

# Phenotypic landscape inference reveals multiple evolutionary paths to C<sub>4</sub> photosynthesis

Ben P. Williams<sup>1,\*</sup>, Iain G. Johnston<sup>2,\*</sup>, Sarah Covshoff<sup>1</sup>, Julian M. Hibberd<sup>1</sup>

<sup>1</sup> Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom

<sup>2</sup> Department of Mathematics, Imperial College London, London, United Kingdom

\* These authors contributed equally to this work.

## Abstract

C<sub>4</sub> photosynthesis has independently evolved from the ancestral C<sub>3</sub> pathway in at least 60 plant lineages, but, as with other complex traits, how it evolved is unclear. Here we show that the polyphyletic appearance of C<sub>4</sub> photosynthesis is associated with diverse and flexible evolutionary paths that group into four major trajectories. We conducted a meta-analysis of 18 lineages containing species that use C<sub>3</sub>, C<sub>4</sub>, or intermediate C<sub>3</sub>-C<sub>4</sub> forms of photosynthesis to parameterise a 16-dimensional phenotypic landscape. We then developed and experimentally verified a novel Bayesian approach based on a hidden Markov model that predicts how the C<sub>4</sub> phenotype evolved. The alternative evolutionary histories underlying the appearance of C<sub>4</sub> photosynthesis were determined by ancestral lineage and initial phenotypic alterations unrelated to photosynthesis. We conclude that the order of C<sub>4</sub> trait acquisition is flexible and driven by non-photosynthetic drivers. This flexibility will have facilitated the convergent evolution of this complex trait.

## Introduction

The convergent evolution of complex traits is surprisingly common, with examples including camera-like eyes of cephalopods, vertebrates and cnidarian [1], mimicry in invertebrates and vertebrates [2, 3] and the different photosynthetic machineries of plants [4]. While the polyphyletic origin of simple traits [5, 6] is underpinned by flexibility in the underlying molecular mechanisms, the extent to which this applies to complex traits is less clear. C<sub>4</sub> photosynthesis is both highly complex, involving alterations to leaf anatomy, cellular ultrastructure and photosynthetic metabolism, and also convergent, being found in at least 60 independent lineages of angiosperms [4]. As the emergence of the entire C<sub>4</sub> phenotype cannot be comprehensively explored experimentally, C<sub>4</sub> photosynthesis is an ideal system for the mathematical modelling of complex trait evolution as transitions on an underlying phenotype landscape. Furthermore, understanding the evolutionary events that have generated C<sub>4</sub> photosynthesis on many independent occasions has the

potential to inform approaches being undertaken to engineer C<sub>4</sub> photosynthesis to C<sub>3</sub> crop species [7].

The C<sub>4</sub> pathway is estimated to have first evolved between 32 and 25 million years ago [8] in response to multiple ecological drivers including decreasing atmospheric CO<sub>2</sub> concentration [9]. C<sub>4</sub> species have since radiated to represent the most productive crops and native vegetation on the planet, because modifications to their leaves increase the efficiency of photosynthesis in the subtropics and tropics [10]. In C<sub>4</sub> plants, photosynthetic efficiency is improved compared with C<sub>3</sub> species because significant alterations to leaf anatomy, cell biology and biochemistry lead to higher concentrations of CO<sub>2</sub> around the primary carboxylase RuBisCO [11]. The morphology of C<sub>4</sub> leaves is typically modified into so-called Kranz anatomy that consists of repeating units of vein, bundle sheath (BS) and mesophyll (M) cells [11] (Appendix Fig. 5). Photosynthetic metabolism becomes modified and compartmentalised between the M and BS, with M cells lacking RuBisCO but instead containing high activities of the alternate carboxylase PEPC to generate C<sub>4</sub> acids. The diffusion of these acids followed by their decarboxylation in BS cells around RuBisCO increases CO<sub>2</sub> supply and therefore photosynthetic efficiency [12]. C<sub>4</sub> acids are decarboxylated by at least one of three enzymes within BS cells; NADP or NAD-dependent malic enzymes (NADP-ME or NAD-ME respectively), or phosphoenolpyruvate carboxykinase (PCK) [13]. Specific lineages of C<sub>4</sub> species have typically been classified into one of three sub-types, based on the activity of these decarboxylases, as well as anatomical and cellular traits that consistently correlate with each other [14].

The genetic mechanisms underlying the evolution of cell-specific gene expression associated with the separation of photosynthetic metabolism between M and BS cells involve both alterations to cis-elements and trans-acting factors [15, 16, 17, 18]. Phylogenetically independent lineages of C<sub>4</sub> plant have co-opted homologous mechanisms to generate cell specificity [16] as well as the altered allosteric regulation of C<sub>4</sub> enzymes [19] indicating that parallel evolution underpins at least part of the convergent C<sub>4</sub> syndrome. However, while a substantial amount of work has addressed the molecular alterations

that generate the biochemical differences between  $C_3$  and  $C_4$  plants [18] much less is known about the order and flexibility with which phenotypic traits important for  $C_4$  photosynthesis are acquired [20]. Clues to this question exist in the form of  $C_3$ - $C_4$  intermediates, species exhibiting characteristics of both  $C_3$  or  $C_4$  photosynthesis, such as the activity or localisation of  $C_4$  cycle enzymes [21], the possession of one or more anatomical or cellular adaptations associated with  $C_4$  photosynthesis [22], or combinations of both. To address these unknown aspects of  $C_4$  evolutionary history, we combined the concept of considering evolutionary paths as stochastic processes on complex adaptive landscapes [23, 24] with the analysis of extant  $C_3$ - $C_4$  intermediate species to develop a predictive model of how the full  $C_4$  phenotype evolved.

## Results

### A meta-analysis of photosynthetic phenotypes

To parameterise the phenotypic landscape underlying photosynthetic phenotypes, data was consolidated from 43 studies encompassing 18  $C_3$ , 18  $C_4$  and 37  $C_3$ - $C_4$  intermediate species from 22 genera. These  $C_3$ - $C_4$  species are from 18 independent lineages likely representing 18 distinct evolutionary origins of  $C_3$ - $C_4$  intermediacy [4] (Appendix Fig. 6). These studies were used to quantify 16 biochemical, anatomical and cellular characteristics associated with  $C_4$  photosynthesis (Fig. 1). Principal components analysis (PCA) was performed to confirm the phenotypic intermediacy of the  $C_3$ - $C_4$  species (Fig. 1A). This result, the sister-group relationships of  $C_3$ - $C_4$  species with congeneric  $C_4$  clades [4, 25, 26] and the prevalence of extant  $C_3$ - $C_4$  species in genera with the most recent origins of  $C_4$  photosynthesis [8] all support the notion that  $C_3$ - $C_4$  species represent phenotypic states through which transitions to  $C_4$  photosynthesis could occur. The combined traits of  $C_3$ - $C_4$  intermediate species therefore represent samples from across the space of phenotypes connecting  $C_3$  to  $C_4$  photosynthesis (Fig. 1B). Within our meta-analysis data,  $C_3$ - $C_4$  phenotypes were available for 33 eudicot and 4 monocot species. 16 and 17 of these species have extant congeneric relatives performing NADP-ME or NAD-ME sub-type  $C_4$  photosynthesis respectively. No  $C_3$ - $C_4$  relatives of PCK sub-type  $C_4$  species are known [4]. Our meta-analysis therefore encompassed a variety of taxonomic lineages, as well as representing close relatives of known phenotypic variants performing  $C_4$  photosynthesis.

We defined each  $C_4$  trait as either being absent (0) or present (1). For quantitative traits the Expectation-Maximization (EM) algorithm and hierarchical clustering were used to impartially assign binary scores (Appendix Fig. 7). This generated a 16-bit string for each of the species (Fig. 1), with a presence or absence score for each of the traits included in our meta-analysis.

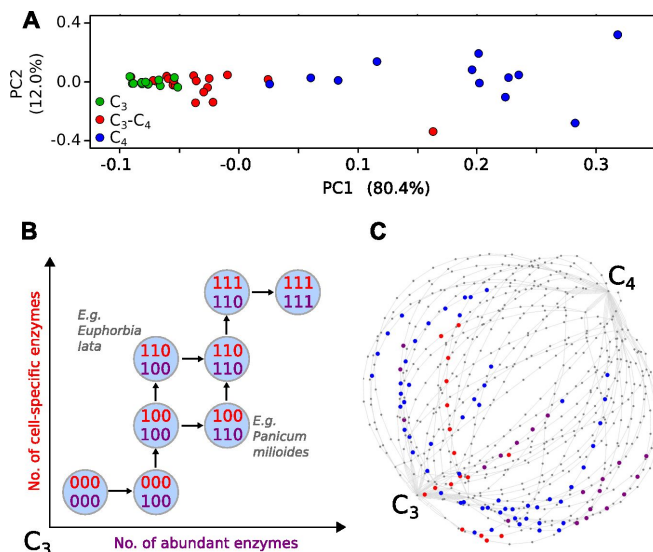


Figure 1: **Evolutionary paths to  $C_4$  phenotype space modelled from a meta-analysis of  $C_3$ - $C_4$  phenotypes.** Principal component analysis (PCA) on data for the activity of five  $C_4$  cycle enzymes confirms the intermediacy of  $C_3$ - $C_4$  species between  $C_3$  and  $C_4$  phenotype spaces (A). Each  $C_4$  trait was considered absent in  $C_3$  species and present in  $C_4$  species, with previously studied  $C_3$ - $C_4$  intermediate species representing samples from across the phenotype space (B). With a dataset of 16 phenotypic traits, a 16-dimensional space was defined. (C) shows a 2D representation of 50 pathways across this space. The phenotypes of multiple  $C_3$  - $C_4$  species were used to identify pathways compatible with individual species (e. g. *Alternanthera ficoides* (red nodes) and *Parthenium hysterophorus* (blue nodes)), and pathways compatible with the phenotypes of multiple species (purple nodes).

This defined a 16-dimensional phenotype space with 216 (65,536) nodes corresponding to all possible combinations of presence (1) and absence (0) scores for each characteristic.

## A novel Bayesian approach for predicting evolutionary trajectories

Many existing methods of inference for evolutionary trajectories rely on phylogenetic information or assumptions about the fitness landscape underlying evolutionary dynamics [27, 28, 29]. In convergent evolution, these properties are not always known, as convergent lineages may be genetically distant and associated with poor phylogenetic reconstructions. In addition, the selective pressures experienced by each may be different and dynamic. We therefore consider the convergent evolution of  $C_4$  fundamentally as the acquisition of the key phenotypic traits identified through our meta-analysis (Fig. 1B). The process of acquisition of these traits can be pictured as a path on the 16-dimensional hypercube (Fig. 1C), from the node labelled with all 0's (the  $C_3$  phenotype, with no  $C_4$  characteristics) to the node labelled with all 1's (the  $C_4$  phenotype, with all  $C_4$  characteristics).

The phenotypic landscape underlying the evolution of  $C_4$  photosynthesis was then modelled as a transition network, with weighted edges describing the probability of transitions occurring between two phenotypic states (two nodes on the hypercube, Appendix Fig. 8) Observed intermediate points were then used to constrain the structure of these phenotypic landscapes. To do this, we developed inferential machinery based on the framework of Hidden Markov Models (HMMs) [30] (Appendix Fig. 8) and simulated an ensemble of Markov chains on trial transition networks. Each of these chains represents a possible evolutionary pathway from  $C_3$  to  $C_4$ , and passes through several intermediate phenotypic states. The likelihood of observing intermediate states with characteristics compatible with the biologically observed data on  $C_3$ - $C_4$  intermediates was recorded for the set of paths supported on each trial network. A Bayesian MCMC procedure was used to sample from the set of networks most compatible with the meta-analysis dataset, and thus most likely to represent the underlying dynamics of  $C_4$  evolution. The order in which phenotypic characteristics were acquired was recorded for paths on each network compatible with the  $C_3$ - $C_4$  species data, and posterior probability distributions (given uninformative priors) for the time-ordered acquisition of each  $C_4$  trait were generated. For further information and mathematical details, see Materials & Methods.

To model the evolutionary paths generating  $C_4$  without requiring additional dimensionality, we imposed that only one  $C_4$  trait may be acquired at a time, and loss of acquired  $C_4$  traits was forbidden. To test if we were nevertheless able to detect traits acquired simultaneously in evolution, we tested our approach on artificial positive

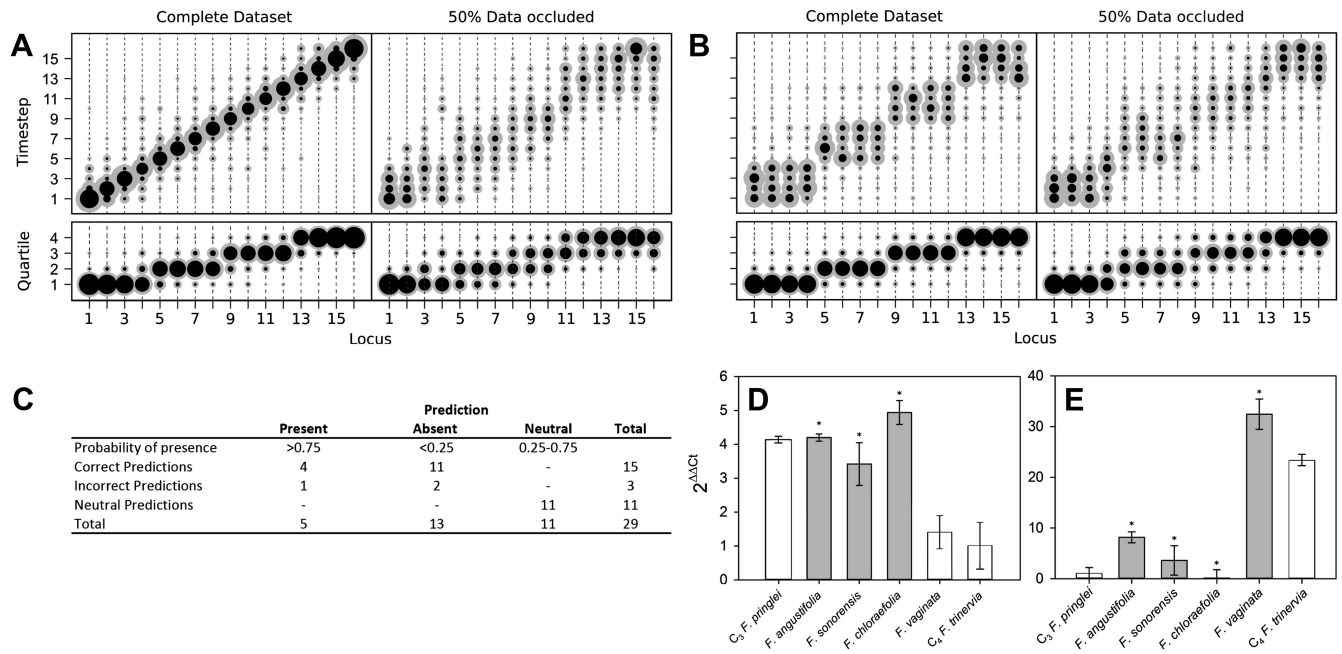
control datasets containing intermediate nodes representing a stepwise evolutionary sequence of events (Fig. 2A) and an evolutionary pathway in which four traits are acquired simultaneously at a time (Fig. 2B). Our approach clearly assigned equal acquisition probabilities to traits whose timing was linked in the underlying dataset, even when 50% of the data was occluded (Fig. 2B). These data are consistent with this approach detecting the simultaneous acquisition of traits in evolution, even though single-trait acquisitions are simulated.

## Verifying prediction accuracy

The presence and absence of unknown phenotypes were predicted by recording all phenotypes encountered along a set of simulated evolutionary trajectories that were compatible with the data from a given species (Appendix Fig. 9), and calculating the posterior distribution of the proportion of these phenotypes with the value 1 for the unknown trait. If the mean of this distribution was  $< 25\%$  or  $> 75\%$ , and that value fell outside one standard deviation of the mean, the missing trait was assigned a strong prediction of absence or presence. To comprehensively test the accuracy of our predictive machinery, we generated 29 occluded datasets, consisting of the original full dataset with one randomly chosen data point removed. The predicted phenotype of each missing trait was then compared with the known phenotype published in the original study. For 29 occluded traits 18 were strongly predicted to be present or absent, and the remaining 11 predictions were neutral. Of the 18 strongly predicted traits (i.e.  $< 25\%$  or  $> 75\%$  probability), 15 were correct, with only 1 false positive and 2 false negative predictions (Fig. 2C). The approach therefore assigns neutral predictions much more frequently than false positive or false negative predictions, suggesting that its outputs are highly conservative, and thus unlikely to produce artefacts. Predictions were generated for phenotypes that have not yet been described in  $C_3$ - $C_4$  species (Appendix Fig. 9). Quantitative real-time PCR experimentally verified a subset of these, relating to abundance of  $C_4$  enzymes not previously measured (Fig. 2D-E). We also found that the model was able to successfully infer evolutionary dynamics in artificially constructed datasets (Fig. 2A-B). Taken together, these prediction and verification studies illustrate that our approach robustly identifies key features of  $C_4$  evolution.

## A high-resolution model for the evolutionary events generating $C_4$

The posterior probability distributions for the acquisition time of each phenotypic trait were combined to produce an objective, computationally-generated blueprint for the order of evolutionary events generating  $C_4$  photosynthesis (Fig. 3). These results were consistent with previous work on subsets of  $C_4$  lineages that proposed



**Figure 2: Verifying a novel Bayesian approach for predicting evolutionary trajectories.** (A, B) Datasets were obtained from an artificially constructed diagonal dynamic matrix (A), and a diagonal matrix with linked timing of locus acquisitions (B). The single, diagonal evolutionary trajectory was clearly replicated in both examples, over a time-scale of 16 individual steps, or four coarse-grained quartiles. We subjected these artificial datasets to our inferential machinery with fully characterised artificial species, and with 50% of data occluded in order to replicate the proportion of missing data from our  $C_3$ - $C_4$  dataset. (C) When applied to our meta-analysis of  $C_3$ - $C_4$  data, predictions were generated for every trait missing from the biological dataset. We tested this predictive machinery by generating 29 artificial datasets, each missing one data point, and comparing the presence/absence of the trait as predicted by our approach with the experimental data from the original study. (D, E) Quantitative real-time PCR (qPCR) was used to verify the predicted phenotypes of four  $C_3$ - $C_4$  species. The abundance RbcS (D) and MDH (E) transcripts was determined from six *Flaveria* species. White bars represent phenotypes already determined by other studies, grey bars those that were predicted by the model and asterisks denote intermediate species phenotypes correctly predicted by our approach (Error bars indicate SEM,  $N=3$ ).

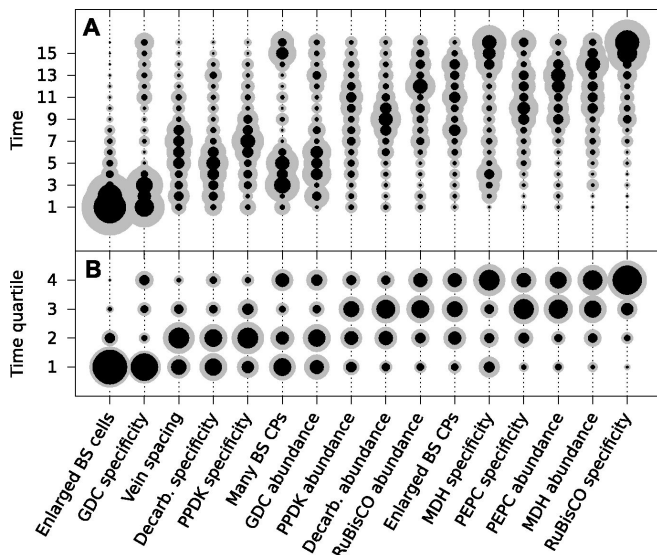


Figure 3: **The mean ordering of phenotypic changes generating  $C_4$  photosynthesis.** EM-clustered data from  $C_3$ - $C_4$  intermediate species were used to generate posterior probability distributions for the timing of the acquisition of  $C_4$  traits in sixteen evolutionary steps (A) or four quartiles (B). Circle diameter denotes the mean posterior probability of a trait being acquired at each step in  $C_4$  evolution (the Bayes estimator for the acquisition probability). Halos denote the standard deviation of the posterior. The 16 traits are ordered from left to right by their probability of being acquired early to late in  $C_4$  evolution. Abbreviations: bundle sheath (BS), glycine decarboxylase (GDC), chloroplasts (CPs), decarboxylase (Decarb.), pyruvate, orthophosphate dikinase (PPDK), malate dehydrogenase (MDH), phosphoenolpyruvate carboxylase (PEPC).

the BS-specificity of GDC occurs prior to the evolution of  $C_4$  metabolism [20, 31], and loss of RuBisCO from M cells occurs late [32, 26], but also provided higher resolution insight into the order of events generating  $C_4$  metabolism. Alterations to leaf anatomy as well as cell-specificity and increased abundance of multiple  $C_4$  cycle enzymes were predicted to evolve prior to any alteration to the primary  $C_3$  and  $C_4$  photosynthetic enzymes RuBisCO and phosphoenolpyruvate carboxylase (PEPC) (Fig. 3).

There was also strong evidence for enlargement of BS cells as an early innovation in most  $C_4$  lineages (Fig. 3), consistent with the suggestion that this was an ancestral state within  $C_3$  ancestors of  $C_4$  grass lineages and that this contributed to the high number of  $C_4$  origins within this family [33, 34]. The compartmentation of PEPC into M cells and its increased abundance compared with  $C_3$  leaves was predicted to occur at similar times, but for all other  $C_4$  enzymes the evolution of increased abundance and cellular compartmentation were

clearly separated by the acquisition of other traits (Fig. 3). This result is consistent with molecular analysis of genes encoding  $C_4$  enzymes that indicates cell-specificity and increased expression are mediated by different cis-elements [15, 17, 35].

Two approaches were taken to verify that these conclusions are robust and accurately reflect biological data. First, the analysis was repeated using scores for presence or absence of traits that were assigned by hierarchical clustering, as opposed to using the EM algorithm (Appendix Fig. 10A). Although hierarchical clustering generated differences in the scoring of a small number of traits, the predicted evolutionary trajectories were not affected, producing highly similar results (Appendix Fig. 10B). Second, we introduced structural changes to the phenotype space, by both adding and subtracting traits from the analysis (Appendix Fig. 11). Removing two independent pairs of traits from the analysis did not affect the predicted timing of the remaining 14 traits (Appendix Fig. 11A-B). However, increased standard deviations were observed in some cases (e.g. for the probabilities of acquiring enlarged BS cells, or decreased vein spacing) likely a consequence of using fewer data. To test if the addition of data might also affect the results, we performed an analysis with two additional traits included (Appendix Fig. 11C). We selected two traits that have been widely observed in  $C_3$ - $C_4$  species, the centripetal positioning of mitochondria and the centrifugal or centripetal position of chloroplasts within BS cells [20]. Despite the widespread occurrence of these traits, their functional importance remains unclear [20]. Consistent with observations made from several genera, we predict that these cellular alterations are acquired early in the evolution of  $C_4$  photosynthesis [31, 36, 37, 38]. Importantly, including these additional early traits in the analysis did not alter the predicted order of the original 16 traits. Together, these analyses did not alter our main conclusions, suggesting that they are robust.

## The order of $C_4$ trait evolution is flexible

In addition to the likely order of evolutionary events generating  $C_4$  photosynthesis, the number of molecular alterations required is also unknown. We therefore aimed to test if multiple traits were predicted to evolve with linked timing, and therefore likely mediated by a single underlying mechanism. To achieve this, we performed a contingency analysis by considering trajectories across phenotype space beginning with a given initial acquisition step. In this analysis, the starting genome had one of the 16 traits acquired and the rest absent, and the contingency of the subsequent trajectory upon the initial step was recorded. This approach was designed to test if acquiring one  $C_4$  trait increased the probability of subsequently acquiring other traits, thus detecting if the evolution of multiple traits is linked by underlying mechanisms. Inflexible linkage between multiple traits was

detected in artificial positive control datasets (Fig. 2B) but not in the  $C_3$ - $C_4$  dataset (Appendix Fig. 12). This result suggests that the order of  $C_4$  trait acquisition is flexible. Multiple origins of  $C_4$  may therefore have been facilitated by this flexibility in the evolutionary pathways connecting  $C_3$  and  $C_4$  phenotypes.

## $C_4$ evolved via multiple distinct evolutionary trajectories

Our Bayesian analysis strongly indicates that there are multiple evolutionary pathways by which  $C_4$  traits are acquired by all lineages of  $C_4$  plants. First, no single sequence of acquisitions was capable of producing intermediate phenotypes compatible with all observations (see Materials and Methods). Second, several traits such as compartmentation of GDC into BS and the increased number of chloroplasts in the BS clearly displayed bimodal probability distributions for their acquisition (Fig. 3). This bimodality is indicative of multiple distinct pathways to  $C_4$  photosynthesis that acquire traits at earlier or later times. To investigate factors underlying this bimodality, we inferred evolutionary pathways generating the  $C_4$  leaf using data from monocot and eudicot lineages, or from lineages using NAD malic enzyme (NAD-ME) or NADP malic enzyme (NADP-ME) as their primary decarboxylase. PCA on the entire set of inferred transition networks for monocot and dicot subsets revealed distinct separation (Fig. 4A), suggesting that the topology of the evolutionary landscape surrounding  $C_4$  is largely different for these two anciently diverged taxa. Performing this PCA including networks that were inferred from the full data set (with both lineages) confirmed that this separation is a robust result and involves posterior variation on a comparable scale to that of the full set of possible networks (Appendix Fig. 13). Analysis of the posterior probabilities of the mean pathways representing either monocots or dicots revealed that this separation is the result of differences in the timing of events generating both anatomical and biochemical traits (Fig. 4C). We propose that the ancient divergence of the monocot and eudicot clades constrained the evolution of  $C_4$  photosynthesis to broadly different evolutionary pathways in each.

There was more overlap between the landscapes generating NAD-ME and NADP-ME species (Fig. 4B), likely reflecting the convergent origins of NAD-ME and NADP-ME sub-types [4, 14]. Despite the traditional definition of these lineages on the basis of biochemical differences, we detected differences in the timing of their anatomical evolution (Fig. 4D). For example, in NAD-ME lineages, increased vein density was predicted to be acquired early in  $C_4$  evolution, while in NADP-ME species this trait showed a broadly different trajectory (Fig. 4D). The proliferation of chloroplasts in the BS was also acquired with different timings between the two sub-types. The alternative evolutionary pathways generating the NADP-ME

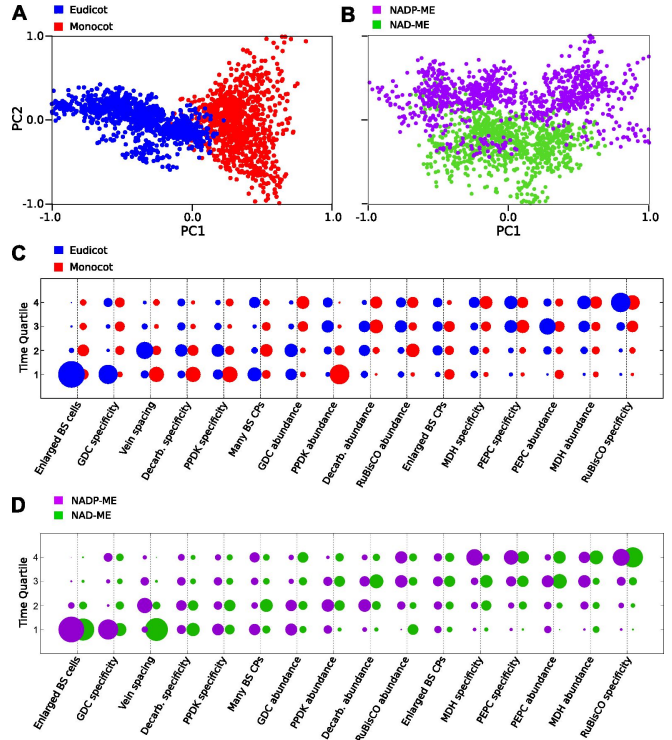


Figure 4: **Differences in the evolutionary events generating different  $C_4$  sub-types and distantly related taxa.** Principal component analysis (PCA) on the entire landscape of transition probabilities using only monocot and eudicot data (A) and data from NADP-ME and NAD-ME sub-type lineages (B) shows broad differences between the evolutionary pathways generating  $C_4$  in each taxon. Monocots and eudicots differ in the predicted timing of events generating  $C_4$  anatomy and biochemistry (C), whereas NADP-ME and NAD-ME lineages differ primarily in the evolution of decreased vein spacing and greater numbers of chloroplasts in BS cells (D).

and NAD-ME subtypes were therefore defined by differences in the timing of anatomical and cellular traits that are predicted to precede the majority of biochemical alterations (Figs. 3, 4D). We therefore conclude that these distinct sub-types evolved as a consequence of alternative evolutionary histories in response to non-photosynthetic pressures. Furthermore, we propose that early evolutionary events determined the downstream phenotypes of  $C_4$  sub-types by restricting lineages to independent pathways across phenotype space.

## Discussion

### A novel Bayesian technique for inferring stochastic trajectories

The adaptive landscape metaphor has provided a powerful conceptual framework within which evolutionary transitions can be modelled [24, 39, 29]. However, the majority of complex biological traits provide numerous challenges in utilising such an approach, including missing phenotypic data, incomplete phylogenetic information and in the case of convergent evolution, variable ancestral states. Here we report the development of a novel, predictive Bayesian approach that is able to infer likely evolutionary trajectories connecting phenotypes from sparsely sampled, highly stochastic data. With this model, we provided insights into the evolution of one of the most complex traits to have arisen in multiple lineages;  $C_4$  photosynthesis. However, as our approach is not dependent on detailed phylogenetic inference, we propose that it could be used to model the evolution of other complex traits, such as those in the fossil record, which are also currently limited by the fragmented nature of data available [40]. Our approach is also not limited by the time-scale over which predicted trajectories occur. As a result, it may be useful in inferring pathways underlying stochastic processes occurring over much shorter timescales, such as disease or tumour progression, or the differentiation of cell types.

### $C_4$ evolution was initiated by non-photosynthetic drivers

A central hypothesis for the ecological drivers of  $C_4$  evolution is that declining  $CO_2$  concentration in the Oligocene decreased the rate of carboxylation by RuBisCO, creating a strong pressure to evolve alternative photosynthetic strategies [9, 41]. According to this hypothesis, alterations to the localisation and abundance of the primary carboxylases PEPC and RuBisCO would be expected to occur early in the evolutionary trajectories generating  $C_4$ . Conversely, our data suggest that alterations to anatomy and cell biology were predicted to precede the majority of biochemical alterations, and that other enzymes of the  $C_4$  pathway are recruited prior to PEPC and RuBisCO (Fig. 3). These enzymes, such as

PPDK and  $C_4$  acid decarboxylases, function in processes not related to photosynthesis within leaves of  $C_3$  plants (Aubry et al. 2011), so the early changes to abundance and localisation of these enzymes within  $C_4$  lineages may have been driven by non-photosynthetic pressures. A recent *in silico* study also predicts that changes to photorespiratory metabolism and GDC in BS cells evolved prior to the  $C_4$  pathway [42]. Our model predicts that BS-specificity of GDC was acquired early in  $C_4$  evolution for the majority of lineages. However, we also note that the predicted timing of GDC BS-specificity is bimodal in our analysis (Fig. 3), and not predicted to be acquired early in monocot lineages (Fig. 4C). These results suggest that this is not a feature of  $C_4$  evolution to have occurred repeatedly in all lineages.

Recent evidence from physiological and ecological studies have identified a number of additional environmental pressures that may have driven the evolution and radiation of  $C_4$  lineages, including high evaporative demands [43] and increased fire frequency [10]. Increased BS volume and vein density have been proposed as likely adaptations to improve leaf hydraulics under drought [43, 33], but nothing is known about how early recruitment of GDC, PPDK and  $C_4$  acid decarboxylases (Fig. 3) may relate to these pressures. A better understanding of the mechanisms underlying the recruitment of these enzymes [16, 17, 35] may help identify the key molecular events facilitating  $C_4$  evolution.

Our data also suggest that modifications to leaf development drove the evolution of diverse  $C_4$  sub-types. For examples, we find that differences in the timing of events altering leaf vascular development and BS chloroplast division occur prior to the appearance of the alternative evolutionary pathways generating the NADP-ME and NAD-ME biochemical sub-types (Fig. 4D). These traits are predicted to evolve prior to any alterations to the  $C_4$  acid decarboxylase enzymes that traditionally define these sub-types [14]. As an homologous mechanism has been shown to regulate the cell-specificity of gene expression in both NADP-ME and NAD-ME gene families in independent lineages [16], it is unlikely that mechanisms underlying the recruitment of these enzymes drove the evolution of distinct sub-types. We therefore conclude that these different sub-types evolved as a consequence of alternative evolutionary histories in leaf development, rather than biochemical or photosynthetic pressures. This may explain why differences in the carboxylation efficiency or photosynthetic performance of different  $C_4$  sub-types have never been detected [14], making the adaptive significance of different decarboxylation mechanisms difficult to explain. Instead, we propose that early evolutionary events determined the downstream phenotypes of  $C_4$  sub-types by restricting lineages to independent pathways across phenotype space. The numerous differences in leaf development and cell biology between  $C_4$  sub-types [14] may provide clues as to which developmental changes underlie subsequent differences in



metabolic evolution.

## Convergent evolution was facilitated by flexibility in evolutionary trajectories

C<sub>4</sub> photosynthesis provides an excellent example of how independent lineages with a wide range of ancestral phenotypes can converge upon similar complex traits. Several studies on more simple traits have demonstrated that convergence upon a phenotype can be specified by diverse genotypes, and thus non-homologous molecular mechanisms in independent lineages [44, 6, 5]. Taken together, our data also indicate that flexibility in the viable series of evolutionary events has also facilitated the convergence of this highly complex trait. First, we show that at least four distinct evolutionary trajectories underlie the evolution of C<sub>4</sub> lineages (Fig. 4). Second, we find no evidence for inflexible linkage between the predicted timing of distinct C<sub>4</sub> traits (Appendix Fig. 12). This diversity in viable pathways also helps explain why C<sub>4</sub> has been accessible to such a wide variety of species and not limited to a smaller subset of the angiosperm phylogeny. A recent model for the evolution of the biochemistry associated with the C<sub>4</sub> leaf also found that C<sub>4</sub> photosynthesis was accessible from any surrounding point of a fitness landscape [42]. Our study of C<sub>4</sub> anatomy, biochemistry and cell biology also suggests the C<sub>4</sub> phenotype is accessible from multiple trajectories. Encouragingly, the trajectories predicted by Heckman et al (2013) were found to pass through phenotypes of C<sub>3</sub>-C<sub>4</sub> species, despite the fact that these species were not used to parameterise the evolutionary landscape. As different mechanisms generate increased abundance and cell-specificity for the majority C<sub>4</sub> enzymes in independent C<sub>4</sub> lineages (reviewed in [11, 18]), it is likely that mechanistic diversity underlies the multiple evolutionary pathways generating C<sub>4</sub> photosynthesis and may be a key factor in facilitating the convergent evolution of complex traits. This may benefit efforts to recapitulate the acquisition of C<sub>4</sub> photosynthesis through the genetic engineering of C<sub>3</sub> species [7], expanding the molecular toolbox available to establish C<sub>4</sub> traits in distinct phenotypic backgrounds.

## Acknowledgements

We thank the BBSRC and MRC for funding, and S. Kelly, J.A. Langdale, H. Griffiths, and N. Jones for advice. B.P.W conducted the meta-analysis and performed qPCR experiments, I.G.J. designed and performed the computational modelling and tested predictions computationally. S.C. extracted RNA for qPCR analysis. B.P.W and I.G.J analysed data and generated Figures and B.P.W, I.G.J and J.M.H wrote the paper. The complete study was supervised by J.M.H. All authors discussed the results and commented on the manuscript. **Note:** source data, and correspondence pertaining to

publication, may be viewed on the journal website associated with this article.

## References

- [1] Kozmik Z., Ruzickova J., Jonasova K., Matsumoto Y., Vopalensky P., and Kozmikova I. et al. Assembly of the cnidarian camera-type eye from vertebrate-like components. *PNAS*, 105:8989–93, 2008.
- [2] Santos J. C., Coloma L. A., and Cannatella D. C. Multiple, recurring origins of aposematism and diet specialization in poison frogs. *PNAS*, 100:12792–7, 2003.
- [3] Wilson J. S., Williams K. A., Forister M. L., von Dohlen C. D., and Pitts J. P. Repeated evolution in overlapping mimicry rings among north american velvet ants. *Nature Communications*, 3:1272, 2012.
- [4] Sage R. F., Christin P.-A., and Edwards E. J. The c<sub>4</sub> plant lineages of planet earth. *J. Exp. Bot.*, 62:3155–69, 2011.
- [5] Steiner C. C., Rmpler H., Boettger L. M., Schneberg T., and Hoekstra H. E. The genetic basis of phenotypic convergence in beach mice: similar pigment patterns but different genes. *Mol. Biol. Evol.*, 26:35–45, 2009.
- [6] Hill R. C., Egydio de Carvalho C., Salogianis J., Schlager B., Pilgrim D., and Haag E. S. Genetic flexibility in the convergent evolution of hermaphroditism in caenorhabditis nematodes. *Developmental Cell*, 10:531–8, 2006.
- [7] Hibberd J. M., Sheehy J. E., and Langdale J. A. Using c<sub>4</sub> photosynthesis to increase the yield of rice-rationale and feasibility. *Curr. Opin. Plant Biol.*, 11:228–31, 2008.
- [8] Christin P.-A., Sage T. L., Edwards E. J., Ogburn R. M., Khoshravesht R., and Sage R. F. Complex evolutionary transitions and the significance of c<sub>3</sub>-c<sub>4</sub> intermediate forms of photosynthesis in molluginaceae. *Evolution*, 65:643–60, 2011.
- [9] Vicentini A., Barber J. C., Aliscioni S. S., Giusani L. M., and Kellogg E. A. The age of the grasses and clusters of origins of c<sub>4</sub> photosynthesis. *Global Change Biology*, 14:2963–77, 2008.
- [10] Edwards E. J., Osborne C. P., Stromberg C. A. E., Smith S. A., C<sub>4</sub> Grasses Consortium, Bond W. J., and et al. The origins of c<sub>4</sub> grasslands: integrating evolutionary and ecosystem science. *Science*, 328:587–91, 2010.
- [11] Langdale J. A. C<sub>4</sub> cycles: past, present, and future research on c<sub>4</sub> photosynthesis. *Plant Cell*, 23:3879–92, 2011.



- [12] Zhu X-G., Long S. P., and Ort D. R. What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Cur. Opin. Biotechnol.*, 19:153–9, 2008.
- [13] Hatch M. D., Kagawa T, and Craig S. Subdivision of c4 pathway species based on differing c4 acid decarboxylating systems and ultrastructural features. *Aus. J. Plant Physiol.*, 2:111, 1975.
- [14] Furbank R. T. Evolution of the c4 photosynthetic mechanism: are there really three c4 acid decarboxylation types? *J. Exp. Bot.*, 62:3103–8, 2011.
- [15] Akyildiz M., Gowik U., Engelmann S., Koczor M., Streubel M., and Westhoff P. Evolution and function of a cis-regulatory module for mesophyll-specific gene expression in the c4 dicot flaveria trinervia. *Plant Cell*, 19:3391–402, 2007.
- [16] Brown N. J., Newell C. A., Stanley S., Chen J. E., Perrin A. J., Kajala K., and Hibberd J. M. Independent and parallel recruitment of preexisting mechanisms underlying c4 photosynthesis. *Science*, 331:1436–9, 2011.
- [17] Kajala K., Brown N. J., Williams B. P., Borrill P., Taylor L. E., and Hibberd J. M. Multiple arabidopsis genes primed for recruitment into c4 photosynthesis. *Plant J.*, 69:47–56, 2012.
- [18] Williams B. P., Aubry S., and Hibberd J. M. Molecular evolution of genes recruited into c4 photosynthesis. *Trends Plant Sci.*, 17:213–20, 2012.
- [19] Christin P., Salamin N., Savolainen V., Duvall M., and Besnard G. C4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Current Biology*, 17:1241–7, 2007.
- [20] Sage R. F., Sage T. L., and Kocacinar F. Photorespiration and the evolution of c4 photosynthesis. *Annu. Rev. Plant Biol.*, 2012.
- [21] Hattersley P. W. and Stone N. E. Photosynthetic enzyme activities in the c3-c4 intermediate neurachne minor. *Aus. J. Plant Physiol.*, 13:399, 1986.
- [22] Moore B. D., Franceschi V. R., Cheng S-H., Wu J., and Ku M. S. B. Photosynthetic characteristics of the c3-c4 intermediate parthenium hysterophorus. *Plant Physiol.*, 85:978–83, 1987.
- [23] Wright S. The roles of mutation, indbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, 1:356–66, 1932.
- [24] Gavrillets S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.*, 12:307–12, 1997.
- [25] McKown A. D, Moncalvo J-M., and Dengler N. G. Phylogeny of flaveria (asteraceae) and inference of c4 photosynthesis evolution. *Am. J. Bot.*, 92:1911–28, 2005.
- [26] Khoshravesh R., Hossein A., Sage T. L., Nordenstam B., and Sage R. F. Phylogeny and photosynthetic pathway distribution in anticharis endl. (*Scrophulariaceae*). *J. Exp. Bot.*, 63:5645–58, 2012.
- [27] Mooers A. O. and Heard S. B. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72:31–54, 2013.
- [28] Weinreich D. M., Watson R. A., and Chao L. Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, pages 1165–74, 2005.
- [29] Lobkovsky A.E., Wolf Y. I., and Koonin E. V. Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput. Biol.*, 7:–1002302, 2011.
- [30] Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–86, 1989.
- [31] Hylton C. M., Rawsthorne S., Smith A. M., Jones D. A., and Woolhouse H. W. Glycine decarboxylase is confined to the bundle-sheath cells of leaves of c3-c4 intermediate species. *Planta*, 175:452–9, 1988.
- [32] Cheng S. H., Moore B. D., Edwards G. E., and Ku M. S. Photosynthesis in flaveria brownii, a c4-like species: leaf anatomy, characteristics of co2 exchange, compartmentation of photosynthetic enzymes, and metabolism of co2. *Plant Physiol.*, 87:867–73, 1988.
- [33] Griffiths H., Weller G., Toy L. F. M., and Dennis R. J. Youre so vein: bundle sheath physiology, phylogeny and evolution in c3 and c4 plants. *Plant Cell Environ.*, 36:249–61, 2013.
- [34] 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. Anatomical enablers and the evolution of c4 photosynthesis in grasses. *PNAS*, 110:1381–6, 2013.
- [35] Wiludda C., Schulze S., Gowik U., Engelmann S., Koczor M., Streubel M., Bauwe H., and Westhoff P. Regulation of the photorespiratory gldpa gene in c4 flaveria: an intricate interplay of transcriptional and post-transcriptional processes. *Plant Cell*, 24:137–51, 2012.
- [36] McKown A. D. and Dengler N. G. Key innovations in the evolution of kranz anatomy and c4 vein pattern in flaveria (asteraceae). *Am. J. Bot.*, 94:382–99, 2007.

- [37] Muhaidat R., Sage T. L., Frohlich M. W., Dengler N. G., and Sage R. F. Characterization of c3-c4 intermediate species in the genus *heliotropium* 1. (*Boraginaceae*): *anatomy, ultrastructure and enzyme activity*. *Plant Cell Environ.*, 34:1723–36, 2011.
- [38] Sage T. L., Sage R. F., Vogan P. J., Rahman B., Johnson D. C., Oakley J. C., and Heckel M. A. The occurrence of c2 photosynthesis in euphorbia subgenus *chamaesyce* (euphorbiaceae). *J. Exp. Bot.*, 62:3183–95, 2011.
- [39] Whibley A. C., Langlade N. B. and aldo C., Hanna A. I., Bangham A., Thbaud C., and Coen E. Evolutionary paths underlying flower color variation in *antirrhinum*. *Science*, 313:963–6, 2006.
- [40] Kidwell S. M. and Holland S. M. The quality of the fossil record: implications for evolutionary analyses. *Ann. Rev. Ecol. Systematics*, 33:561–88, 2002.
- [41] Christin P., Besnard G., Samaritani E., Duvall M., Hodkinson T., Savolainen V., and Salamin N. Oligocene co2 decline promoted c4 photosynthesis in grasses. *Current Biology*, 18:37–43, 2008.
- [42] Heckman D., Schulze S., Denton A., Gowik U., Westhoff P., Weber A. P. M., and Lercher M. J. Predicting c4 photosynthesis evolution: modular, individually adaptive steps on a mount fuji fitness landscape. *Cell.*, 153:1579–1588, 2013.
- [43] Osborne C. P. and Sack L. Evolution of c4 plants: a new hypothesis for an interaction of co2 and water relations mediated by plant hydraulics. *Phil. Trans. Roy. Soc. B: Biol. Sci.*, 367:583–600, 2012.
- [44] Wittkopp P. J., Williams B. L., Selegue J. E., and Carroll S. B. *Drosophila* pigmentation evolution: divergent genotypes underlying convergent phenotypes. *PNAS*, 100:1808–13, 2003.
- [45] Abramoff M. D., Magalhaes P. J., and Ram S. J. Image processing with imagej. *Biophotonics International*, 11:36–42, 2004.
- [46] Roweis S. Em algorithms for pca and spca. *In Advances in Neural Information Processing Systems*, 62632. 10. Cambridge, MA:—, 1998.
- [47] Revuz D. and Yor M. Continuous martingales and brownian motion. *Springer. Rosche E., Streubel M., and Westhoff P.*, 1994. Primary structure of the photosynthetic pyruvate orthophosphate dikinase of the C3 plant *Flaveria pringlei* and expression analysis of pyruvate orthophosphate dikinase sequences in C3, C3-C4 and C4 *Flaveria* species. *Plant Mol. Bio.* 26:763–9, 1999.
- [48] Johnson M. T. J., Carpenter E. J., Tian Z., Bruskiwich R., Burris J. N., Carrigan C. T., and et al. Evaluating methods for isolating total rna and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS ONE*, 7:–50226, 2012.
- [49] Bustin S. A. Absolute quantification of mrna using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.*, 25:169–93, 2000.
- [50] Livak K. J. and Schmittgen T. D. Analysis of relative gene expression data using real-time quantitative pcr and the 2-ddct method. *Methods*, 25:402–8, 2001.

# Appendix

## Materials & Methods

**Biological data from  $C_4$  intermediates.** Data from eighteen  $C_3$ , seventeen  $C_4$  and thirty-seven  $C_3$ - $C_4$  species were consolidated from 43 studies that have examined the phenotypic characteristics of  $C_3$ - $C_4$  species since their discovery in 1974 (Supplementary File 1). Values for sixteen of the most widely-studied  $C_3$  characteristics were recorded for each intermediate species, as well as congeneric  $C_3$  and  $C_4$  relatives where available. The majority of data on enzyme abundance and the number and size of bundle sheath (BS) chloroplasts were obtained from studies employing the same methodology and were thus cross-comparable. These cross-comparable quantitative data were partitioned into presence absence scores using two clustering techniques, the expectation-maximisation (EM) algorithm and hierarchical clustering (Appendix Fig. 10). EM clustering was performed using a one-dimensional mixture model with two assigned components (e.g. presence and absence clusters), allowing for variable variance between the two components of the model, and variable population size between the two components. Hierarchical clustering was performed using a complete-linkage agglomerative approach, partitioning clusters by maximum distance according to a Euclidean distance metric. This approach identifies clusters with common variance, thus contrasting with the clusters of variable variance identifiable by EM.

For quantitative data not comparable with other studies, values obtained for intermediate species were scored as 1 or 0 if they were closer to the values for the respective  $C_4$  or  $C_3$  controls used in the original study. For qualitative abundance data from immunoblots, relative band intensity was measured using ImageJ software [45] and abundance was scored as 1 or 0 if the band intensity value was closer to the  $C_4$  or  $C_3$  control respectively. For qualitative cell-specificity data from immunolocalisations, a presence score was only assigned if the enzyme appeared completely absent from either mesophyll (M) or BS cells. We represent the phenotypic properties of each intermediate species as a string of  $L = 16$  numbers (Fig. 1). We will refer to these strings as *phenotype strings* of  $L$  loci, with each locus describing data on the corresponding phenotypic trait. In a given locus, 0 denotes the absence of a  $C_4$  trait, 1 denotes the presence of a  $C_4$  trait, and 2 denotes missing data.

**Principal component analysis (PCA).** PCA was performed on five variables for  $C_4$  cycle enzyme activity, with missing values estimated using the EM algorithm for PCA as described by [46].

**Model transition networks.** The fundamental

element underlying our analysis is a transition network  $P$ , consisting of a directed graph with  $2^L = 65536$  nodes, and the weight of the edge  $P_{ij}$  denoting the probability of a transition occurring from node  $i$  to node  $j$ . Each node corresponds to a possible phenotype: we labeled nodes with labels  $l_i$  so that  $l_i$  was the binary representation of the phenotype at node  $i$ , and  $P_{ij}$  takes on the specific meaning of *the probability of a transition from phenotype  $l_i$  to phenotype  $l_j$* . We made several restrictions on the structure of  $P$ . We allowed only transitions that change a given phenotype at one locus, so  $P_{ij} = 0$  if  $H(l_i, l_j) \neq 1$ , where  $H(b_1, b_2)$  is the Hamming distance between bitstrings  $b_1$  and  $b_2$ . Transitions that changed loci with value 1 to value 0 (steps back towards the  $C_3$  state) were forbidden, so  $P_{ij} = 0$  if  $H(l_i, l_0) > H(l_j, l_0)$ , where  $l_0$  is the phenotypic string containing only zeroes. We assume that the possibility of events involving backwards steps, and multiple simultaneous trait acquisitions, constitute second-order effects which will not strongly influence the inferred evolutionary dynamics.

**Evolutionary trajectories.** Given the transition network  $P$ , we modelled the evolutionary trajectories that may give rise to  $C_4$  photosynthesis through the picture of a discrete analogue to a Brownian bridge, i.e. as a stochastic process on  $P$  with constrained start and end positions [47]. We enforced the start state of the process to be  $l_{C_3} \equiv l_0 = 0\dots 0$  (the phenotype string of all zeroes) and the end state, through the imposed structure of  $P$ , to be  $l_{C_4} \equiv l_{2^L-1} = 1\dots 1$  (the string of all ones). The dynamics of the process between these points consisted of  $L$  steps, with a phenotypic trait being acquired at each step, and a step from node  $i$  to node  $j$  occurring with probability  $P_{ij}$ .

**Sampling intermediates.** As many evolutionary trajectories may lead to the acquisition of the required phenotypic traits, we considered an ensemble of evolutionary trajectories for each transition network. Each member of this ensemble is started at  $l_{C_3}$  and allowed to step across the network according to probabilities  $P$ .

To compare the dynamics of a given transition network to the properties of observed biological intermediates, we pictured this ensemble of trajectories as a modification of a hidden Markov model (HMM [30]). At each timestep in each individual trajectory, the process may with some probability emit a signal to the observer, with that signal being simply  $l_i$ , the label of the node at which the process currently resides. Over an ensemble of trajectories, a set of randomly-emitted signals is thus built up (Appendix Fig. 8).

We define a *compatibility function* between two strings as

ensembles to broadly sample available paths on networks.

$$C(s, t) = \prod_{i=1}^L c(s_i, t_i) \quad (1)$$

$$c(s_i, t_i) = \begin{cases} 1 & \text{if } s_i = t_i \text{ or } s_i = 2 \text{ or } t_i = 2; \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

$C(s, t)$  thus returns 1 if a signal comprising string  $s$  could be responsible for observation  $t$  once some of the loci within  $s$  have been obscured: signal  $s$  is compatible with observation  $t$ .

**Likelihood of observing biological data.** We wish to compute the likelihood of observing biological data  $B$  given a transition network  $P$ . Under our model, this likelihood is calculated by considering the compatibility of randomly emitted signals from processes supported by  $P$  with the observed data  $B$ . We write

$$\mathcal{L}(P|B) = \prod_i \sum_{\text{chains } x} \sum_{\text{signals } s} \mathbb{P}_{\text{chain}}(x|P) \mathbb{P}_{\text{emission}}(s|x) C(s, B_i) \quad (3)$$

Here,  $\mathbb{P}_{\text{chain}}(x|P)$  is the probability of specific trajectory  $x$  arising on network  $P$ ;  $\mathbb{P}_{\text{emission}}(s|x)$  is the probability that trajectory  $x$  emits signal  $s$ , and  $C(s, B_i)$  gives the compatibility of signal  $s$  with intermediate state  $B_i$ . The term within the product operator thus describes the probability that evolutionary dynamics on network  $P$  give rise to a signal that is compatible with species  $B_i$ , with the overall likelihood being the product of this probability over all observed species.

**Simulation.** The uniform and random nature of signal emission means that  $\mathbb{P}_{\text{emission}}(s|x)$  is a constant if signal  $s$  can be emitted from trajectory  $x$ , and zero otherwise. Our simulation approach only produces signals which can be emitted from the trajectory under consideration, so  $\mathbb{P}_{\text{emission}}(s|x)$  will always take the same constant value (which depends on the probability of signal emissions). As we will be considering ratios of network likelihoods and will not be concerned with absolute likelihoods we will ignore this term henceforth. For each network  $P$  we simulate an ensemble of  $N_{\text{chain}}$  trajectories and, for each node encountered throughout this ensemble, we record compatibilities with each of the biologically observed intermediates. We sum these compatibilities over the ensemble, obtaining  $\sum_{\text{chains } x} \mathbb{P}_{\text{chain}}(x|P) C(s, B_i)$ . A network that does not encounter any node compatible with a particular intermediate will thus be assigned zero likelihood; networks that encounter compatible nodes many times will be assigned high likelihoods.

For each transition network, we simulated  $N_{\text{chain}} = 2 \times 10^4$  individual trajectories running from  $C_3$  to  $C_4$ . This value was chosen after preliminary investigations to analyse the ability of trajectory

**Bayesian MCMC over compatible networks.** Given uninformative prior knowledge about the evolutionary dynamics leading to  $C_4$  photosynthesis (specifically, our prior involves each possible transition from a given node being assigned equal probability), we aimed to build a posterior distribution over a suitable description of the evolutionary dynamics. We represented the dynamics supported on a network  $P$  through a matrix  $\pi$ , where  $\pi_{i,n}$  describes the probability that acquisition of trait  $i$  occurs at the  $n$ th step in an evolutionary trajectory. The values of matrix  $\pi_{i,n}$  were built up from sampling over the ensemble of trajectories simulated on  $P$ .

We used Bayesian MCMC to sample networks based on their associated likelihood values (Wasserman, 2004). At each iteration, we perturbed the transition probability of the current network  $P$  a small amount (see below) to yield a new trial network  $P'$ . We calculated  $\mathcal{L}(P'|B)$  and accepted  $P'$  as the new network if  $\frac{\mathcal{L}(P'|B)}{\mathcal{L}(P|B)} > r$ , where  $r$  was taken from  $\mathcal{U}(0, 1)$ . For practical reasons we implemented this scheme using log-likelihoods.

The perturbations we applied to transition probabilities are Normally distributed in logarithmic space: for each edge  $w_{ij}$  we used  $w'_{ij} = \exp(\ln w_{ij} + \mathcal{N}(0, \sigma^2))$ . To show that this scheme obeys detailed balance, consider two states  $A$  and  $A'$ , for simplicity described by a one-dimensional scalar quantity. Consider the proposed move from  $A$  when  $\Delta$  is the result of the random draw. This proposal is  $A \rightarrow A'$  if  $A' = \exp(\ln A + \Delta) = Ae^\Delta$ , implying that  $A = A'e^{-\Delta}$ . The probability of proposing move  $A \rightarrow A'$  is thus  $\mathcal{N}(x = \Delta|0, \sigma^2)$ , and the probability of proposing  $A' \rightarrow A$  is  $\mathcal{N}(x = -\Delta|0, \sigma^2)$ . By the symmetry of the Normal distribution, these two probabilities are equal.

We started each MCMC run with a randomly initialised transition matrix. We allowed  $2 \times 10^4$  burn-in steps then sampled over a further  $2 \times 10^5$  steps. The value  $\sigma = 0.1$  was chosen for the perturbation kernel. These values were chosen through an initial investigation to analyse the convergence of MCMC runs under different parameterisations. We performed 40 MCMC runs for each experiment and confirmed that the resulting posterior distributions had converged and yielded consistent results.

**Summary dynamic matrices.** We report the posterior distributions  $\mathbb{P}(\pi_{i,n})$  inferred from sampling compatible networks as above. In the coarse-grained time representation, we use  $\pi_{i,n'}^{CG} = \mathbb{P}(\sum_{n=1+4(n'-1)}^{4n'} \pi_{i,n})$ , summing over sets of ordinals of size 4.

We used the transition network  $P$ , rather than a more coarse-grained representation of the evolutionary

dynamics (for example, the summary matrices  $\pi$ ), as the fundamental element within our simulations so as not to discard possible details that would be lost in a coarse-grained approach – for example, the presence of multiple distinct pathways, which may be averaged over in a summary matrix.

**Proofs of principle.** To verify our approach, we constructed artificial data sets, consisting of sets of strings in which phenotypic traits were acquired in a single ordering. Specifically,  $\pi_{i,n} = \delta_{i,n}$ , so the first step always resulted in acquiring the first trait, and so on. To test the approach in a pleiotropic setting, where multiple traits were acquired simultaneously, we also constructed data sets where traits were acquired at only four timesteps, each corresponding to the simultaneous acquisition of four traits. We subjected these datasets to our inferential machinery with all data intact, and with 50% of data points occluded, to determine the sensitivity and robustness of our approach (Fig. 2A-B). The approach accurately determines the ordering of events in both the bare and occluded cases and assigns very similar posterior probability distributions to the ordering of those traits acquired simultaneously.

**Comparing the evolution of multiple C<sub>4</sub> subtypes.** To compare the pathways generating C<sub>4</sub> in monocots and eudicots, and in NADP-ME and NAD-ME sub-type lineages, we performed inference on two data sets:  $B_1$  and  $B_2$ , each comprising phenotype measurements from one of the groups of interest. We reported the posteriors on the resulting summary dynamics  $\mathbb{P}(\pi_{i,n})$  as before, and for the principal components analysis (PCA) we sampled  $10^3$  summary dynamic matrices  $\pi_{i,n}$  from the inferred posterior distribution during the Bayesian MCMC procedure, and performed PCA on these sampled matrices.

**Predictions.** When a simulated chain encountered a phenotypic node compatible with a given biological intermediate, the values of traits corresponding to missing data in the biological data were recorded. These recorded values, sampled over the sampled set of networks, allowed us to place probabilities on the values of biologically unobserved traits inferred from the encounters of compatible dynamics with the corresponding phenotypic possibilities. For example, if 70% of paths on network  $P$  pass through point 101 and 30% pass through point 001, we infer a 70% probability that the missing trait in biological intermediate 201 takes the value 1. Predictions were presented if the inferred probability of a ‘1’ value was  $> 75\%$  (predicting a ‘1’) or  $< 25\%$  (predicting a ‘0’). If one of these inequalities held and the limiting value fell outside one standard deviation of the inferred probability (i.e. for mean  $\mu$  and standard deviation  $\sigma$ ,  $\mu > 0.75$  and  $\mu - \sigma > 0.75$  (predicting a ‘1’) or  $\mu < 0.25$  and  $\mu + \sigma < 0.25$  (predicting a ‘0’)), the

prediction was presented as ‘strict’.

**Acquisition ordering and evidence against a single pathway.** We used a dynamic programming approach to explore whether a deterministic sequence of events, with a trait  $T_n$  always being acquired at timestep  $n$  ( $\pi_{i,n} = \delta_{T_n,n}$ ) was compatible with the biological data. Performing an exhaustive search over sequences of single transitions that were compatible with the observed data, we identified several such sequences that accounted for all but one trait acquisition, but no single sequence exists that accounts for all the data.

**Contingent trait acquisition.** To explore the possibility of multiple traits being acquired simultaneously, we tracked acquisition probabilities for later traits given that a certain trait was acquired first. This tracking was performed over all sampled compatible networks, building up ‘contingent’ acquisition tables  $\gamma$  with the  $i, j$ th element given by  $\mathbb{P}(\pi_{j,2} | \pi_{i,1} = 1)$ ,  $j \neq i$ . If a pair of traits  $i$  and  $j$  were acquired simultaneously, we would expect  $\gamma_{ij}$  and  $\gamma_{ji}$  to both be higher than expected in the non-contingent case (as  $j$  should always appear to be immediately acquired after  $i$  and vice versa).

**Quantitative real-time PCR (qPCR).** RNA was extracted from mature leaves of six *Flaveria* species as part of the One Thousand Plants Consortium ([www.onekp.com](http://www.onekp.com)), using the hot acid phenol protocol as described by [48] (protocol no. 12). cDNA was synthesised from 0.5 g RNA using *Superscript<sup>TM</sup> II* (Life Technologies, Glasgow, U.K.) following manufacturer’s instructions. An oligo dT primer (Roche, Basel, Switzerland) was used to selectively transcribe polyadenylated transcripts. To each RNA sample, 1 fmol GUS transcript was added for use as an exogenous control or ‘RNA spike’, against which measured transcript abundance was normalised as described by Smith et al. (2003).

qPCR was performed as described by [49] using the DNA-binding marker SYBR Green (Sigma Aldrich, St. Louis, MO, USA) according to manufacturer’s instructions. Primers were designed using cDNA sequences for *Flaveria* species available at Genbank (<http://www.ncbi.nlm.nih.gov/genbank>) and synthesised by Life Technologies. Amplification was performed using a Rotor-Gene Q instrument (Qiagen, Hilden, Germany), using the following cycling parameters: 94 °C for 2 minutes, followed by 40 cycles of 94 °C for 20 seconds, 60 °C for 30 seconds, 72 °C for 30 seconds, followed by a 5 minute incubation of 72 °C. Relative transcript abundance was calculated as described by [50].

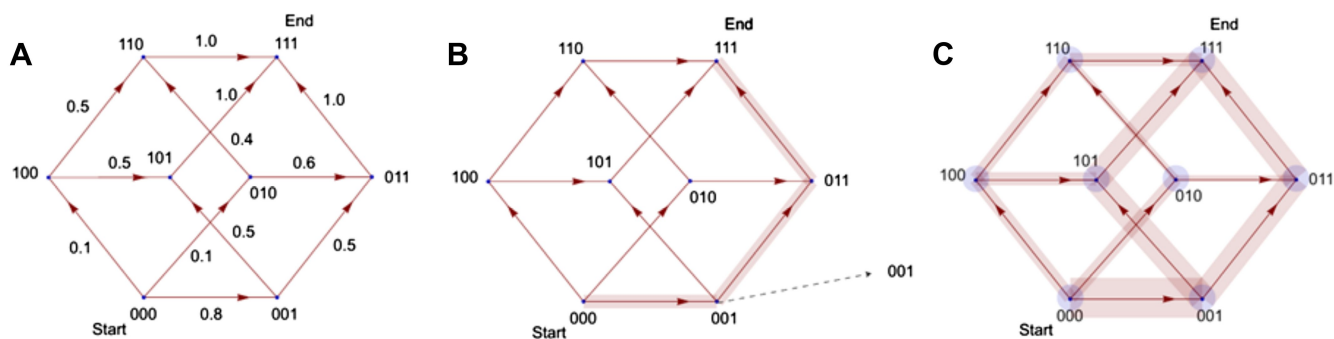


Figure 8: **Illustration of the principle by which evolutionary pathways emit intermediate signals.** In this illustration, the phenotype consists of 3 traits, yielding a simple (hyper)cubic transition network. Simulated trajectories on this network evolve according to the weights of network edges (A). Probabilities were calculated from the signals emitted by simulated trajectories at intermediate nodes (B). Ensembles of trajectories were simulated to obtain probabilities from these signals for every possible evolutionary transition (C).

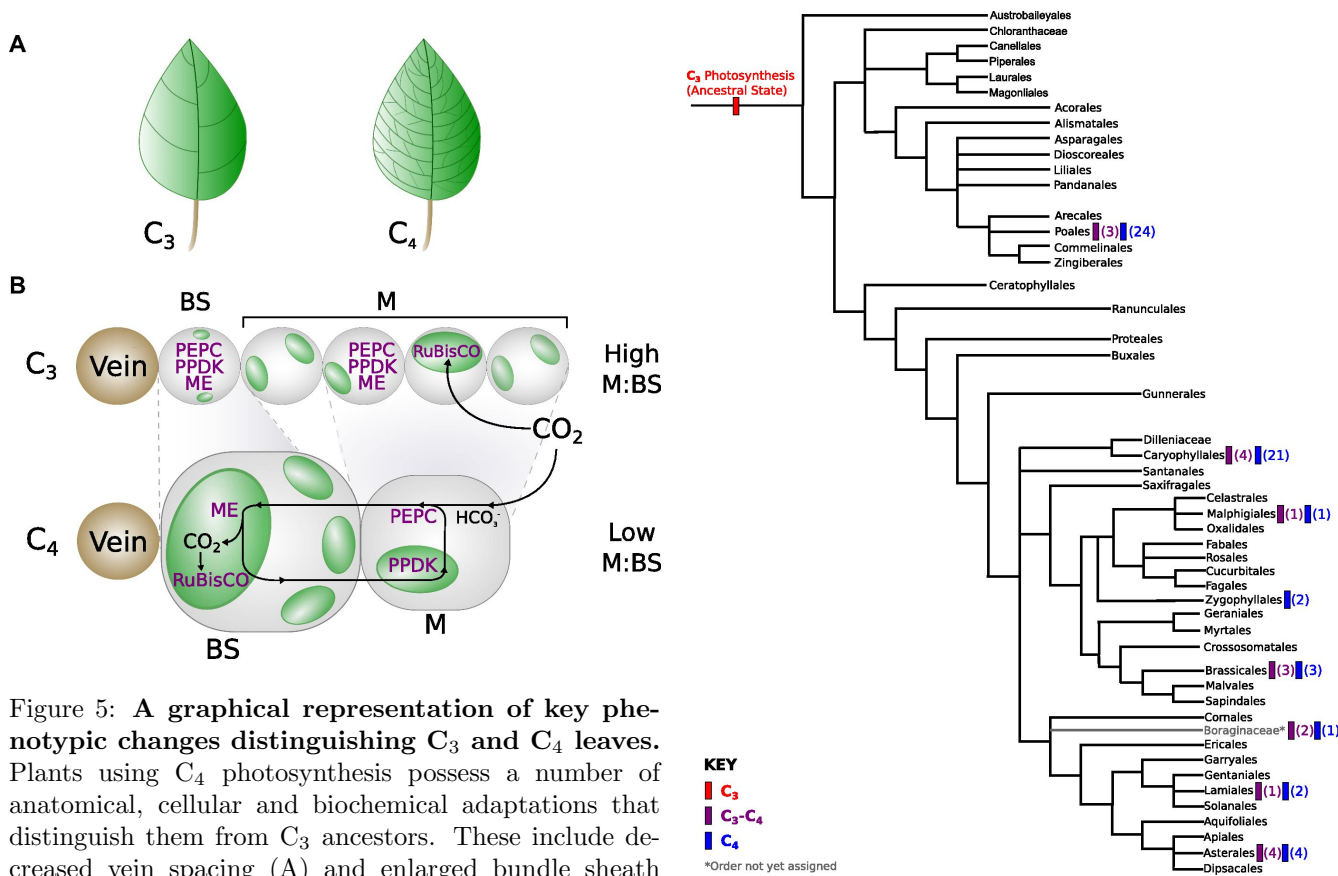


Figure 5: **A graphical representation of key phenotypic changes distinguishing  $C_3$  and  $C_4$  leaves.** Plants using  $C_4$  photosynthesis possess a number of anatomical, cellular and biochemical adaptations that distinguish them from  $C_3$  ancestors. These include decreased vein spacing (A) and enlarged bundle sheath (BS) cells, which lie adjacent to veins (B). Together, these adaptations decrease the ratio of mesophyll (M) to BS cell volume.  $C_4$  metabolism is generated by the increased abundance and M or BS-specific expression of multiple enzymes (shown in purple), which are expressed in both M and BS cells of  $C_3$  leaves. Abbreviations: ME - Malic enzymes, RuBisCO - Ribulose1-5,Bisphosphate Carboxylase Oxygenase, PEPC - phosphoenolpyruvate carboxylase, PPDK - pyruvate,orthophosphate dikinase.

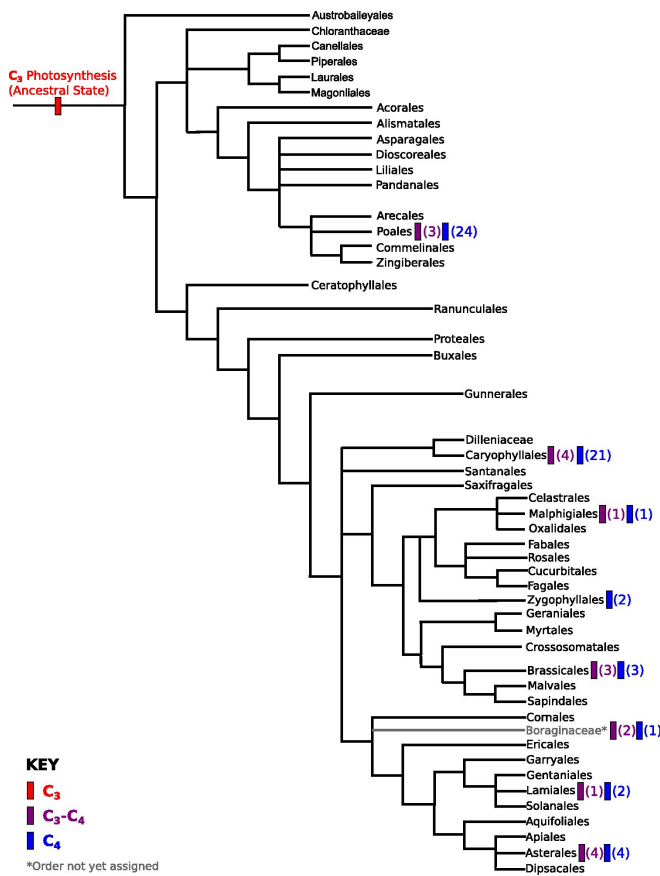


Figure 6: **Phylogenetic distribution of  $C_3$ - $C_4$  lineages included in this study.** A phylogeny of angiosperm orders is shown, based on the classification by the Angiosperm Phylogeny Group. The phylogenetic distribution of known two-celled  $C_4$  photosynthetic lineages are annotated, together with the distribution of  $C_3$ - $C_4$  lineages that we used in this study. The numbers of independent  $C_3$ - $C_4$  or  $C_4$  lineages present in each order are shown in parentheses.

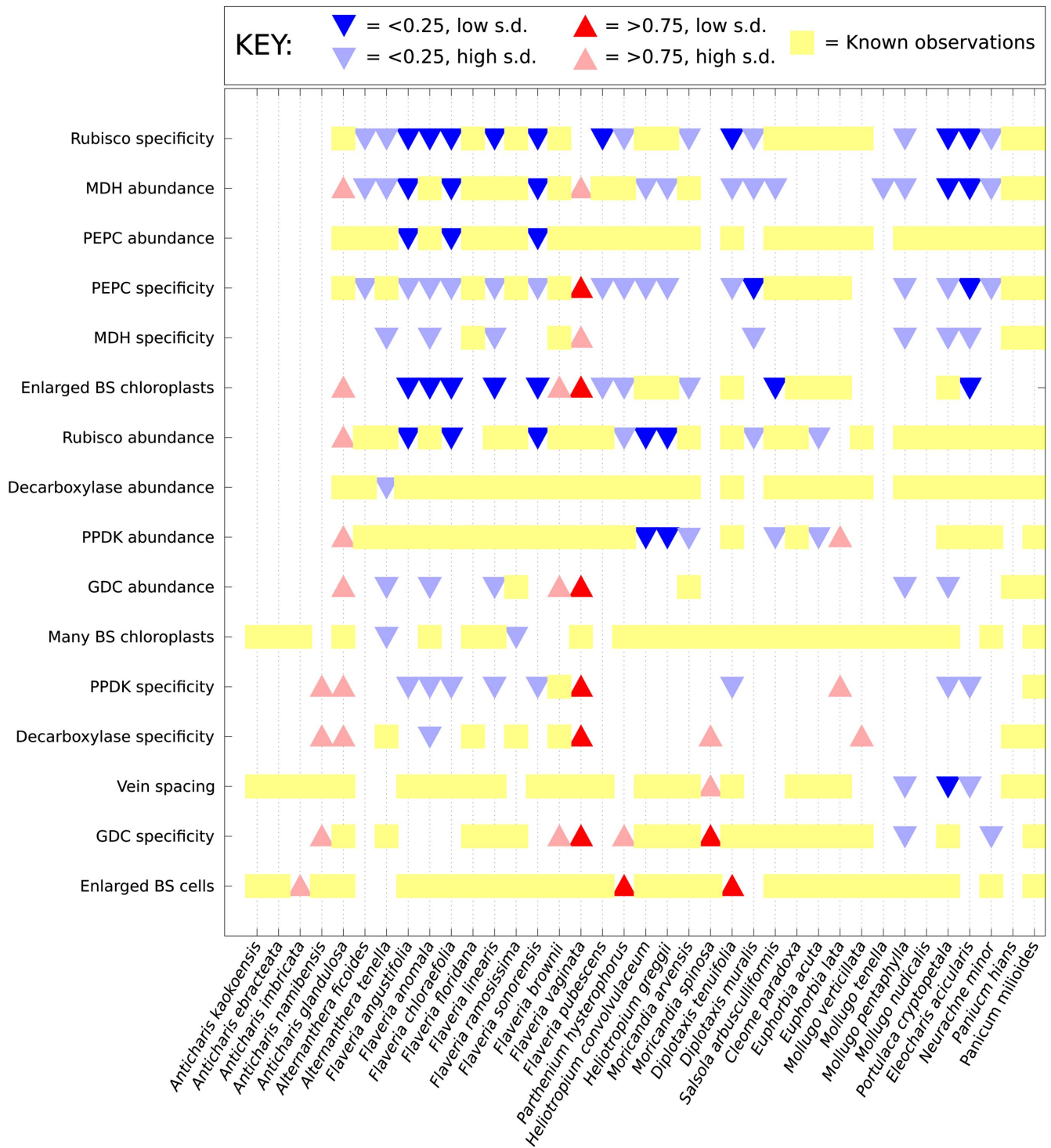
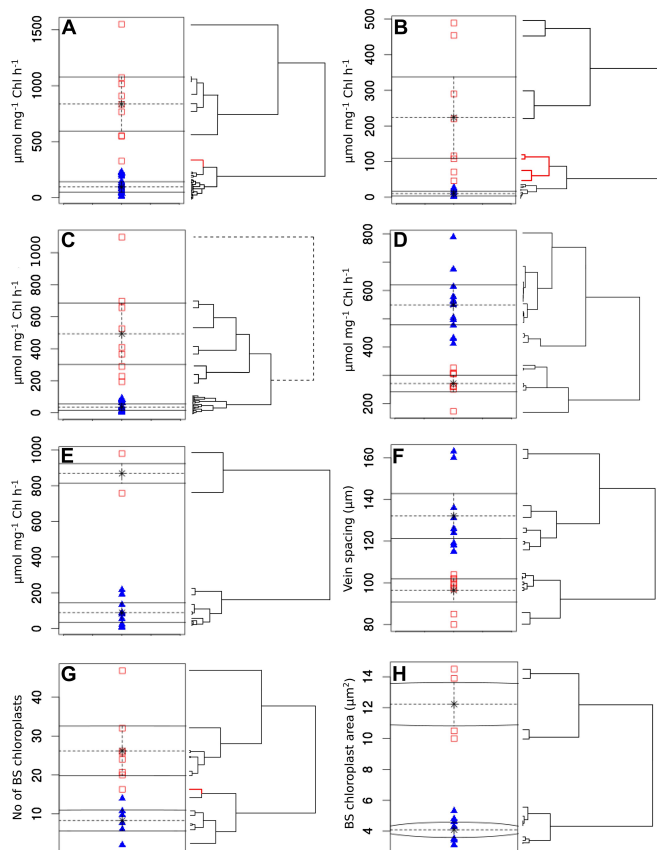
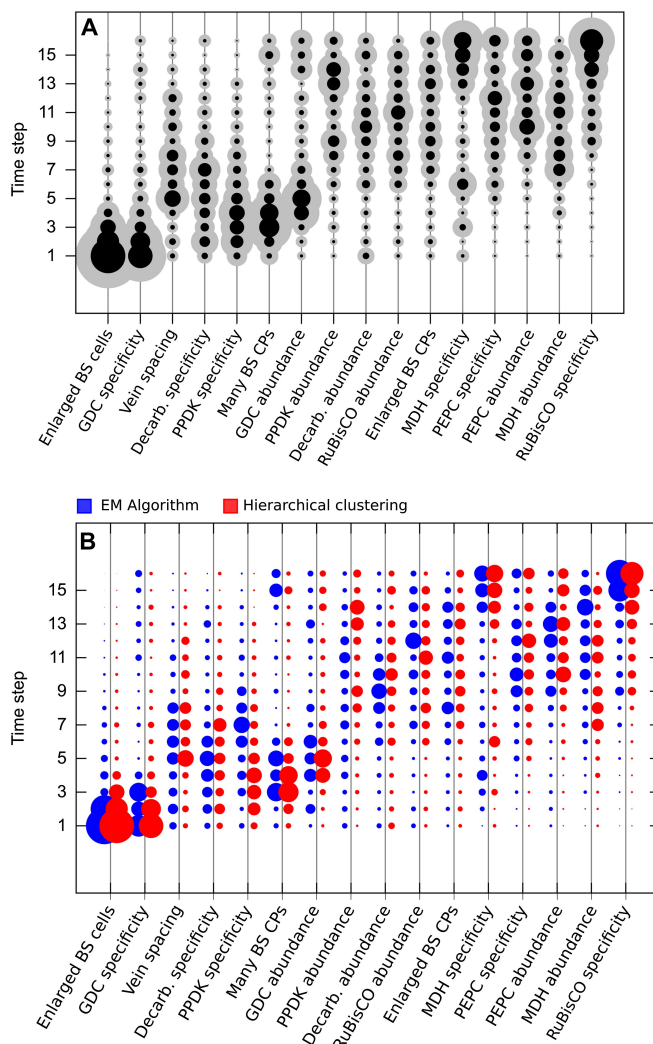


Figure 9: **Computational prediction of C<sub>3</sub>-C<sub>4</sub> intermediate phenotypes.** A probability for the presence of unobserved phenotypic characters was generated for every characteristic not yet studied in each of the C<sub>3</sub>-C<sub>4</sub> species included in this study. Red (upward triangles) predict a posterior mean probability of > 0.75 for the presence of a C<sub>4</sub> trait; blue (downward triangles) predict a posterior mean probability of < 0.25. Darker triangles represent probabilities whose standard deviations (s.d.) are lower than 0.25. Yellow blocks correspond to known data: no symbol is present for traits for which presence and absence have an equal probability (0.25 – 0.75).





**Figure 7: Clustering quantitative traits by EM algorithm and hierarchical clustering.** Quantitative variables were assigned binary scores using two-data clustering techniques. Each panel depicts the assignment of presence (red squares) and absence (blue triangles) scores by the EM algorithm. Adjacent to the right are cladograms depicting the partitioning of the same values into clusters by hierarchical clustering. Red cladogram branches denote values partitioned into a different group to that assigned by EM. The variables depicted in each panel are PEPC activity (A), PPK activity (B),  $C_4$  acid decarboxylase activity (C), RuBisCO activity (D), MDH activity (E), vein spacing (F), number of BS chloroplasts (G), BS chloroplast size (H).



**Figure 10: Results obtained using data clustered by hierarchical clustering.** Traits were also assigned presence/absence scores by hierarchical clustering. Analysis of data partitioned by hierarchical clustering predicted a similar sequence of evolutionary events to that shown in Figure 3 (A). Direct comparison of posterior probabilities reveals a high degree of similarity between results from the data clustered by hierarchical clustering versus the EM algorithm (B). These results suggest our conclusions are not affected by the different methods of assigning binary scores to traits.

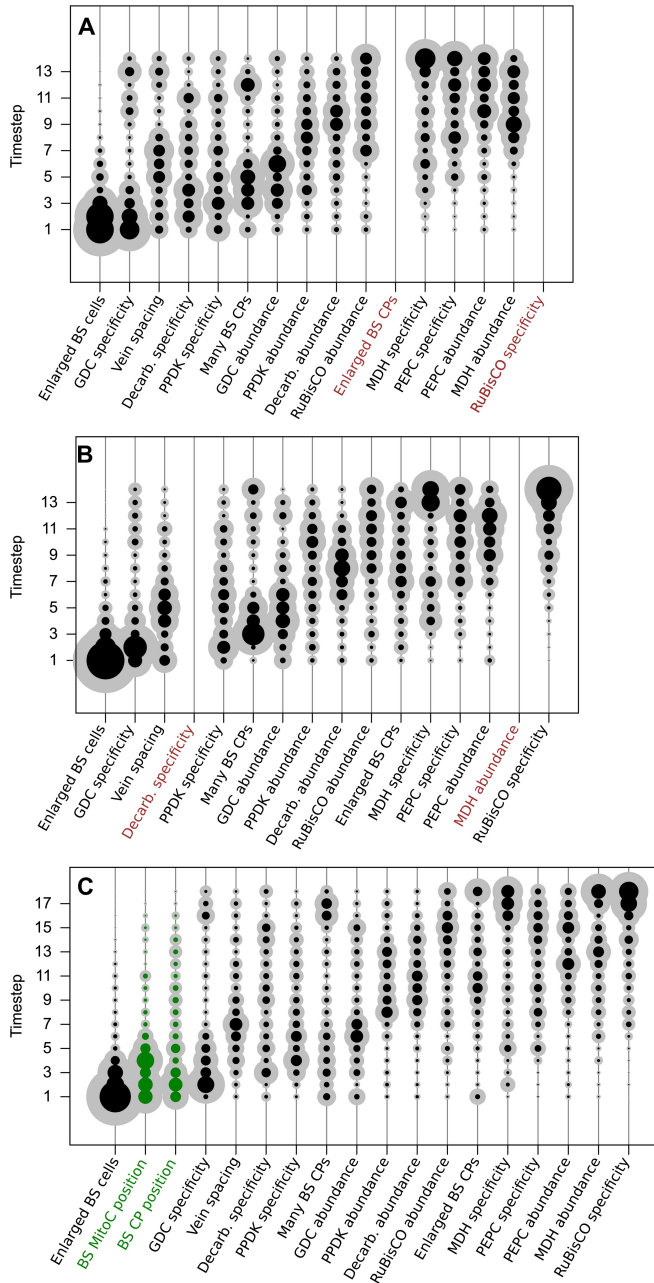


Figure 11: **Adding or removing traits does not affect the predicted order of evolutionary events.** Two independent pairs of traits were randomly selected and deleted from the analysis. In both cases, removing two traits did not affect the predicted timing of the remaining 14 traits in the analysis (A-B). Furthermore, including two additional traits associated with  $C_4$  photosynthesis also did not alter the predicted timing of other traits (C). Together, these data suggest our results are robust to both the removal and addition of traits from the phenotype space. Abbreviations: bundle sheath (BS), glycine decarboxylase (GDC), chloroplasts (CPs),  $C_4$  acid decarboxylase (Decarb.), mitochondria (MitoC) pyruvate,orthophosphate dikinase (PPDK), malate dehydrogenase (MDH), phosphoenolpyruvate carboxylase (PEPC).

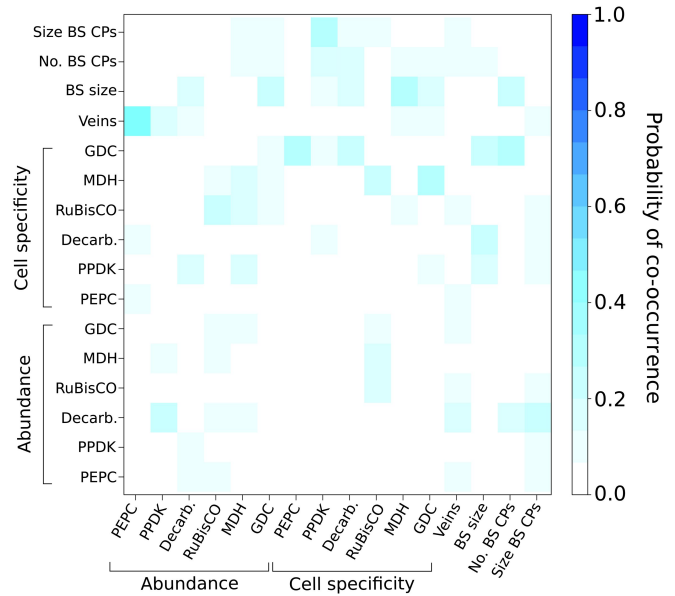


Figure 12: **Probabilities of  $C_4$  traits being acquired simultaneously.** The extent to which  $C_4$  traits are linked in evolution was assessed by modelling  $C_4$  evolution from a start phenotype with one trait already acquired. Linked traits would have a high probability of being acquired in the next event. Artificially acquired traits are listed on the x-axis and the probability of each additional  $C_4$  trait being subsequently acquired (y-axis) is denoted in each pixel of the heatmap. There is overall very low probability for multiple traits being linked in their acquisition in the evolution of  $C_4$ .

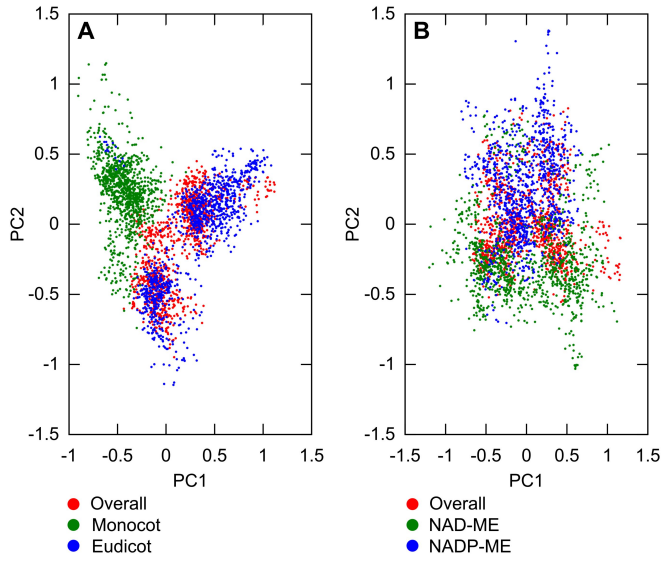


Figure 13: **Variation between lineages compared to variance of overall dataset.** PCA was performed on sampled transition networks from the sets compatible with the overall dataset and each of the two subsets corresponding to different lineages: overall/monocot/eudicot (A) overall/NAD-ME/NADP-ME (B). In (A) the variation between monocot and eudicot lineages is observed to be preserved when the overall transition networks are included, and on a similar quantitative scale to the variation in the overall set, embedded mainly on the first principal axis. In (B) the variation is of a similar scale but less distinct, correlating more with the second principal axis.