eLIFE
elifesciences.org

Figures and figure supplements

Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis

Ben P Williams, et al.

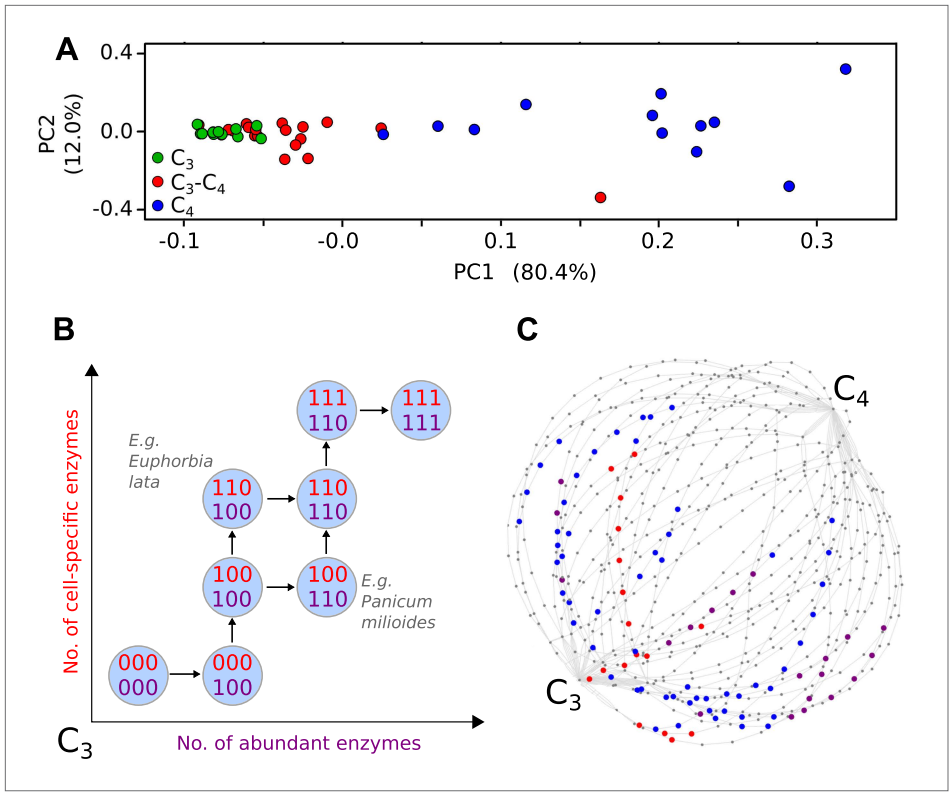

Figure 1. Evolutionary paths to C_4 phenotype space modelled from a meta-analysis of C_3–C_4 phenotypes. Principal component analysis (PCA) on data for the activity of five C_4 cycle enzymes confirms the intermediacy of C_3–C_4 species between C_3 and C_4 phenotype spaces (**A**). Each C_4 trait was 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**) 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).

DOI: 10.7554/eLife.00961.004

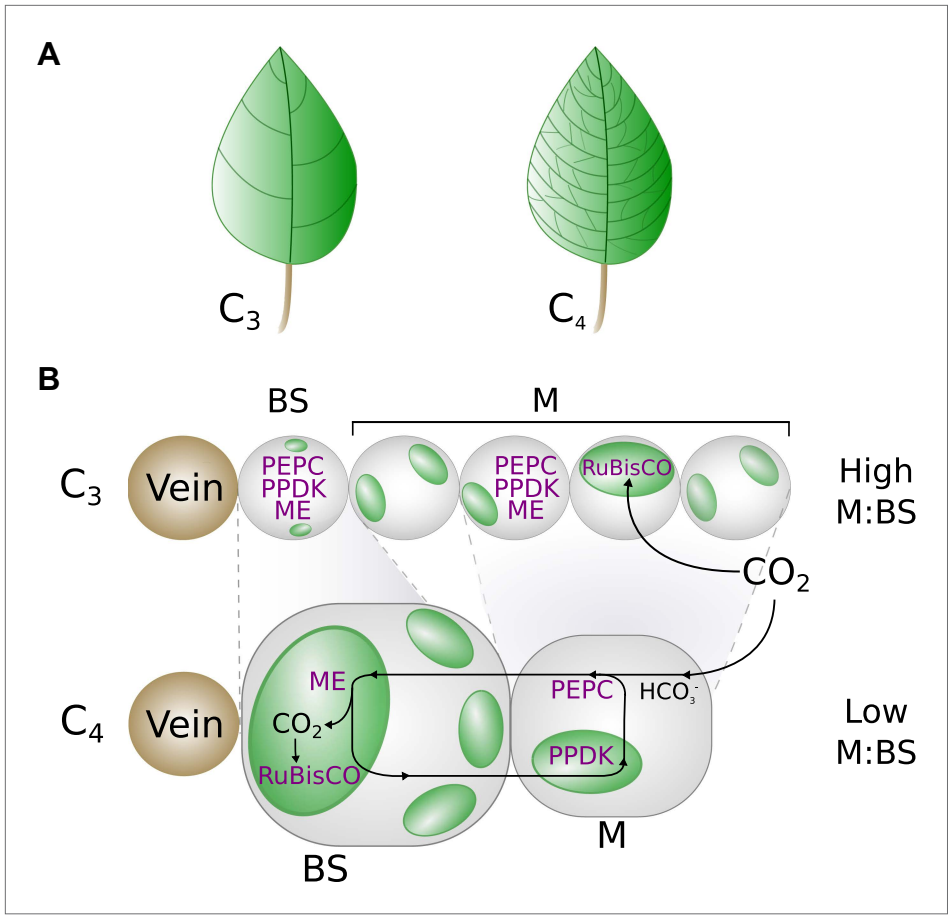

Figure 1—figure supplement 1. A graphical representation of key phenotypic changes distinguishing C_3 and C_4 leaves. Plants using C_4 photosynthesis possess a number of anatomical, cellular, and biochemical adaptations that distinguish them from C_3 ancestors. These include decreased vein spacing (**A**) 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–phospho*enol*pyruvate carboxylase, PPDK–pyruvate,orthophosphate dikinase.

DOI: 10.7554/eLife.00961.006

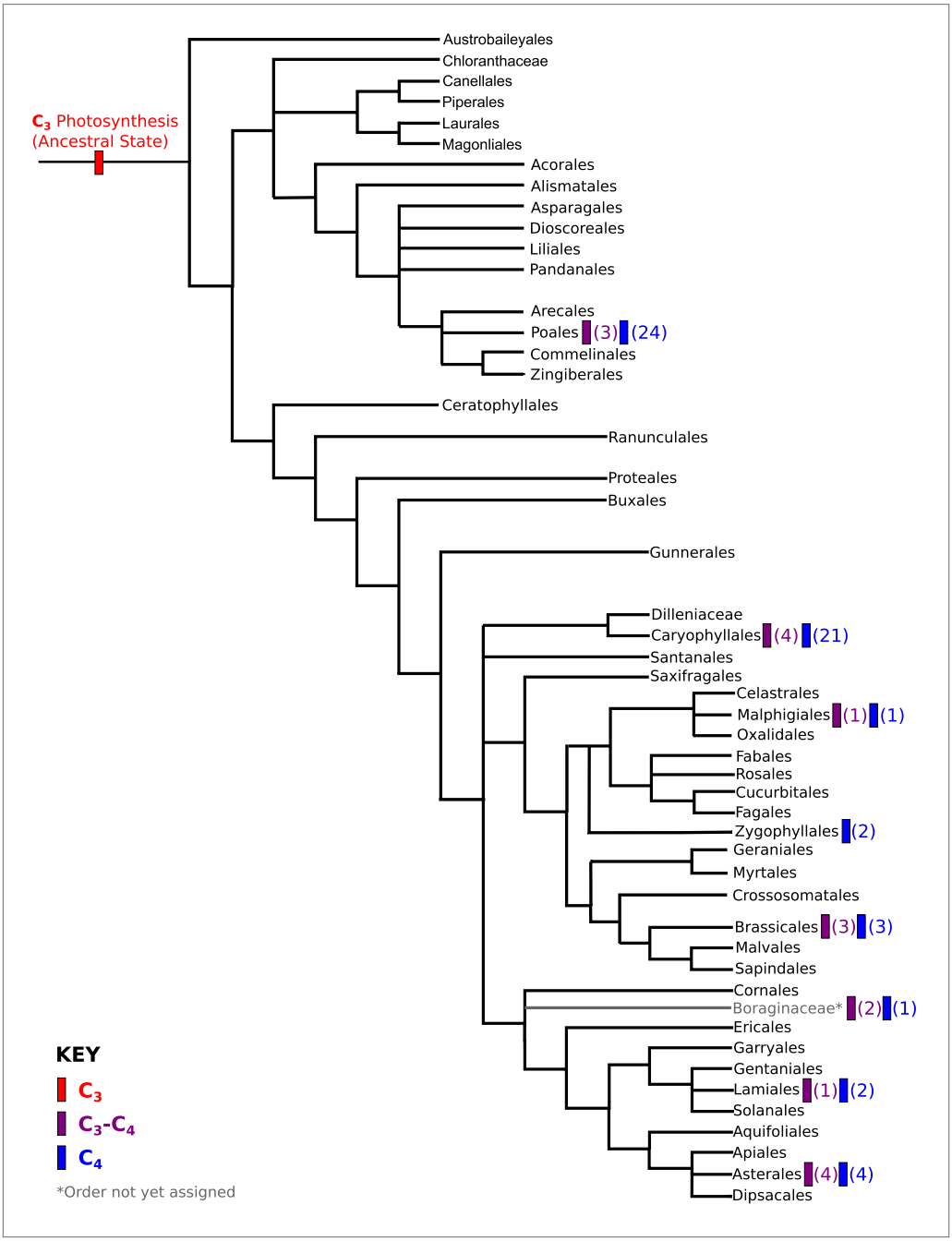

Figure 1—figure supplement 2. Phylogenetic distribution of C_4 and C_3–C_4 lineages across the angiosperm phylogeny. A phylogeny of angiosperm orders is shown, based on the classification by the Angiosperm Phylogeny Group. The phylogenetic distribution of known two-celled C_4 photosynthetic lineages 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.

DOI: 10.7554/eLife.00961.007

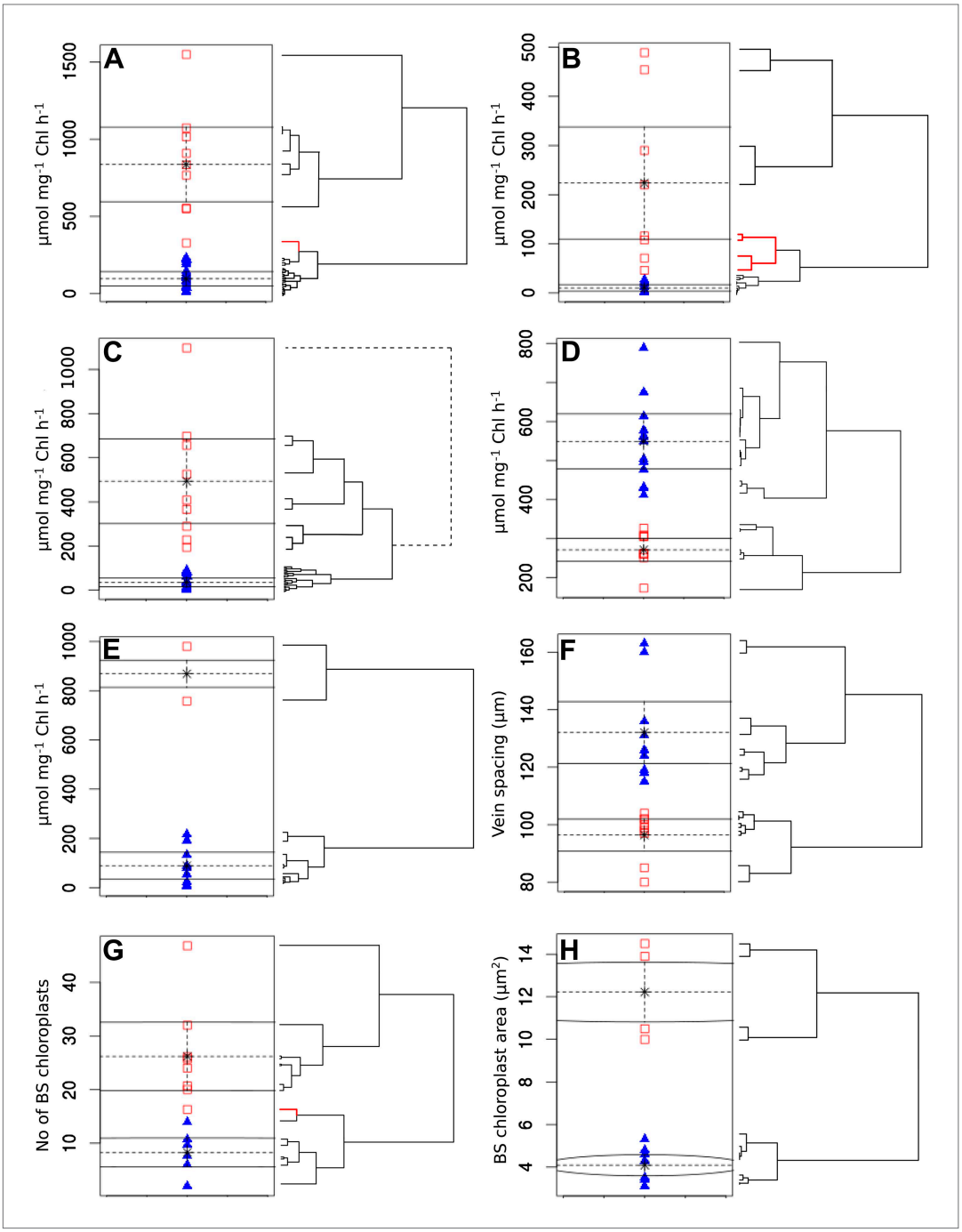

Figure 1—figure supplement 3. Clustering quantitative traits by EM algorithm and hierarchical clustering. Quantitative variables were assigned binary scores using two-data clustering techniques. Each panel depicts the assignation of presence (red squares) and absence (blue triangles) scores by the EM 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**), PPDK 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**).

DOI: 10.7554/eLife.00961.008

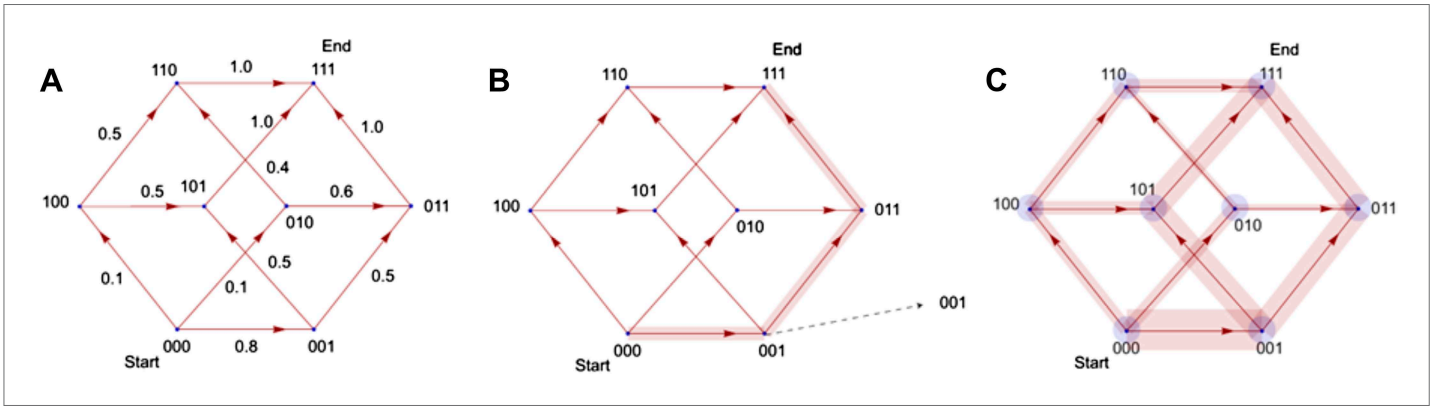

Figure 1—figure supplement 4. Illustration of the principle by which evolutionary pathways emit intermediate signals. In this illustration, the phenotype consists of three traits, yielding a simple (hyper)cubic transition network. Simulated trajectories on this network evolve according to the weights 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**).

DOI: 10.7554/eLife.00961.009

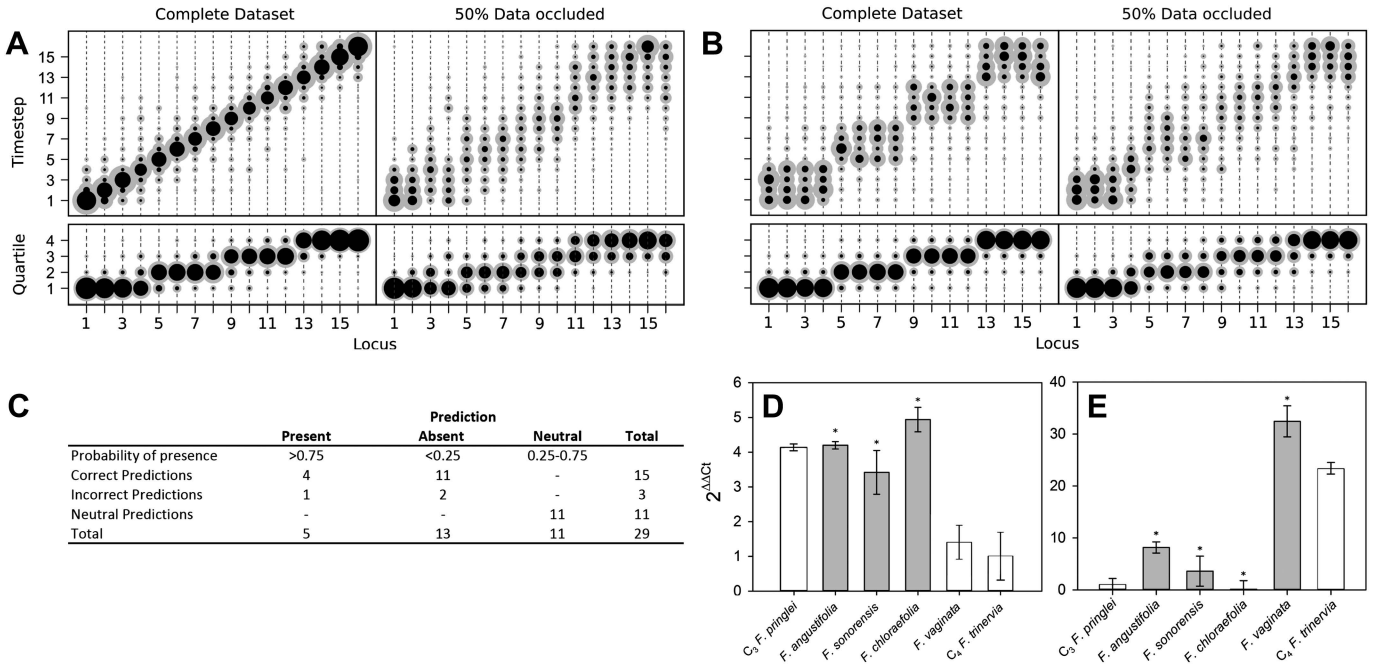

Figure 2. Verifying a novel Bayesian approach for predicting evolutionary trajectories. (**A** and **B**) Datasets were obtained from an artificially constructed diagonal dynamic matrix (**A**), and a diagonal matrix with linked timing of locus acquisitions (**B**). The single, diagonal evolutionary 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** and **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 were 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).

DOI: 10.7554/eLife.00961.010

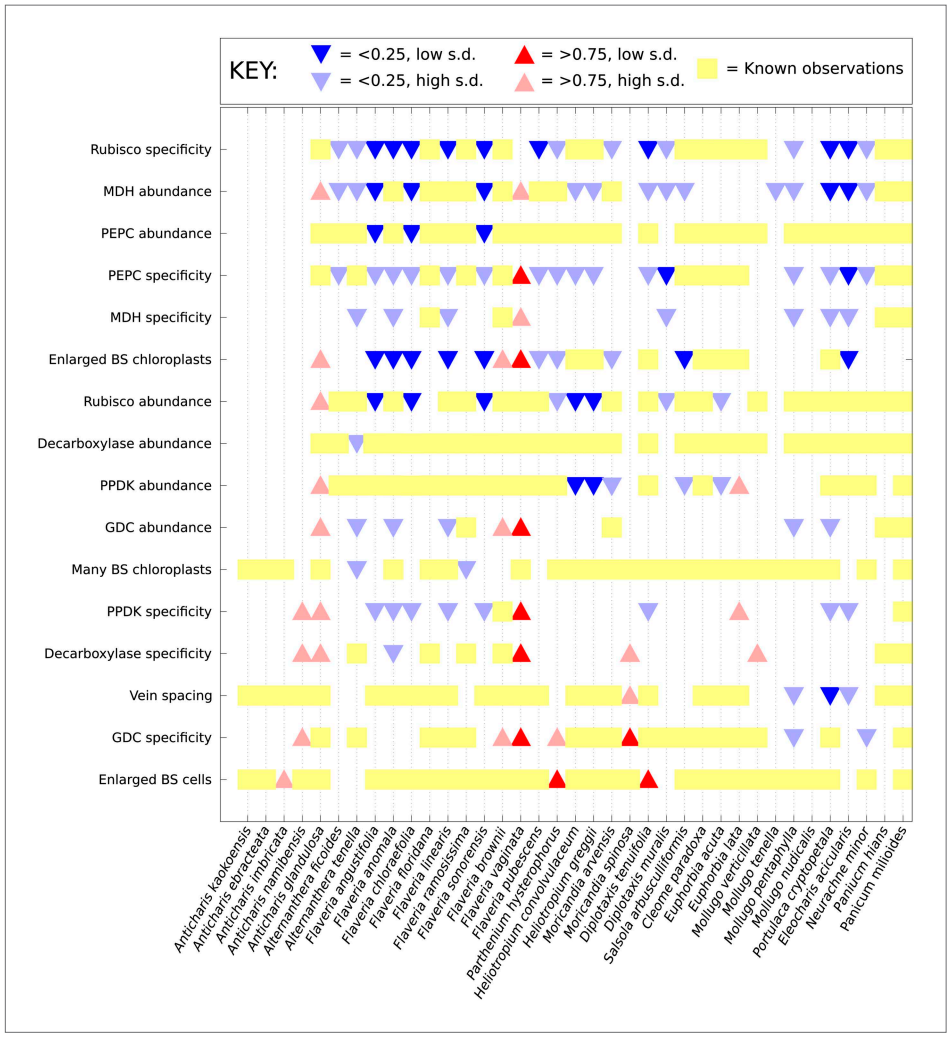

Figure 2—figure supplement 1. Computational prediction of C_3–C_4 intermediate phenotypes. A probability for the presence of unobserved phenotypic characters was generated for every characteristic not yet studied in each of the C_3–C_4 species included in this study. Red (upward triangles) predict a posterior mean probability of >0.75 for the presence of a C_4 trait; blue (downward triangles) predict a posterior mean probability of <0.25. Darker triangles represent probabilities whose standard deviations (SD) 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).

DOI: 10.7554/eLife.00961.011

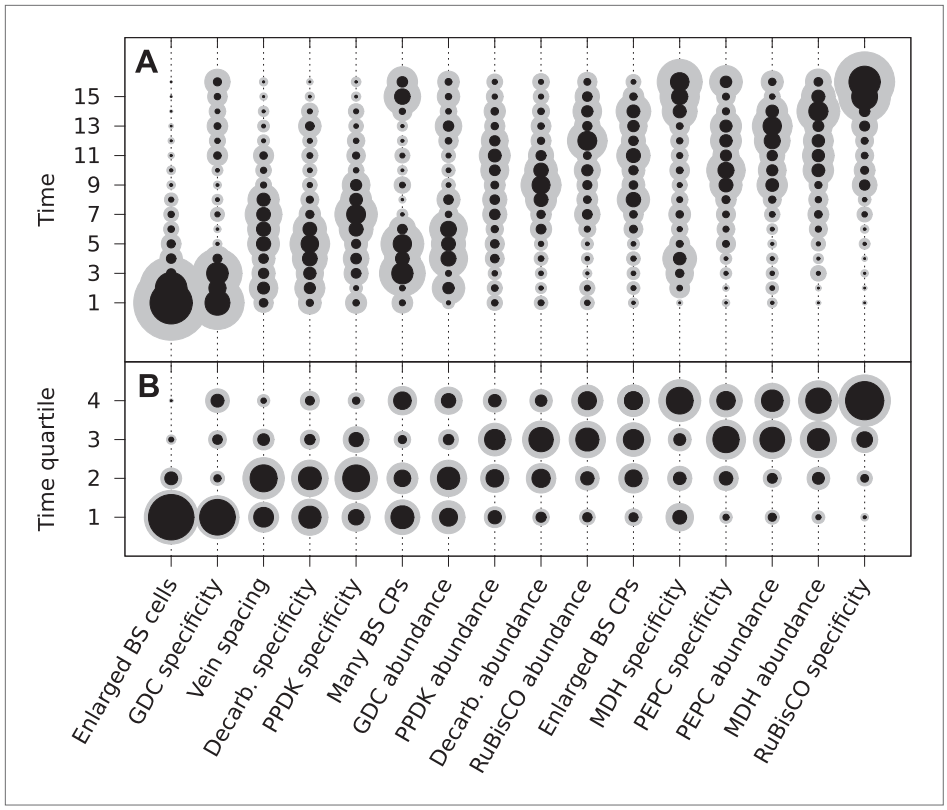

Figure 3. The mean ordering of phenotypic changes generating C_4 photosynthesis. EM-clustered data from C_3–C_4 intermediate species were used to generate posterior probability distributions for the timing of the acquisition of C_4 traits in sixteen evolutionary steps (**A**) or four quartiles (**B**). 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).

DOI: 10.7554/eLife.00961.012

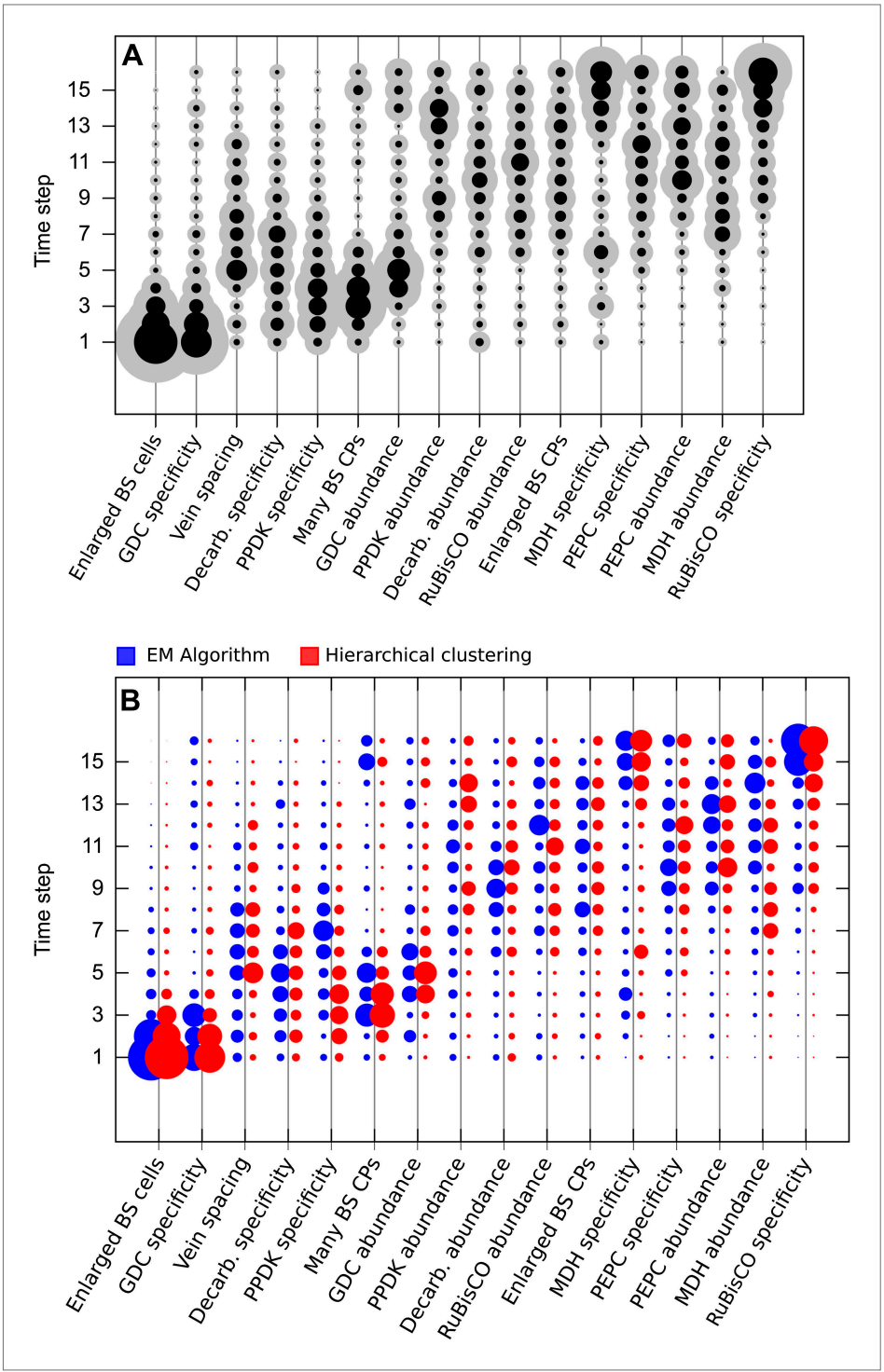

Figure 3—figure supplement 1. Results obtained using data clustered by hierarchical clustering. Traits were also assigned presence/absence scores by hierarchical clustering. Analysis of data partitioned by hierarchical clustering predicted a similar sequence of evolutionary events to that shown in ***Figure 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.

DOI: 10.7554/eLife.00961.013

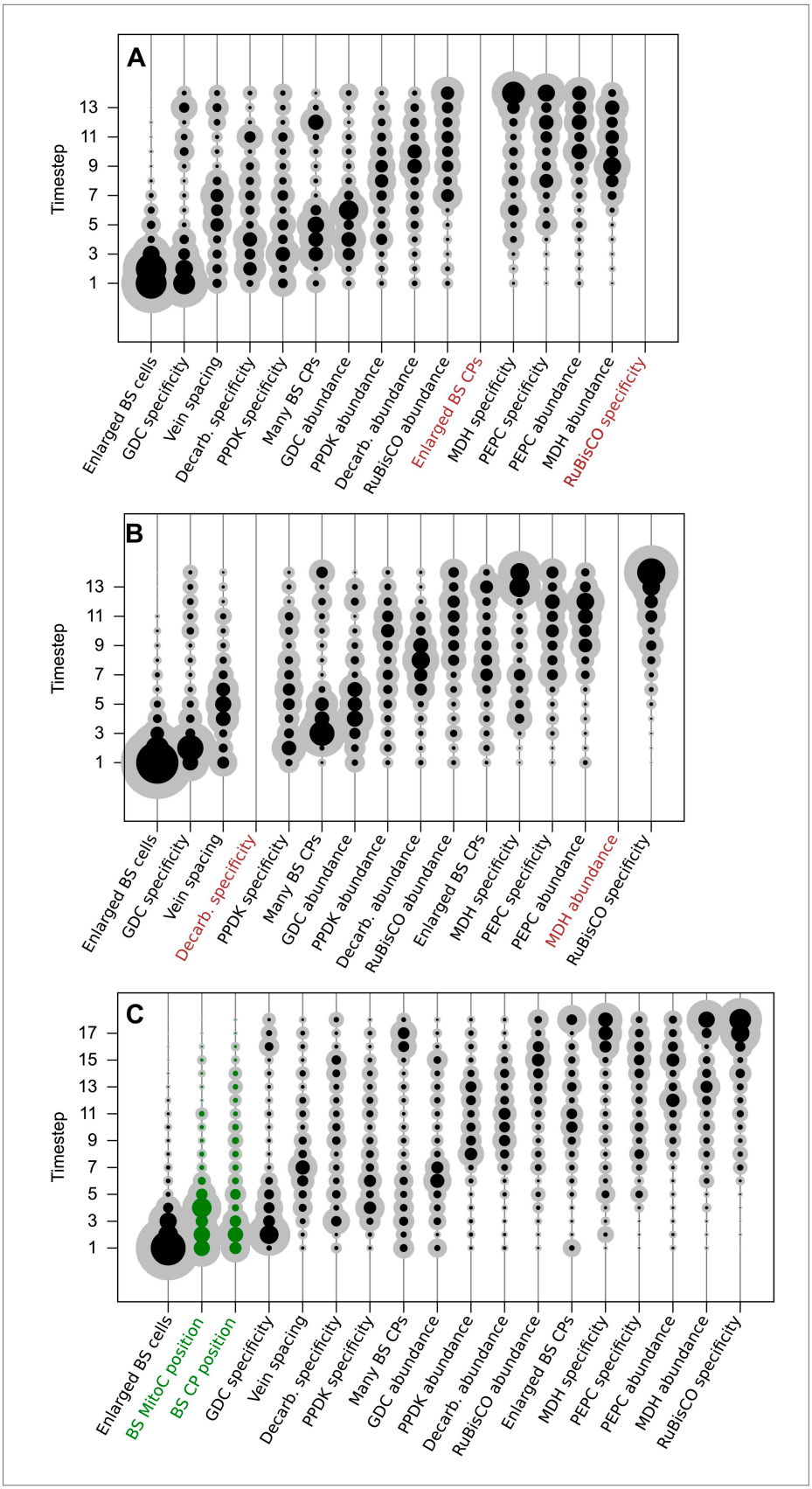

Figure 3—figure supplement 2. Adding or removing traits does not affect the predicted order of evolutionary events. Two independent pairs of traits were randomly selected and deleted from the analysis. In both cases,

Figure 3—figure supplement 2. Continued on next page

Figure 3—figure supplement 2. Continued

removing two traits did not affect the predicted timing of the remaining 14 traits in the analysis (**A** and **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).

DOI: 10.7554/eLife.00961.014

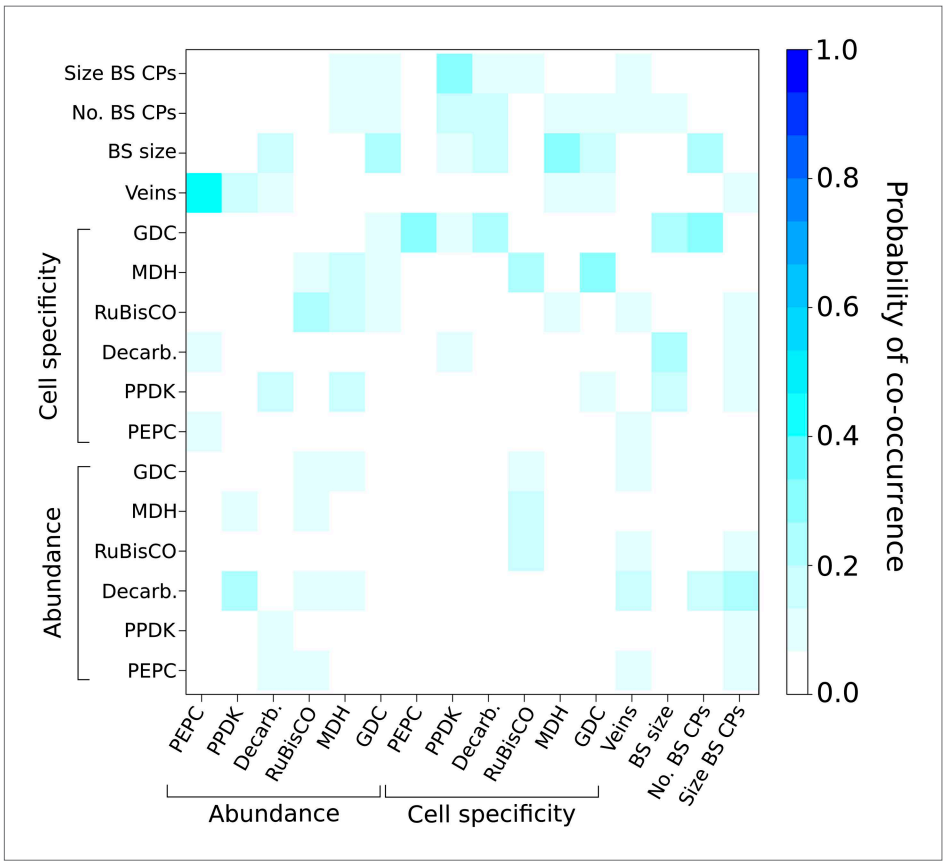

Figure 3—figure supplement 3. Probabilities of C_4 traits being acquired simultaneously. The extent to which C_4 traits are linked in evolution was assessed by modelling C_4 evolution from a start phenotype with one trait already acquired. Linked traits would have a high probability of being acquired 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 heat map. There is overall very low probability for multiple traits being linked in their acquisition in the evolution of C_4.

DOI: 10.7554/eLife.00961.015

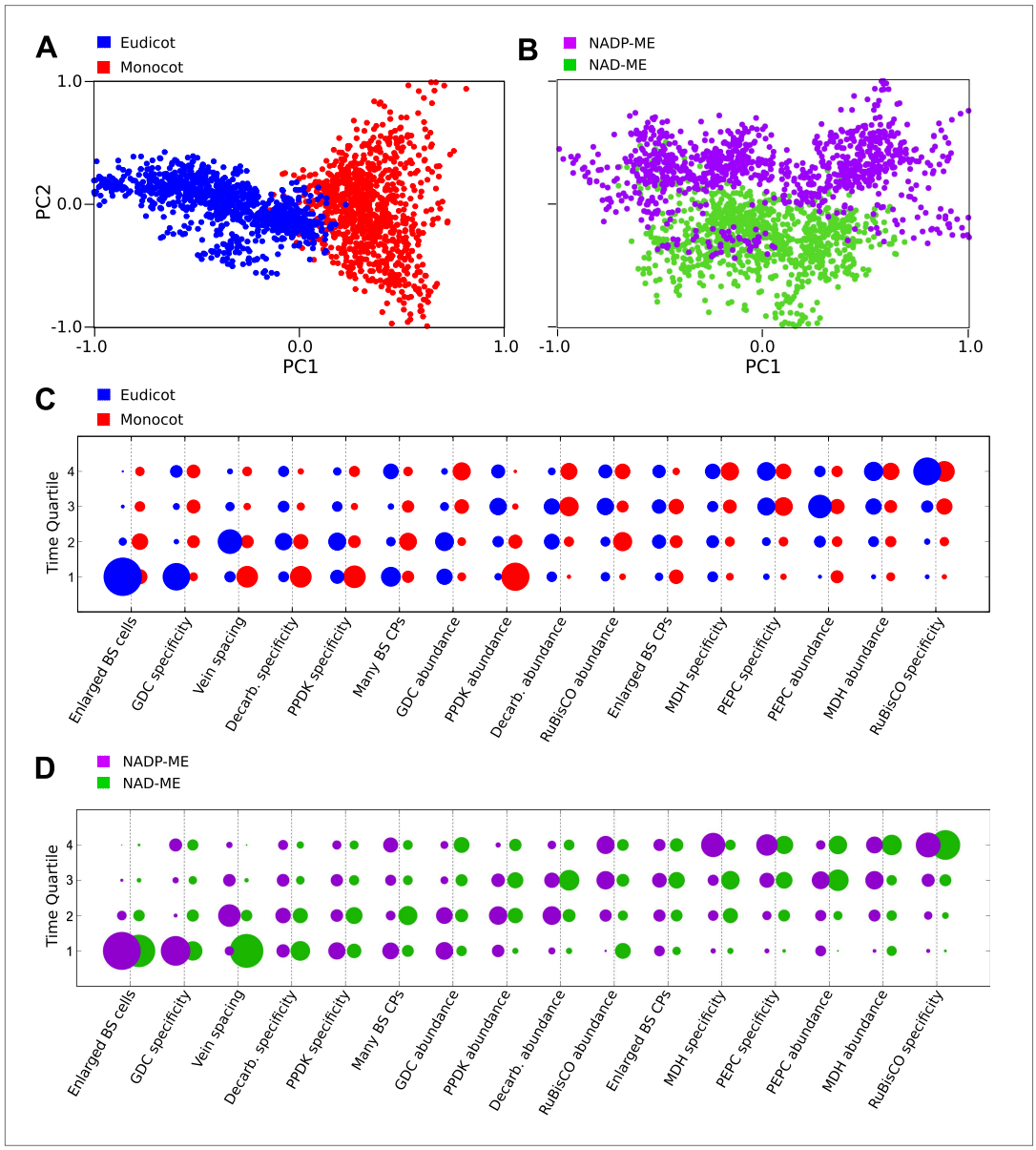

Figure 4. Differences in the evolutionary events generating different C_4 sub-types and distantly related taxa. Principal component analysis (PCA) on the entire landscape of transition probabilities using only monocot and eudicot data (**A**) and data from NADP-ME and NAD-ME sub-type lineages (**B**) 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**).

DOI: 10.7554/eLife.00961.016

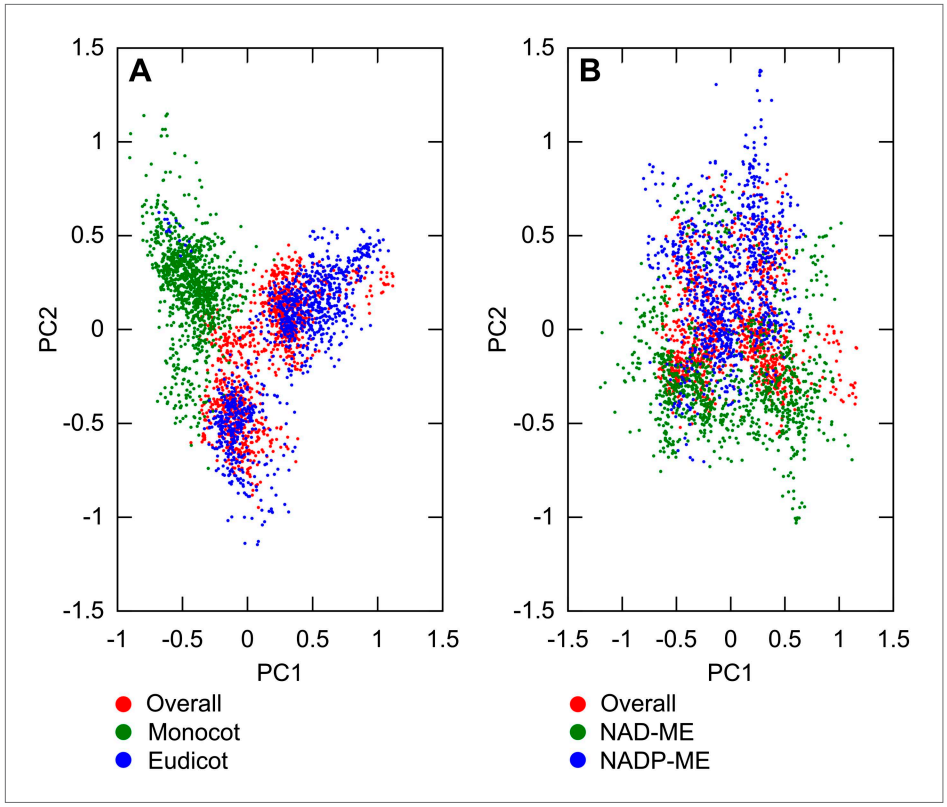

Figure 4—figure supplement 1. Variation between lineages compared to variance of overall dataset. PCA was performed on sampled transition networks from the sets compatible with the overall dataset and each of the two subsets corresponding to different lineages: overall/monocot/eudicot (**A**) overall/NAD-ME/NADP-ME (**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.

DOI: 10.7554/eLife.00961.017