub> cycle by using a low relative mutational probability.

### A Simple, Mt. Fuji-like Biochemical Fitness Landscape

We found that the biochemical fitness landscape is exceedingly smooth: there are no local maxima besides the C<sub>4</sub> endpoint, as there is always at least one parameter change toward the C<sub>4</sub> value that increases the CO<sub>2</sub> fixation rate.

Comparison to experimental data from C<sub>3</sub>-C<sub>4</sub> intermediate species indicates that our model indeed captures their evolutionary dynamics. The single-peaked fitness landscape suggests that these species are transitory states rather than evolutionary dead ends, continuously evolving toward the full C<sub>4</sub> syndrome as long as selective environmental conditions persist. The origin of *Flaveria* C<sub>4</sub> traits in the past 5 million years, together with the unusually large number of C<sub>3</sub>-C<sub>4</sub> intermediate species in this genus (Sage et al., 2012), is consistent with this notion.

Half of the parameters in our model exhibit sign epistasis (Figure 3). Certain evolutionary trajectories thus involve reductions in fitness and are deemed not accessible (Weinreich et al., 2005); their inaccessibility contributes to the clustering of evolutionary trajectories. The paucity of reciprocal sign epistasis provides a partial explanation for the smooth landscape structure (Poelwijk et al., 2011).

Fitness landscapes resulting from interactions of mutations within the same gene can be rough and multi-peaked (Weinreich et al., 2006). However, experimental fitness landscapes spanned by independently encoded functional units are similar in structure to the biochemical fitness landscape observed here: inter-

actions among alleles of different genes rarely exhibit sign epistasis and often lead to simple, single-peaked landscapes (Chou et al., 2011; Khan et al., 2011; but see Kvitek and Sherlock, 2011).

### Evolutionary Trajectories

Due to extensive sign epistasis among mutations within the same coding sequence, it was concluded that protein evolution may be largely reproducible and even predictable (Lozovsky et al., 2009; Weinreich et al., 2006). Despite the relatively low incidence of sign epistasis, we find that the same is true for the evolution of a complex biochemical system. Thus, different plants that independently “replay the tape of evolution” toward C<sub>4</sub> photosynthesis tend to follow similar trajectories of phenotypic changes (Figure 5). This resembles the high level of phenotypic and often genotypic parallelism in microbial evolution observed in experiments (Hindré et al., 2012) and predicted based on stoichiometric metabolic modeling (Fong and Palsson, 2004; Ibarra et al., 2002).

To explain the polyphyly of the C<sub>4</sub> syndrome, it has been hypothesized that each evolutionary step comes with a fitness gain (Gowik and Westhoff, 2011; Sage, 2004). We found that reality may be even more extreme: the fitness gain achieved by each individual change remained comparable along evolutionary trajectories (Figure 4). Accordingly, realized fitness advantages were much more similar across parameters than expected for random trajectories (Figure 3). This differs markedly both from theoretical expectations (Fisher, 1930; Orr, 2005) and from experimental observations in some genetic landscapes (Chou et al., 2011; Khan et al., 2011), which find diminishing fitness increases and a slowdown of adaptation along adaptive trajectories.

In the case of C<sub>4</sub> evolution, late-changing parameters (C<sub>4</sub> cycle kinetics, BS conductance) benefit from an already optimized background provided by previous evolution. Because everything else required for C<sub>4</sub> photosynthesis is already in place, their potential to contribute favorably to fitness is increased. Accordingly, we find no clear pattern of decelerated evolution along simulated trajectories, except for the last steps in PEPC kinetics and for late-occurring fixations of the now-superfluous photorespiratory pump (Figure S6). Conversely, the first few steps in C<sub>4</sub> evolution (initial establishment of CO<sub>2</sub> pumps) are only weakly selected, as only little RuBisCO is available in the BS at this time (Figure 4). Their fixation thus takes substantially longer than later changes (Figure S6): the first step is the most difficult one, also in C<sub>4</sub> evolution.

Why do C<sub>3</sub> plants still dominate many habitats, despite the simple, single-peaked fitness landscape and the substantial fitness gains resulting from individual evolutionary changes toward C<sub>4</sub> metabolism? A partial explanation is provided by weak selection on the first mutations. Furthermore, C<sub>4</sub> metabolism is strongly favored by selection only under specific environmental conditions, such as drought, high temperatures, and high light (excluding, for example, plants in woodlands). Finally, the potentiating Kranz-like anatomy in the C<sub>3</sub> ancestors of C<sub>4</sub> lineages (Christin et al., 2011, 2013; Sage et al., 2012) is not present in many other lineages, making the evolution of C<sub>4</sub> metabolism in these species unlikely.

The evolutionary dynamics uncovered above may shed light onto plans for experimental evolution of  $C_4$  photosynthesis in  $C_3$  plants through the application of increased selection pressure (Sage and Sage, 2007). Our results indicate that this endeavor may be accelerated by genetically engineering the first, slow steps of  $C_4$  evolution. In particular, it may be advisable to pre-establish a photorespiratory  $CO_2$  pump by knocking out M-specific GDC expression.

## EXPERIMENTAL PROCEDURES

### Biochemical Model and Fitness Landscape

The steady-state enzyme-limited net  $CO_2$  assimilation rate ( $A_c$ ) was used as a proxy for fitness of  $C_3$ ,  $C_4$ , and intermediate evolutionary phenotypes. To predict  $A_c$  from phenotypic parameters, we slightly modified a mechanistic biochemical model for  $C_3$ - $C_4$  intermediates developed by von Caemmerer (2000) (Figure 1).

The  $CO_2$  assimilation rates in the M and in the BS are calculated from the respective rates of carboxylation, oxygenation, and mitochondrial respiration (in addition to photorespiration). We assume constant concentrations of  $CO_2$  (250  $\mu$ bar) and  $O_2$  (200 mbar) in M cells. Carboxylation and oxygenation are modeled as inhibitory Michaelis-Menten kinetics. RuBisCO kinetic parameters were shown to be subject to trade-offs (Savir et al., 2010); accordingly, we model these parameters as a function of RuBisCO maximal turnover rate ( $k_{ccat}$ ). Activity of the  $C_4$  cycle is assumed to be limited by PEPC activity and to follow Michaelis-Menten kinetics. The parameterization corresponds to a temperature of 25°C. The resulting set of equations can be solved for  $A_c$  in closed form. Equations, parameters, and further details are given in Extended Experimental Procedures and Tables S1 and S2.

For each evolving model parameter, we obtained representative  $C_3$  and  $C_4$  values (see below). The resulting range was subdivided into equidistant steps, leading to a discrete six-dimensional phenotype space. Based on the biochemical model, we calculated  $A_c$  for each parameter combination.

### Calculation of Evolutionary Trajectories

We simulated a set of 5,000 evolutionary trajectories on the discrete fitness landscape, starting with the  $C_3$  state. At each step, a trait (parameter) to be changed was chosen at random, with relative probabilities derived from current qualitative knowledge about the genetic complexity of the trait (Extended Experimental Procedures). We estimated selection coefficients ( $s$ ) as the relative difference in  $A_c$  between ancestral and derived state, calculated using the biochemical model. We assumed a randomly mating population of diploid hermaphrodites, with incomplete dominance of mutations. The derived state was accepted with its probability of fixation, estimated using a formula first derived by Kimura (1957). We repeated the simulation process until reaching the  $C_4$  parameter set.

To calculate a mean path from the set of 5,000 simulated trajectories, we averaged each parameter at each step (i.e.,  $\beta$  at the first step of the mean path is the average of  $\beta$  values across the first steps of all simulated trajectories, etc.). Parameters were normalized to the interval [0,1]. Clustering of trajectories was quantified by calculating for each trajectory the mean of the normalized point-wise Manhattan distances to this mean trajectory. This measure is closely related to the recently introduced mean path divergence (Lobkovsky et al., 2011).

To estimate evolutionary modularity for each parameter, we used a distance measure defined as the number of other fixation events that occurred between two subsequent fixation events of the same parameter.  $V_{pmax}$  and  $\beta$  evolve together and were treated as a joint parameter in this context.

To determine a greedy trajectory through the landscape, we changed at each step the parameter that maximized the product of mutational probability and probability of fixation.

### Comparison to Experimental Data

Data for the partitioning of RuBisCO between M and BS cells ( $\beta$ ) and RuBisCO turnover rates ( $k_{ccat}$ ), as well as PEPC activities ( $V_{pmax}$ ) and decarboxylation of

M-derived glycine in the BS ( $\xi$ ) for *Moricandia* and *Panicum*, were obtained from the literature. We assayed PEPC activity in leaf extracts (summarized by Ashton et al., 1990) from 14 *Flaveria* species as a proxy for  $V_{pmax}$ .  $\xi$  was estimated for 14 *Flaveria* species by comparing the transcript levels of glycine decarboxylase P subunit genes that are expressed specifically in the BS (*glcpA*) to those expressed in all inner leaf tissues (*glcpD*). GldP transcript levels in leaves of 14 *Flaveria* species were determined by RNA sequencing. Data on  $K_p$  and  $g_s$  in intermediate species were not available. See Extended Experimental Procedures for more details on experimental data. We mapped experimental parameter values to the closest point in the discrete space of the model fitness landscape.

### Random Null Model and Statistical Methods

To assess the statistical significance of our findings, we used a random null model to predict evolutionary trajectories. In this model, each trajectory starts with the  $C_3$  state and evolves randomly, i.e., with equal probability for each directed parameter change, until the  $C_4$  state is reached.

All simulations and statistical analyses were performed in the R environment (R Development Core Team, 2010). Statistical significance was assessed using Fisher's exact test and the median test implemented in the coin package (Hothorn et al., 2006).

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and two tables and can be found with this article online at <http://dx.doi.org/10.1016/j.cell.2013.04.058>.

## ACKNOWLEDGMENTS

We thank Veronica Maurino, Itai Yanai, Eugene Koonin, Joachim Krug, and Andrea Bräutigam for helpful discussions. Computational support and infrastructure was provided by the "Center for Information and Media Technology" (ZIM) at the Heinrich Heine University Düsseldorf. This work was supported by the Deutsche Forschungsgemeinschaft (IRTG 1525 to D.H. and A.D.; FOR 1186 to S.S.; EXC 1028 to M.J.L., A.P.M.W., and P.W.; and CRC 680 to M.J.L.).

Received: December 18, 2012

Revised: March 21, 2013

Accepted: April 23, 2013

Published: June 20, 2013

## REFERENCES

- Ashton, A.R., Burnell, J.N., Furbank, R.T., Jenkins, C.L.D., and Hatch, M.D. (1990). The enzymes in  $C_4$  photosynthesis. In *Enzymes of Primary Metabolism*, P.M. Dey and J.B. Harboene, eds. (London, UK: Academic Press), pp. 39–72.
- Aubry, S., Brown, N.J., and Hibberd, J.M. (2011). The role of proteins in  $C_3$  plants prior to their recruitment into the  $C_4$  pathway. *J. Exp. Bot.* 62, 3049–3059.
- Beale, C.V., and Long, S.P. (1995). Can perennial  $C_4$  grasses attain high efficiencies of radiant energy-conversion in cool climates. *Plant Cell Environ.* 18, 641–650.
- Berry, J.A., and Farquhar, G.D. (1978). The  $CO_2$  concentrating function of  $C_4$  photosynthesis: a biochemical model. In *Proceedings of the Fourth International Congress on Photosynthesis Biochemical Society*, London, pp. 119–131.
- Brodribb, T.J., Feild, T.S., and Sack, L. (2010). Viewing leaf structure and evolution from a hydraulic perspective. *Funct. Plant Biol.* 37, 488–498.
- Brown, R.H. (1978). A difference in N use efficiency in  $C_3$  and  $C_4$  plants and its implications in adaptation and evolution. *Crop Sci.* 18, 93–98.
- Brown, R.H., Byrd, G.T., and Black, C.C. (1991). Assessing the degree of  $c_4$  photosynthesis in  $c_3$ - $c_4$  species using an inhibitor of phosphoenolpyruvate carboxylase. *Plant Physiol.* 97, 985–989.

- Chou, H.H., Chiu, H.C., Delaney, N.F., Segrè, D., and Marx, C.J. (2011). Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* 332, 1190–1192.
- Christin, P.A., Sage, T.L., Edwards, E.J., Ogburn, R.M., Khoshraves, R., and Sage, R.F. (2011). Complex evolutionary transitions and the significance of C<sub>3</sub>-C<sub>4</sub> intermediate forms of photosynthesis in Molluginaceae. *Evolution* 65, 643–660.
- Christin, P.A., Osborne, C.P., Chatelet, D.S., Columbus, J.T., Besnard, G., Hodkinson, T.R., Garrison, L.M., Vorontsova, M.S., and Edwards, E.J. (2013). Anatomical enablers and the evolution of C<sub>4</sub> photosynthesis in grasses. *Proc. Natl. Acad. Sci. USA* 110, 1381–1386.
- Dal'Molin, C.G., Quek, L.E., Palfreyman, R.W., Brumley, S.M., and Nielsen, L.K. (2010). C4GEM, a genome-scale metabolic model to study C<sub>4</sub> plant metabolism. *Plant Physiol.* 154, 1871–1885.
- Drincovich, M.F., Lara, M.V., Andreo, C.S., and Mauro, V.G. (2011). Evolution of C<sub>4</sub> decarboxylases: Different solutions for the same biochemical problem: provision of CO<sub>2</sub> in Bundle Sheath Cells. In C<sub>4</sub> photosynthesis and related CO<sub>2</sub> concentration mechanisms, A.S. Raghavendra and R.F. Sage, eds. (Dordrecht: Springer), pp. 277–300.
- Edwards, G.E., and Ku, M.S.B. (1987). Biochemistry of C<sub>3</sub>-C<sub>4</sub> intermediates. In The biochemistry of plants, Volume 10 (New York: Academic Press, Inc.), pp. 275–325.
- Edwards, E.J., Osborne, C.P., Strömberg, C.A.E., Smith, S.A., Bond, W.J., Christin, P.A., Cousins, A.B., Duvall, M.R., Fox, D.L., Freckleton, R.P., et al.; C4 Grasses Consortium. (2010). The origins of C<sub>4</sub> grasslands: integrating evolutionary and ecosystem science. *Science* 328, 587–591.
- Ehleringer, J., and Pearcy, R.W. (1983). Variation in Quantum Yield for CO<sub>2</sub> Uptake among C<sub>3</sub> and C<sub>4</sub> Plants. *Plant Physiol.* 73, 555–559.
- Ehleringer, J.R., Sage, R.F., Flanagan, L.B., and Pearcy, R.W. (1991). Climate change and the evolution of C<sub>4</sub> photosynthesis. *Trends Ecol. Evol.* 6, 95–99.
- Ellis, R.J. (1979). Most abundant protein in the world. *Trends Biochem. Sci.* 4, 241–244.
- Engel, N., van den Daele, K., Kolkisaoglu, U., Morgenthal, K., Weckwerth, W., Pärnik, T., Keerberg, O., and Bauwe, H. (2007). Deletion of glycine decarboxylase in Arabidopsis is lethal under nonphotorespiratory conditions. *Plant Physiol.* 144, 1328–1335.
- Farquhar, G.D., Caemmerer, S., and Berry, J.A. (1980). A biochemical model of photosynthetic CO<sub>2</sub> assimilation in leaves of C<sub>3</sub> species. *Planta* 149, 78–90.
- Fisher, R.A. (1930). *The Genetical Theory of Natural Selection* (Oxford: Oxford Univ. Press).
- Fong, S.S., and Palsson, B.O. (2004). Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.* 36, 1056–1058.
- Furbank, R.T. (2011). Evolution of the C<sub>4</sub> photosynthetic mechanism: are there really three C<sub>4</sub> acid decarboxylation types? *J. Exp. Bot.* 62, 3103–3108.
- Ghannoum, O., von Caemmerer, S., and Conroy, J.P. (2001). Carbon and water economy of Australian NAD-ME and NADP-ME C<sub>4</sub> grasses. *Funct. Plant Biol.* 28, 213–223.
- Ghannoum, O., Evans, J.R., and von Caemmerer, S. (2011). Nitrogen and water use efficiency of C<sub>4</sub> plants. In C<sub>4</sub> Photosynthesis and Related CO<sub>2</sub> Concentrating Mechanisms, A.S. Raghavendra and R.F. Sage, eds. (Dordrecht, The Netherlands: Springer), pp. 129–146.
- Gillespie, J.H. (1983). A simple stochastic gene substitution model. *Theor. Popul. Biol.* 23, 202–215.
- Gowik, U., and Westhoff, P. (2011). The path from C<sub>3</sub> to C<sub>4</sub> photosynthesis. *Plant Physiol.* 155, 56–63.
- Hatch, M.D., Kagawa, T., and Craig, S. (1975). Subdivision of C<sub>4</sub>-pathway species based on differing C<sub>4</sub> acid decarboxylating systems and ultrastructural features. *Funct. Plant Biol.* 2, 111–128.
- Hibberd, J.M., Sheehy, J.E., and Langdale, J.A. (2008). Using C<sub>4</sub> photosynthesis to increase the yield of rice-rationale and feasibility. *Curr. Opin. Plant Biol.* 11, 228–231.
- Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. (2012). New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nat. Rev. Microbiol.* 10, 352–365.
- Hothorn, T., Hornik, K., van de Wiel, M.A., and Zeileis, A. (2006). A Lego system for conditional inference. *Am. Stat.* 60, 257–263.
- Hylton, C.M., Rawsthorne, S., Smith, A.M., Jones, D.A., and Woolhouse, H.W. (1988). Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C<sub>3</sub>-C<sub>4</sub> intermediate species. *Planta* 175, 452–459.
- Ibarra, R.U., Edwards, J.S., and Palsson, B.O. (2002). *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 420, 186–189.
- Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E., and Cooper, T.F. (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* 332, 1193–1196.
- Kiirats, O., Lea, P.J., Franceschi, V.R., and Edwards, G.E. (2002). Bundle sheath diffusive resistance to CO<sub>2</sub> and effectiveness of C<sub>4</sub> photosynthesis and refixation of photorespired CO<sub>2</sub> in a C<sub>4</sub> cycle mutant and wild-type *Amaranthus edulis*. *Plant Physiol.* 130, 964–976.
- Kimura, M. (1957). Some problems of stochastic processes in genetics. *Ann. Math. Stat.* 28, 882–901.
- Ku, M.S.B., Monson, R.K., Littlejohn, R.O., Jr., Nakamoto, H., Fisher, D.B., and Edwards, G.E. (1983). Photosynthetic characteristics of C<sub>3</sub>-C<sub>4</sub> intermediate *Flaveria* species: I. Leaf anatomy, photosynthetic responses to O<sub>2</sub> and CO<sub>2</sub>, and activities of key enzymes in the C<sub>3</sub> and C<sub>4</sub> pathways. *Plant Physiol.* 71, 944–948.
- Kvitek, D.J., and Sherlock, G. (2011). Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet.* 7, e1002056.
- Lobkovsky, A.E., Wolf, Y.I., and Koonin, E.V. (2011). Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput. Biol.* 7, e1002302.
- Lozovsky, E.R., Chookajorn, T., Brown, K.M., Imwong, M., Shaw, P.J., Kamchonwongpaisan, S., Neafsey, D.E., Weinreich, D.M., and Hartl, D.L. (2009). Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc. Natl. Acad. Sci. USA* 106, 12025–12030.
- Mauro, V.G., and Peterhansel, C. (2010). Photorespiration: current status and approaches for metabolic engineering. *Curr. Opin. Plant Biol.* 13, 249–256.
- McKown, A.D., and Dengler, N.G. (2007). Key innovations in the evolution of Kranz anatomy and C<sub>4</sub> vein pattern in *Flaveria* (Asteraceae). *Am. J. Bot.* 94, 382–399.
- McKown, A.D., Moncalvo, J.-M., and Dengler, N.G. (2005). Phylogeny of *Flaveria* (Asteraceae) and inference of C<sub>4</sub> photosynthesis evolution. *Am. J. Bot.* 92, 1911–1928.
- Muhaidat, R., Sage, T.L., Frohlich, M.W., Dengler, N.G., and Sage, R.F. (2011). Characterization of C<sub>3</sub>–C<sub>4</sub> intermediate species in the genus *Heliotropium* L. (Boraginaceae): anatomy, ultrastructure and enzyme activity. *Plant Cell Environ.* 34, 1723–1736.
- Oberhardt, M.A., Palsson, B.O., and Papin, J.A. (2009). Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5, 320.
- Orr, H.A. (2005). The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* 6, 119–127.
- Pál, C., Papp, B., Lercher, M.J., Csermely, P., Oliver, S.G., and Hurst, L.D. (2006). Chance and necessity in the evolution of minimal metabolic networks. *Nature* 440, 667–670.
- Papp, B., Notebaart, R.A., and Pál, C. (2011). Systems-biology approaches for predicting genomic evolution. *Nat. Rev. Genet.* 12, 591–602.
- Peisker, M. (1986). Models of carbon metabolism in C<sub>3</sub>-C<sub>4</sub> intermediate plants as applied to the evolution of C<sub>4</sub> photosynthesis. *Plant Cell Environ.* 9, 627–635.
- Pick, T.R., Bräutigam, A., Schlüter, U., Denton, A.K., Colmsee, C., Scholz, U., Fahnenstich, H., Pieruschka, R., Rascher, U., Sonnewald, U., and Weber, A.P. (2011). Systems analysis of a maize leaf developmental gradient redefines the

- current C<sub>4</sub> model and provides candidates for regulation. *Plant Cell* 23, 4208–4220.
- Poelwijk, F.J., Tănase-Nicola, S., Kiviet, D.J., and Tans, S.J. (2011). Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J. Theor. Biol.* 272, 141–144.
- R Development Core Team. (2010). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
- Sage, R.F. (2004). The evolution of C<sub>4</sub> photosynthesis. *New Phytol.* 161, 341–370.
- Sage, R.F., and Sage, T.L. (2007). Learning from nature to develop strategies for directed evolution of C<sub>4</sub> rice. In *Charting New Pathways to C<sub>4</sub> Rice*, J.E. Sheehy, P.L. Mitchell, and B. Hardy, eds. (Hackensack, NJ, USA: World Scientific Publishing), pp. 195–216.
- Sage, R.F., Christin, P.A., and Edwards, E.J. (2011). The C<sub>4</sub> plant lineages of planet Earth. *J. Exp. Bot.* 62, 3155–3169.
- Sage, R.F., Sage, T.L., and Kocacinar, F. (2012). Photorespiration and the evolution of C<sub>4</sub> photosynthesis. *Annu. Rev. Plant Biol.* 63, 19–47.
- Savir, Y., Noor, E., Milo, R., and Tlustý, T. (2010). Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc. Natl. Acad. Sci. USA* 107, 3475–3480.
- Stern, D.L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? *Evolution* 62, 2155–2177.
- von Caemmerer, S. (1989). A model of photosynthetic CO<sub>2</sub> assimilation and carbon-isotope discrimination in leaves of certain C<sub>3</sub>-C<sub>4</sub> intermediates. *Planta* 178, 463–474.
- von Caemmerer, S. (2000). *Biochemical Models of Leaf Photosynthesis* (Collingwood, Australia: CSIRO Publishing).
- Weinreich, D.M., Watson, R.A., and Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59, 1165–1174.
- Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312, 111–114.
- Yin, X., and Struik, P.C. (2009). C<sub>3</sub> and C<sub>4</sub> photosynthesis models: an overview from the perspective of crop modelling. *NJAS-Wagen. J. Life Sci.* 57, 27–38.
- Yizhak, K., Tuller, T., Papp, B., and Ruppin, E. (2011). Metabolic modeling of endosymbiont genome reduction on a temporal scale. *Mol. Syst. Biol.* 7, 479.

## EXTENDED EXPERIMENTAL PROCEDURES

### Biochemical Model

In the following, we list the equations for the biochemical model. The model is slightly modified from the model of enzyme-limited C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis given by von Caemmerer (2000). Parameter dimensions are listed in Table S1.

The maximal RuBisCO activity per leaf area in the mesophyll ( $V_{mmax}$ ) and the bundle sheath ( $V_{smax}$ ) are given as a function of the fraction of RuBisCO active sites in the mesophyll ( $\beta$ ), the total leaf RuBisCO concentration ( $E_{tot}$ ), and the maximal RuBisCO turnover rate ( $k_{ccat}$ ):

$$V_{mmax} = \beta E_{tot} k_{ccat}$$

$$V_{smax} = (1 - \beta) E_{tot} k_{ccat}$$

The CO<sub>2</sub> assimilation rate in the mesophyll ( $A_m$ ) comprises the rate of carboxylation ( $V_{cm}$ ) and oxygenation ( $V_{om}$ ) and the mitochondrial respiration other than photorespiration ( $R_m$ ) in the mesophyll.  $V_{cm}$  and  $V_{om}$  are modeled as inhibitory Michaelis-Menten kinetics:

$$V_{cm} = \frac{C_m V_{mmax}}{C_m + K_C \left(1 + \frac{O_m}{K_O}\right)}$$

$$V_{om} = \frac{O_m V_{mmax}}{O_m + K_O \left(1 + \frac{C_m}{K_C}\right)}$$

$$A_m = V_{cm} - 0.5V_{om} - R_m = \frac{(C_m - \gamma^* O_m) V_{mmax}}{C_m + K_C + O_m \frac{K_C}{K_O}} - R_m$$

$C_m$  and  $O_m$  represent the CO<sub>2</sub> and O<sub>2</sub> partial pressure in the mesophyll chloroplasts, respectively.  $K_C$  and  $K_O$  are Michaelis-Menten constants of RuBisCO for CO<sub>2</sub> and O<sub>2</sub>, respectively.  $\gamma^*$  is a function of the RuBisCO specificity ( $S_{C/O}$ ):

$$\gamma^* = \frac{0.5}{S_{C/O}}$$

Activity of the C<sub>4</sub> cycle is assumed to be limited by PEPC activity and is given by Michaelis-Menten kinetics:

$$V_p = \frac{C_m V_{pmax}}{C_m + K_p}$$

CO<sub>2</sub> assimilation in the bundle sheath ( $A_s$ ) is also catalyzed by RuBisCO, so kinetics as for  $A_m$  are applied.  $A_s$  can also be expressed as a function of the activity of the photorespiratory pump ( $\xi$ ), the amount of photorespiration in the mesophyll ( $V_{om}$ ), the C<sub>4</sub> cycle, and the rate of CO<sub>2</sub> leakage from the bundle sheath ( $L$ ):

$$A_s = \frac{(C_s - \gamma^* O_s) V_{smax}}{C_s + K_C + O_s \frac{K_C}{K_O}} - R_s = \xi(0.5V_{om}) + V_p - L,$$

where  $L$  is given by

$$L = g_s(C_s - C_m),$$

and  $g_s$  is the bundle sheath conductance for CO<sub>2</sub>. As  $C_s$  is given by

$$C_s = \frac{V_p + 0.5\xi V_{om} - A_s}{g_s} + C_m,$$

and  $O_s$  is

$$O_s = \frac{A_s}{0.047g_s} + O_m,$$

a second degree polynomial with respect to  $A_s$  is obtained. The smaller solution for  $A_s$  is chosen and given by:

$$a = 1 - \frac{1}{0.047} \frac{K_C}{K_O}$$

$$b = -V_p - 0.5\xi V_{om} - g_s C_m - V_{smax} + R_s - g_s \left( K_C + O_m \frac{K_C}{K_O} \right) - \frac{\gamma_* V_{smax} + \frac{K_C}{K_O} R_s}{0.047}$$

$$c = (V_{smax} - R_s)(V_p + 0.5\xi V_{om} + g_s C_m) - V_{smax} g_s \gamma_* O_m - R_s g_s \left( K_C + O_m \frac{K_C}{K_O} \right)$$

$$A_s = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

The enzyme limited net  $CO_2$  assimilation rate  $A_c$  equals the sum of net assimilation in mesophyll and bundle sheath:

$$A_c = A_s + A_m$$

$E_{tot}$  was set to  $19.35 \mu\text{mol m}^{-2}$  and mitochondrial respiration was scaled to RuBisCO activity as suggested by von Caemmerer (2000). The mesophyll  $CO_2$  and  $O_2$  partial pressures ( $C_m$ ,  $O_m$ , respectively) in the model were set to 250  $\mu\text{bar}$  and 200 mbar, respectively; parameterization corresponds to a temperature of  $25^\circ\text{C}$ . Heat and a high  $O_2/CO_2$  ratio promote photorespiration in an exponential manner (e.g., Ehleringer et al., 1991), so extreme environmental conditions may further increase the benefit of  $CO_2$  concentration mechanisms.

### RuBisCO Kinetic Constants

Savir et al. (2010) showed that constraints on the evolution of RuBisCO allow the description of its kinetic parameters through simple power laws. Thus it would not be adequate to treat the maximal carboxylation rate ( $k_{ccat}$ ), the Michaelis-Menten constants for  $CO_2$  ( $K_C$ ) and  $O_2$  ( $K_O$ ), and the specificity ( $S_{C/O}$ ) as independent evolutionary parameters in the model. Data from Savir et al. (2010) excluding form II RuBisCOs and the extreme *Synechococcus* 6301 form were used to deduce power laws that are more suitable for land plants (Figure S1):

$$K_C = 16.07 k_{ccat}^{2.36}$$

$$\frac{K_C}{K_O} = 3.7 \cdot 10^{-4} k_{ccat}^{1.16}$$

$$\gamma_* = \frac{0.5}{S_{C/O}} = \frac{0.5}{5009.76 k_{ccat}^{-0.6}}$$

Inserting the resulting power laws into the model described above reduces the number of evolutionary parameters to six, namely  $\beta$ ,  $V_{pmax}$ ,  $K_p$ ,  $g_s$ ,  $\xi$ , and  $k_{ccat}$ . The resulting model thus spans a six-dimensional fitness landscape.

### Population Genetics Model

The selection coefficient ( $s$ ) is calculated using the net  $CO_2$  assimilation rate of the ancestral state ( $A_{C1}$ ) and the net  $CO_2$  assimilation rate of the derived state ( $A_{C2}$ ). We assume that fitness is proportional to net  $CO_2$  assimilation rate:

$$S = \frac{A_{C2} - A_{C1}}{A_{C2}}$$

The probability of fixation ( $\pi$ ) of the derived state in a population of randomly mating diploid hermaphrodites, where mutations are incompletely dominant (i.e., heterozygous effect  $h = 1/2$ ), is given by (Kimura, 1957):

$$\pi = \begin{cases} \frac{1}{2N^s} & s = 0 \\ \frac{1 - e^{-s}}{1 - e^{-2Ns}} & s \neq 0 \end{cases}$$

### Comparison of Model Predictions to Data from Experimental Inhibition of PEPC

Brown et al. (1991) evaluated the effect of the PEPC inhibitor DCDP (3,3-dichloro-2-dihydroxyphosphinoylmethyl-2-propenoate) on steady state net photosynthesis in  $C_3$ ,  $C_4$  and  $C_3$ - $C_4$  species from the genera *Flaveria*, *Panicum* and *Moricandia*. DCDP is expected to inhibit PEPC activity by 80% to 100% (Jenkins et al., 1989). In order to validate our model of steady state photosynthesis, we parameterized it for the species used in Brown et al. (1991) and evaluated the effect on  $A_c$  when reducing PEPC activity ( $V_{pmax}$ ) by 80% and by 100%. Where experimental parameters were unavailable (see section “Comparison to experimental data”), we used  $C_3$  parameters for  $C_3$  and  $C_3$ - $C_4$  intermediates, and  $C_4$  parameters for  $C_4$  species. Where  $\beta$  was not available, the value that maximizes  $A_c$  (given the remaining parameters) was used.

### Coupling the Mechanistic Model with a Genome-Scale Metabolic Reconstruction

In order to show that the choice of biochemical model operates at the right resolution, we coupled the mechanistic model presented above with a genome scale metabolic reconstruction of  $C_4$  metabolism, C4GEM (Dal’Molin et al., 2010). C4GEM accounts for 1,755 metabolites and 1,588 unique reactions and contains a complex biomass reaction including carbohydrates, cell wall components, amino acids, and nucleotides (Dal’Molin et al., 2010). Flux Balance Analysis (FBA) was conducted using the C4GEM model:

maximize  $cv$

subject to  $Sv = 0$

$V_{min} \leq v \leq V_{max}$

where  $c$  is the vector of coefficients in the objective function, here the leaf biomass production.  $v$  is the vector of fluxes through the networks reaction,  $S$  is the stoichiometric matrix of the metabolic network, and  $v_{min}$  and  $v_{max}$  represent constraints on the respective fluxes. In addition to the constraints used in C4GEM, the following reactions were constrained using the values predicted by the mechanistic model: net  $CO_2$  uptake, RuBisCO carboxylation and oxygenation in mesophyll and bundle sheath,  $CO_2$  leakage from the bundle sheath, PEPC activity in the mesophyll, activity of the respective decarboxylating enzyme in the bundle sheath, plasmodesmata flux of glycine and serine and decarboxylation by the GDC complex.

We sampled the parameter space given by the mechanistic model 1,000 times, each time calculating the solution for  $A_c$ , constrained the FBA model using the predicted values and optimized biomass production under these constraints (Figure S2). This procedure was repeated for NADPME, NADME and PEPC subtype constraints.

### Analysis of the Fitness Landscape

In order to analyze the model, the six evolutionary parameters were constrained to ranges given by representative  $C_3$  and  $C_4$  values (Table S2). For  $\beta$ ,  $k_{ccat}$ , and  $\xi$ , parameter ranges were chosen based on the data set from the genera *Flaveria*, *Moricandia* and *Panicum* presented below. Comparison of measurements for  $V_{pmax}$  with data on other proxies for  $C_4$  cycle activity in *Flaveria* (such as  $\delta^{13}C$  [Apel et al., 1988; Monson et al., 1988; Sudderth et al., 2007],  $CO_2$  compensation point [Vogan and Sage, 2011], %  $^{14}C$  in  $C_4$  acids after 8-10 s pulse [Vogan and Sage, 2011]) showed saturation above PEPC activities of about  $130 \mu mol m^{-2} s^{-1}$ , and the parameter range for  $V_{pmax}$  was thus chosen from zero to  $130 \mu mol m^{-2} s^{-1}$ .

Data on bundle sheath conductivity are very sparse. We used  $3 mmol m^{-2} s^{-1}$  for the  $C_4$  value (as suggested by von Caemmerer [2000]) and a 15-fold higher value for the  $C_3$  state, although this parameter was to our knowledge never measured for  $C_3$  plants.

(Bauwe, 1986) used kinetic progress curves to estimate  $K_p$  in different species, and these results were used to estimate ranges for this parameter.

All parameter ranges were divided into five equidistant steps.

### Analysis of Evolutionary Trajectories

The ultimate cause of evolutionary phenotypic changes are genomic mutations. As we currently lack a precise genotype-phenotype map for this system, we used qualitative reasoning when choosing relative mutational probabilities. This yielded the following hierarchy of mutational probabilities  $\mu$ :

$$\mu(\xi) > \mu(k_{ccat}) > \mu(K_p) = \mu(g_s) = \mu(\beta) > \mu(V_{pmax})$$

As discussed above, loss of the chlorenchymatous isoforms of GLDP are sufficient to divert glycine decarboxylation to the bundle sheath specific forms, a comparatively minor molecular change (Sage, 2004). We thus placed the highest mutational probability on the activity of the photorespiratory pump,  $\xi$  (see discussion in the main text).

It was shown that a single mutation in the *rbcL* gene can act as a switch between C<sub>3</sub>-like and C<sub>4</sub>-like catalytic properties in *Flaveria* RuBisCO (Whitney et al., 2011). Although the underlying mechanism to gain C<sub>4</sub>-like kinetics seems to differ between species (Whitney et al., 2011), this result suggests a rather high mutational probability for  $k_{ccat}$ . A large mutational target for changes in  $\xi$  is further supported by the fact that active GDC is a multi-enzyme system consisting of four distinct subunits, and downregulation of any of these will result in reduced GDC activity (Engel et al., 2007). Furthermore, M expression of each subunit is likely regulated by several transcription factor binding sites, each with several nucleotides important for binding. Random mutations at any of these sites are likely to downregulate M GDC expression. This inactivation is sufficient to establish a photorespiratory CO<sub>2</sub> pump, as we assume a low diffusional distance between M and BS cells, and a specific subcellular distribution of organelles in the BS (proto-Kranz anatomy).

We assigned the lowest mutational probability to  $V_{pmax}$ . Implementation of the C<sub>4</sub> cycle can vary between species (Furbank, 2011), and incomplete C<sub>4</sub> cycles can be operational (Monson and Moore, 1989). This increases the size of the C<sub>4</sub> cycle as a mutational target. Nevertheless, increased and localized expression of the respective rate-limiting gene is required. In the case of *Flaveria*, two *cis*-regulatory elements responsible for C<sub>4</sub>-like expression of the *ppcA* gene coding for PEPC were identified (Crona et al., 2013). This suggests a higher complexity of changes needed when compared to loss of expression of an isoform or change in kinetic properties of an enzyme.

Although there is some insight into the coordinated expression of RuBisCO subunits (Rodermeil, 2001), the molecular mechanisms for changes in  $\beta$ ,  $g_s$ , and  $K_p$  are largely unknown. We set the corresponding mutational probabilities to equal values intermediate between those of  $k_{ccat}$  and  $V_{pmax}$ .

To rule out that wrong assumptions about the probabilities of changes and number of equidistant steps affect our results, we ran a sensitivity analysis against these factors. The simulation of 1,000 evolutionary trajectories was repeated 30,000 times with randomly chosen sets of parameters for probabilities of changes and number of equidistant steps. Mutational probabilities were each drawn uniformly between 0 and 1, and then normalized to sum up to one. Numbers of steps for each parameter were drawn uniformly between 1 and 10.

For each parameter in the biochemical model, the normalized mean of the step numbers at which fixation occurred was used to characterize each simulation run (Figure S3). The qualitative patterns of our specific parameter set are reproduced for almost all biochemical parameters. The only exception is  $\xi$ , which is a very late change in most scenarios, indicating that the photorespiratory pump needs a high probability of change in order to play a role in the evolutionary process. As discussed above, the underlying mechanism for increasing  $\xi$  justifies this high probability in our assumptions.

### Comparison to Experimental Data

The dicotyledonous genera *Flaveria* (Asteraceae) and *Moricandia* (Brassicaceae), as well as the monocotyledonous *Panicum* (Poaceae), each contain C<sub>3</sub>-C<sub>4</sub> intermediate species. In order to validate the evolutionary model we obtained data on species from these genera from the literature and complemented it with further measurements.

PEPC activity in leaf extracts was used as a proxy for C<sub>4</sub> cycle activity ( $V_{pmax}$ ). *F. robusta*, *F. chloraefolia*, *F. pringlei*, *F. angustifolia*, *F. cronquistii*, *F. anomala*, *F. floridana*, *F. ramosissima*, *F. linearis*, *F. brownii*, *F. vaginata*, *F. trinervia*, *F. bidentis*, and *F. australasica* were grown in 17 cm pots on soil (C-400 with Cocopor [Stender Erden, Schermbeck, Germany] fertilized with 3 g/l Osmocote exact standard 3 – 4 M [Scotts, Nordhorn, Germany]) in May 2012 in the greenhouse. Additional light was given 16h per day. The first and second youngest fully expanded leaves were harvested from about 2 month old plants of comparable sizes. Four biological replicates were used per species, each containing material of three individuals. PEPC activity was determined as summarized by Ashton et al. (1990).

Additional PEPC activities for one *Moricandia* and three *Panicum* species were obtained from Winter et al. (1982) and Ku et al. (1976). The values from Ku et al. (1976) were converted to leaf area basis using data from Ku and Edwards (1978).

Data on RuBisCO distribution ( $\beta$ ) for six *Flaveria* species and three *Panicum* species were obtained from cell separation experiments (Edwards and Gutierrez, 1972; Holaday et al., 1988; Ku et al., 1976; Moore et al., 1988, 1989), and in the case of the data from Ku et al. (1976) and Holaday et al. (1988), corrected for mesophyll to bundle sheath area ratio (Hattersley, 1984; McKown and Dengler, 2007; Wilson et al., 1983). For four *Flaveria* species,  $\beta$  was estimated from immunofluorescence studies (Bauwe, 1984). Immunofluorescence data were evaluated visually and corrected for mesophyll to bundle sheath cell ratio (McKown and Dengler, 2007).

RuBisCO turnover rate ( $k_{ccat}$ ) for 11 *Flaveria* species was taken from Wessinger et al. (1989).

The fraction of mesophyll derived photorespirational glycine decarboxylated in the bundle sheath ( $\xi$ ) in *Flaveria* was estimated from transcriptome data. The transcriptomes of photosynthetically active leaves from 14 *Flaveria* species (see above) were analyzed by RNA-seq via Illumina sequencing. The resulting reads (from one to four RNaseq experiments per species with 30 to 51 million reads per experiment) were mapped to the sequences of the *F. trinervia glpA* and *glpD* genes (GenBank accession: Z99767.1 and Z99768.1) with the software package CLC Genomic Workbench using standard settings and allowing nonambiguous mapping only. The P subunit of the glycine decarboxylase is an essential component of glycine decarboxylation. While the *glpA* gene is

known to be transcribed exclusively in the bundle sheath in *Flaveria pringlei* (C<sub>3</sub>) and *F. trinervia* (C<sub>4</sub>), *gldpD* is transcribed throughout all inner leaf tissues in *F. pringlei*.  $\xi$  was calculated according to:

$$c = \frac{A+D}{2} - D$$

$$\xi = \frac{c}{c+D}$$

where  $A$  is the sum of reads mapped to *gldpA* and  $D$  is the sum of reads mapped to *gldpD*.

Estimates for  $\xi$  in one *Moricandia* and one *Panicum* species were obtained from immunogold labeling experiments (Hylton et al., 1988), corrected for mesophyll to bundle sheath distribution of mitochondria (Brown and Hattersley, 1989).

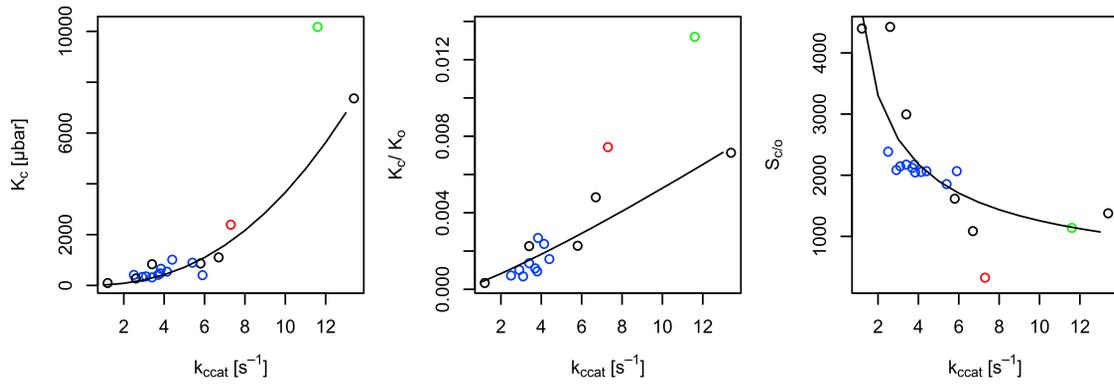
Bundle sheath conductance was estimated in some C<sub>4</sub> species using inhibitors of PEPC (Brown, 1997; Jenkins et al., 1989). These methods rely on the assumption that RuBisCO activity is confined to the bundle sheath, and  $g_s$  has to our knowledge never been measured for C<sub>3</sub>-C<sub>4</sub> intermediates or C<sub>3</sub> species, where this assumption does not hold.

We used data from Bauwe (1986) to define the parameter range for  $K_p$ , but further data were not available.

The data set was compared to the predicted set of trajectories. Data points were mapped to the closest point in the discrete 6-dimensional space given by the model. This allowed counting the number of species that are crossed by each predicted path. Results were compared to the random null model described above.

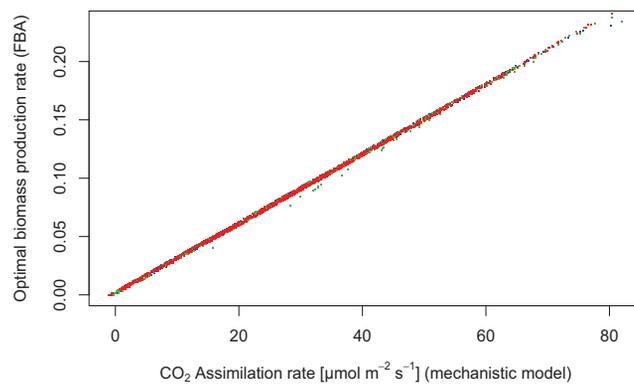
#### SUPPLEMENTAL REFERENCES

- Apel, P., Bauwe, H., Bassüner, B., and Maass, I. (1988). Photosynthetic properties of *Flaveria cronquistii*, *F. palmeri*, and hybrids between them. *Biochem. Physiol. Pflanz.* 183, 291–299.
- Bauwe, H. (1984). Photosynthetic enzyme activities and immunofluorescence studies on the localization of ribulose-1, 5-bisphosphate carboxylase/oxygenase in leaves of C<sub>3</sub>, C<sub>4</sub>, and C<sub>3</sub>-C<sub>4</sub> intermediate species of *Flaveria* (Asteraceae). *Biochem. Physiol. Pflanz.* 179, 253–268.
- Bauwe, H. (1986). An efficient method for the determination of  $K_m$  values for HCO<sub>3</sub><sup>-</sup> of phosphoenolpyruvate carboxylase. *Planta* 169, 356–360.
- Brown, R.H. (1997). Analysis of bundle sheath conductance and C<sub>4</sub> photosynthesis using a PEP-carboxylase inhibitor. *Aust. J. Plant Physiol.* 24, 549–554.
- Brown, R.H., and Hattersley, P.W. (1989). Leaf anatomy of C<sub>3</sub>-C<sub>4</sub> species as related to evolution of C<sub>4</sub> photosynthesis. *Plant Physiol.* 91, 1543–1550.
- Crona, K., Greene, D., and Barlow, M. (2013). The peaks and geometry of fitness landscapes. *J. Theor. Biol.* 317, 1–10.
- Edwards, G.E., and Gutierrez, M. (1972). Metabolic activities in extracts of mesophyll and bundle sheath cells of *Panicum miliaceum* (L.) in relation to the C<sub>4</sub> dicarboxylic acid pathway of photosynthesis. *Plant Physiol.* 50, 728–732.
- Hattersley, P.W. (1984). Characterization of C<sub>4</sub> type leaf anatomy in grasses (Poaceae). Mesophyll: bundle sheath area ratios. *Ann. Bot. (Lond.)* 53, 163–180.
- Holaday, A.S., Brown, R.H., Bartlett, J.M., Sandlin, E.A., and Jackson, R.C. (1988). Enzymic and photosynthetic characteristics of reciprocal F<sub>1</sub> hybrids of *Flaveria pringlei* (C<sub>3</sub>) and *Flaveria brownii* (C<sub>4</sub>-like species). *Plant Physiol.* 87, 484–490.
- Jenkins, C.L.D., Furbank, R.T., and Hatch, M.D. (1989). Inorganic carbon diffusion between C<sub>4</sub> mesophyll and bundle sheath cells: direct bundle sheath CO<sub>2</sub> assimilation in intact leaves in the presence of an inhibitor of the C<sub>4</sub> pathway. *Plant Physiol.* 91, 1356–1363.
- Ku, S.B., and Edwards, G.E. (1978). Photosynthetic efficiency of *Panicum hians* and *Panicum milioides* in relation to C<sub>3</sub> and C<sub>4</sub> plants. *Plant Cell Physiol.* 19, 665–675.
- Ku, S.B., Edwards, G.E., and Kanai, R. (1976). Distribution of enzymes related to C<sub>3</sub> and C<sub>4</sub> pathway of photosynthesis between mesophyll and bundle sheath cells of *Panicum hians* and *Panicum milioides*. *Plant Cell Physiol.* 17, 615–620.
- Monson, R.K., and Moore, B.D. (1989). On the significance of C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis to the evolution of C<sub>4</sub> photosynthesis. *Plant Cell Environ.* 12, 689–699.
- Monson, R.K., Teeri, J.A., Ku, M.S.B., Gurevitch, J., Mets, L.J., and Dudley, S. (1988). Carbon-isotope discrimination by leaves of *Flaveria* species exhibiting different amounts of C<sub>3</sub>- and C<sub>4</sub>-cycle co-function. *Planta* 174, 145–151.
- Moore, B.D., Monson, R.K., Ku, M.S.B., and Edwards, G.E. (1988). Activities of principal photosynthetic and photorespiratory enzymes in leaf mesophyll and bundle sheath protoplasts from the C<sub>3</sub>-C<sub>4</sub> intermediate *Flaveria ramosissima*. *Plant Cell Physiol.* 29, 999–1006.
- Moore, B.D., Ku, M.S.B., and Edwards, G.E. (1989). Expression of C<sub>4</sub>-like photosynthesis in several species of *Flaveria*. *Plant Cell Environ.* 12, 541–549.
- Rodermel, S. (2001). Pathways of plastid-to-nucleus signaling. *Trends Plant Sci.* 6, 471–478.
- Sudderth, E.A., Muhaidat, R.M., McKown, A.D., Kocacinar, F., and Sage, R.F. (2007). Leaf anatomy, gas exchange and photosynthetic enzyme activity in *Flaveria kochiana*. *Funct. Plant Biol.* 34, 118–129.
- Vogan, P.J., and Sage, R.F. (2011). Water-use efficiency and nitrogen-use efficiency of C<sub>3</sub>-C<sub>4</sub> intermediate species of *Flaveria* Juss. (Asteraceae). *Plant Cell Environ.* 34, 1415–1430.
- Wessinger, M.E., Edwards, G.E., and Ku, M.S.B. (1989). Quantity and kinetic properties of ribulose 1, 5-bisphosphate carboxylase in C<sub>3</sub>, C<sub>4</sub>, and C<sub>3</sub>-C<sub>4</sub> intermediate species of *Flaveria* (Asteraceae). *Plant Cell Physiol.* 30, 665–671.
- Whitney, S.M., Sharwood, R.E., Orr, D., White, S.J., Alonso, H., and Galmés, J. (2011). Isoleucine 309 acts as a C<sub>4</sub> catalytic switch that increases ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in *Flaveria*. *Proc. Natl. Acad. Sci. USA* 30, 14688–14693.
- Wilson, J.R., Brown, R.H., and Windham, W.R. (1983). Influence of Leaf Anatomy on the Dry Matter Digestibility of C<sub>3</sub>, C<sub>4</sub>, and C<sub>3</sub>/C<sub>4</sub> Intermediate Types of *Panicum* Species. *Crop Sci.* 23, 141–146.
- Winter, K., Usuda, H., Tsuzuki, M., Schmitt, M., Edwards, G.E., Thomas, R.J., and Evert, R.F. (1982). Influence of Nitrate and Ammonia on Photosynthetic Characteristics and Leaf Anatomy of *Moricandia arvensis*. *Plant Physiol.* 70, 616–625.



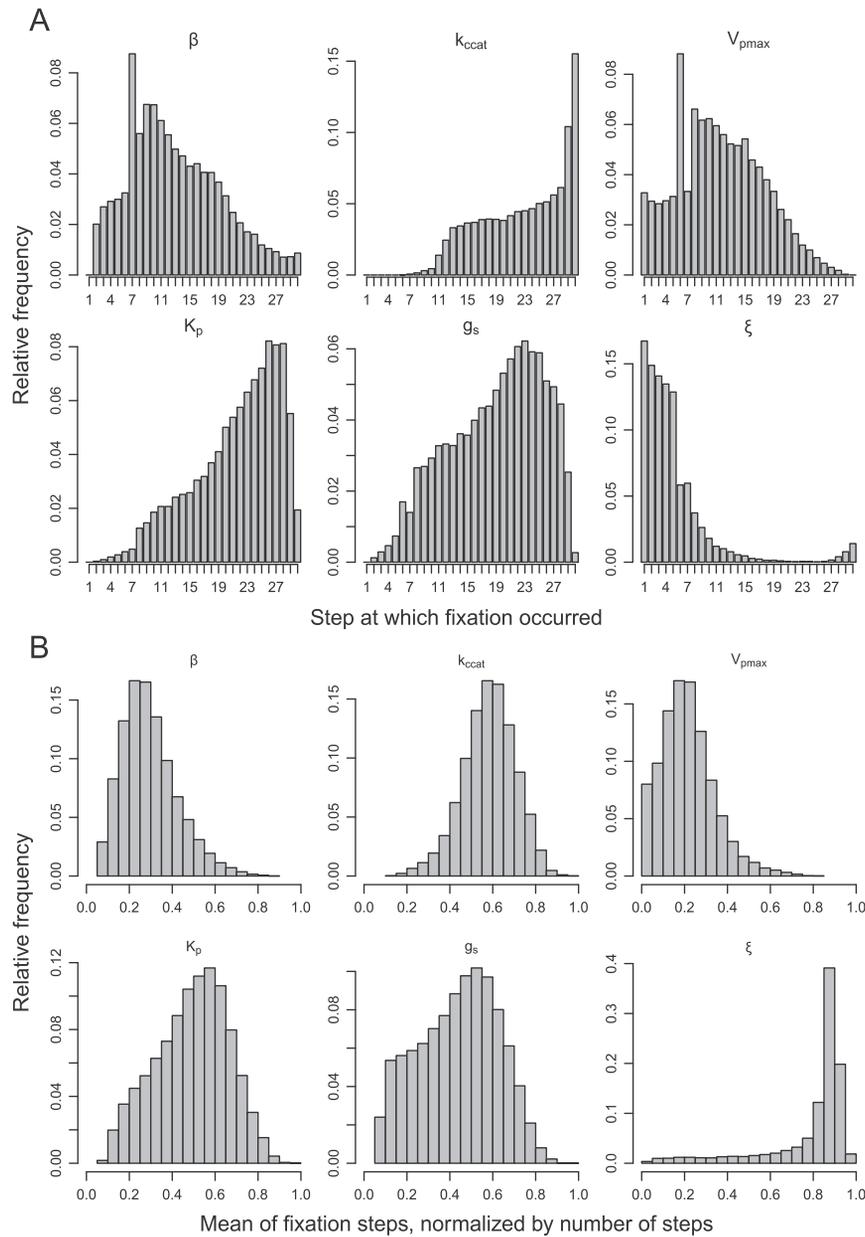
**Figure S1. Nonindependence of RuBisCO Kinetic Constants, Related to Figure 1 and Extended Experimental Procedures**

The figure shows two-dimensional fits to RuBisCO kinetic constants obtained from Savir et al. (2010). Least-squares fitting of power laws was conducted using the `optim()` function of the R environment. The resulting power laws reflect trade-offs, and were used to predict the other RuBisCO kinetic parameters from  $k_{ccat}$ . Blue, Land plants; red, Form II RuBisCO from *Rhodospirillum rubum*, not used for fitting; green, *Synechococcus 6301*, not used for fitting.



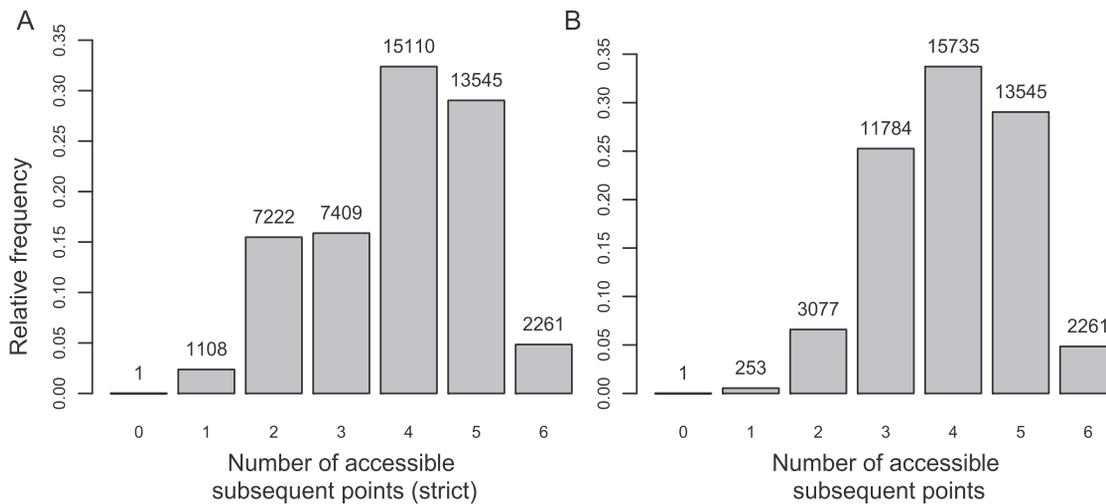
**Figure S2. The Biomass Production Rate Predicted from Genome-wide Flux-Balance Analysis Is Directly Proportional to the Rate of Carbon Fixation,  $A_c$ , Related to Figure 1**

C<sub>4</sub> subtypes are shown in different colors: green, NAD malic enzyme (NAD-ME); blue, NADP malic enzyme (NADP-ME); and red, phosphoenolpyruvate carboxykinase. The slopes obtained from linear regressions for the three C<sub>4</sub> subtypes were statistically indistinguishable ( $p = 0.38$ , ANCOVA), demonstrating the robustness of the model.



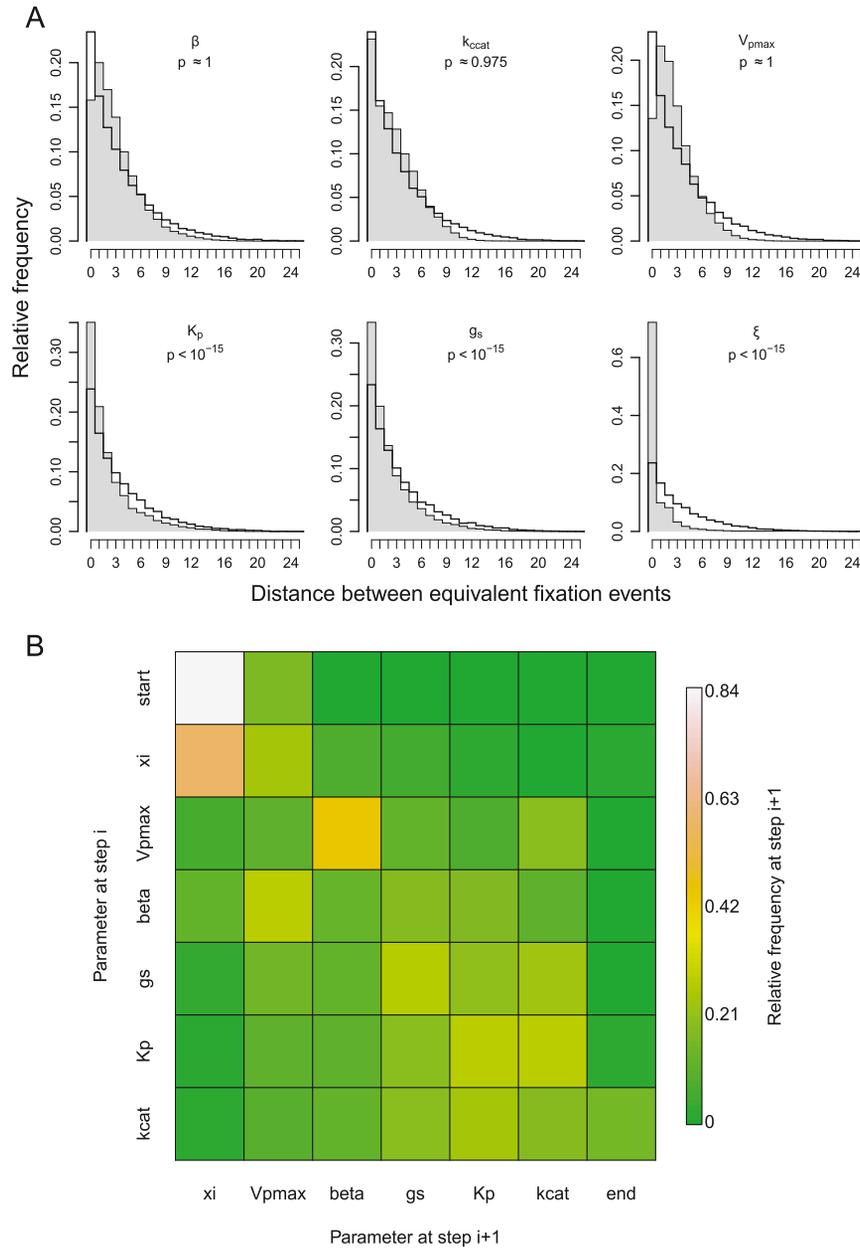
**Figure S3. The Distribution of Fixation Times for Each Model Parameter, Related to Figure 4**

(A and B) In most simulations, establishment of the photorespiratory pump ( $\xi$ ) is the first change to occur. The  $C_4$  cycle ( $V_{pmax}$ ) and shift of RuBisCO activity to the bundle sheath ( $\beta$ ) are also fixed in early stages. In our simulations, reduction of the conductance ( $g_s$ ) is adaptive as soon as one of the pumps is established, but mainly occurs in later stages when the  $C_4$  cycle is fully operating.  $K_p$  also changes late. Except for the last two changes,  $k_{cat}$  shows the most uniform distribution along evolutionary trajectories. The same general pattern is seen with the discretizations and relative mutational probabilities assumed in our simulations (A) and in a sensitivity analysis that combines results from 1,000 simulations each of 30,000 randomly chosen parameter combinations (B). The only exception is the early establishment of the photorespiratory pump ( $\xi$ ), which only happens in our simulations because the respective mutational probability is high.



**Figure S4. Evolutionary Accessibility of Subsequent Points in the Fitness Landscape, Related to Figure 3**

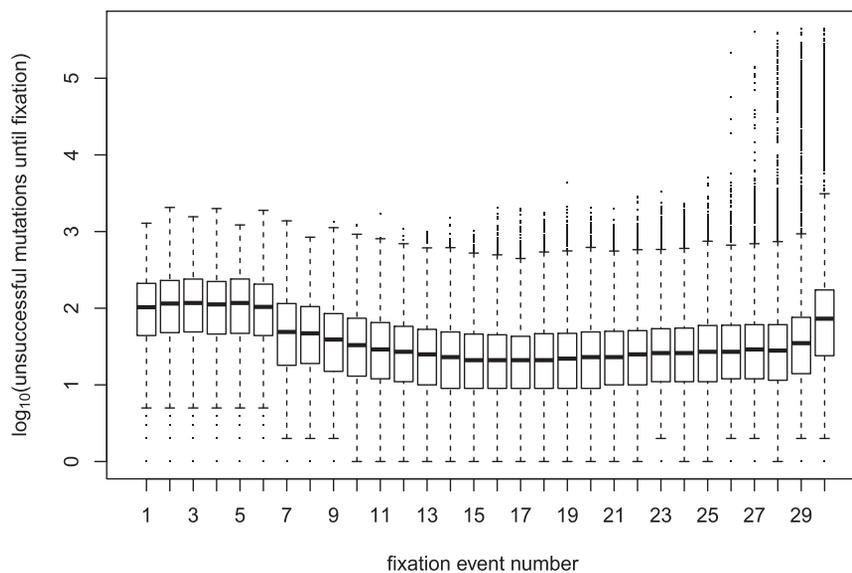
(A and B) Subsequent points are defined as accessible if they come with a fitness change that is strictly positive (A) or at least zero (B); *i.e.*, for a point with  $n$  accessible subsequent points,  $n$  different parameters can be increased alternatively while increasing (A) or not decreasing (B) fitness. The only location lacking accessible subsequent points is the global maximum, the  $C_4$  state.



**Figure S5. Relationships between Changes in Individual Parameters in the Stochastic Simulations, Related to Figure 4**

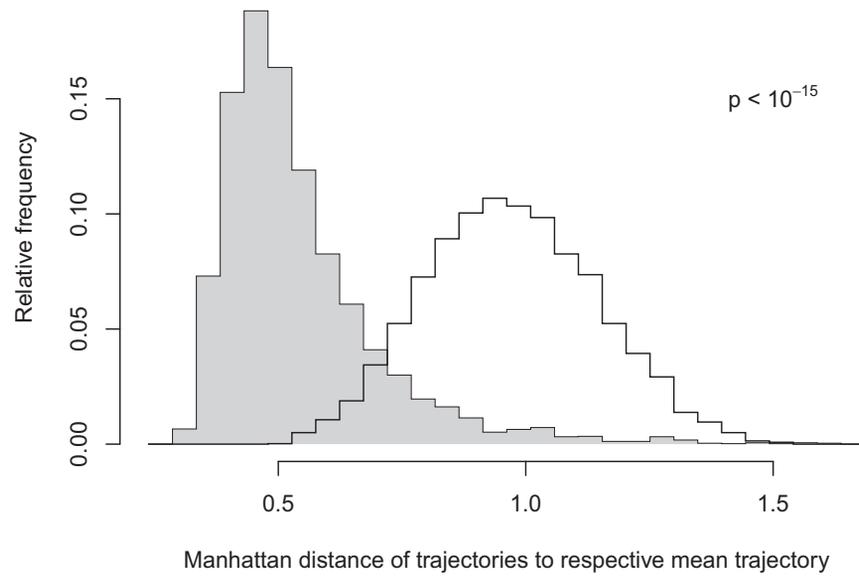
(A) Distributions of distances between two changes in the same parameter. A distance of zero indicates two immediately successive steps. For the parameters in the bottom row ( $K_p$ ,  $g_s$ ,  $\xi$ ), immediately successive steps (distance=0) are much more common than expected by chance ( $p < 10^{-15}$  in each case, Fisher's exact test); the same is true for  $\beta$  and  $V_{pmax}$  when treated as a combined parameter set. The only trait that does not evolve in a modular fashion is thus  $K_{cat}$ , which is significantly more dispersed than expected by chance ( $p < 10^{-15}$ , median test).

(B) Transition matrix for evolutionary trajectories in stochastic simulations. Colors indicate the relative frequency with which a change in parameter  $Y$  at step  $i$  is followed by a change in parameter  $X$  at step  $i + 1$ .



**Figure S6. No Systematic Slowdown of Evolution, Related to Figure 5**

Boxplot for the number of parameter changes that were attempted before a change was fixed according to the population genetic model, based on 5,000 simulated evolutionary trajectories from  $C_3$  to  $C_4$ . The first six steps – mostly shifts in photorespiration to the bundle sheath ( $\xi$ ) and the first establishment of  $C_4$  cycle activity ( $V_{pmax}$ ) – take substantially longer than later steps. Except for the very last steps, there is no clear trend of decelerating evolution, contrasting previous observations in experimental studies and theoretical expectations (see [Discussion](#) in the main text).



**Figure S7. Simulated Paths Cluster, Related to Figure 5**

Histograms of pointwise distances of simulated trajectories to the mean path. Gray: Evolutionary model. White: Random model. The two distributions are significantly different ( $p < 10^{-15}$ , Wilcoxon rank sum test).

Table S1. Parameter Dimensions in the Biochemical Model, Related to Extended Experimental Procedures

Parameters	Dimension
$V_{mmax}, V_{smax}, V_{pmax}$	$\mu\text{mol m}^{-2} \text{s}^{-1}$
$C_m, O_m, C_s, O_s$	$\mu\text{bar}$
$\beta$	-
$E_{tot}$	$\mu\text{mol m}^{-2}$
$K_c, K_o, K_p$	$\mu\text{bar}$
$k_{ccat}$	$\text{s}^{-1}$
$R_m, R_s$	$\mu\text{mol m}^{-2} \text{s}^{-1}$
$\xi$	-
$S_{c/o}$	-
$g_s, g_o$	$\mu\text{mol m}^{-2} \text{s}^{-1}$

Parameter descriptions:  $V_{mmax}, V_{smax}$ : maximal RuBisCO activity per leaf area in the mesophyll and bundle sheath, respectively;  $V_{pmax}$ : Activity of the C<sub>4</sub> cycle;  $C_m, C_s$ : CO<sub>2</sub> partial pressure in the mesophyll and bundle sheath chloroplasts, respectively;  $O_m, O_s$ : O<sub>2</sub> partial pressure in the mesophyll and bundle sheath chloroplasts, respectively;  $\beta$ : fraction of RuBisCO active sites in the mesophyll;  $E_{tot}$ : total leaf RuBisCO concentration;  $K_c, K_o$ : Michaelis-Menten constants of RuBisCO for CO<sub>2</sub> and O<sub>2</sub>, respectively;  $K_p$ : Michaelis-Menten constant of PEPC for bicarbonate;  $k_{ccat}$ : maximal rate of carboxylation for RuBisCO;  $R_m, R_s$ : mitochondrial respiration other than photorespiration in the mesophyll and the bundle sheath, respectively;  $\xi$ : activity of the photorespiratory pump;  $S_{c/o}$ : RuBisCO specificity for CO<sub>2</sub>;  $g_s, g_o$ : bundle sheath conductance for CO<sub>2</sub> and O<sub>2</sub>, respectively.

Table S2. Ranges and Discretization of Evolving Parameters, Related to Figure 1

Parameter	C <sub>3</sub> value	C <sub>4</sub> value	Dimension	Number of steps	Mutational probability
$V_{pmax}$	0	130	$\mu\text{mol m}^{-2} \text{s}^{-1}$	5	1/75
$\beta$	0.95	$2.0 \cdot 10^{-3}$	-	5	2/75
$K_p$	200	80	$\mu\text{bar}$	5	2/75
$k_{ccat}$	3.4	8.8	$\text{s}^{-1}$	5	4/75
$\xi$	0	0.98	-	5	64/75
$g_s$	$1.5 \cdot 10^{-2}$	$1.0 \cdot 10^{-3}$	$\mu\text{mol m}^{-2} \text{s}^{-1}$	5	2/75

Sources for parameter values are given in the text. Parameter descriptions:  $V_{pmax}$ : activity of the C<sub>4</sub> cycle;  $\beta$ : fraction of RuBisCO active sites in the mesophyll;  $K_p$ : Michaelis-Menten constant of PEPC for bicarbonate;  $k_{ccat}$ : maximal rate of carboxylation for RuBisCO;  $\xi$ : activity of the photorespiratory pump;  $g_s$ : bundle sheath conductance for CO<sub>2</sub>.