The University Of Sheffield.

Molecular evolution of C_4 photosynthesis in grasses using comparative transcriptomics

By:

Jose J. Moreno-Villena

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Sciences
Department of Animal and Plant Sciences

December 2017.

Table of Contents

Acknowledgements

First of all I would like to thanks my supervisor Dr Pascal-Antoine Christin who has provided immeasurable support and encouragement during my PhD. His creativity and excitement have kept me motivated, but also his close supervision and hard work made this project possible. The same can be said for my co-supervisor Prof. Colin Osborne, who has often been a source of inspiration and made the best work experience. Both have significantly contributed to my personal and professional journey.

I have been a lucky PhD student and had all the support from many colleagues and I would like to thank all of them. Dr Luke Dunning shared with me many of his tricks and tips, and together we have made most of this project. He has always provided valuable comments and helped direct the course of work when needed. His input has increased the quality and quantity of my research. Similarly, Dr Marjorie Lundgren has taught me lots about photosynthesis itself, and shared with me her precious plant collection. Dr Jill Olofsson has helped clarify much of the genetic theory in my mind, and has helped resolve any methodological doubts. Dr Claudia Solis-lemus has also provided helpful comments and support from far, particularly for the phylonetworks part of my work.

A special thanks to my comrades Matheus Bianconi and Danny Woods. Together we learnt how to cope in a new, highly demanding environment, and did so with a lot of humour, fun and collaboration. I would also like to acknowledge Dr Emanuel Samaritani, Anne-Lise Liabot and Lamia Munshi for their work with the plants, and the rest of the Osborne lab group for sharing their cool research in our lab meetings, collaborating when needed, and for the beautiful Christmas walks.

The University of Sheffield, and particularly Animal and Plant Sciences is an amazing place to work. Researchers from all over the planet find their base here and the result is a vibrant and stimulating atmosphere. I would like to thanks the university staff, specially those at the Controlled Environment Facilities and the High Performance Computing resource. Also, I need to thanks the members of Dr Patrik Nosil's lab for the exchange of ideas, the Molecular Ecology Lab, and particularly Rachel Tucker for her excellent management of the research facilities and for making me feel at home.

An adventure always has a beginning, mine started at the University of Stirling, in

Scotland, where I worked in Dr Alistair Jump and Dr Mario-Vallejo's labs as part of the Erasmus Program (European Community Action Scheme for the Mobility of University Students). There, under the excellent guidance in molecular lab-work provided by Dr Jennifer Sjölund, I was initiated in research and they very much contributed to my career progression. To all of them, and the other members of their groups, a big and warm thanks.

My days in Sheffield and in the university have been flavoured by good colleagues and friends that provided much needed distraction and stimulation. Dr Bianca Santini showed me how to be a PhD student and improved my critical thinking. Dr Jason Griffiths, Ema Jardine and Bethan Hindle, who livened up many evenings. Pepa, Laura Gonzalez and Carmen made Sheffield colourful and sparkling, and the same goes for Ian Berriman. To all of them, thanks.

Finally, I would like to thank my friends and family at home for their constant warm support, specially to my sibling Ade, my grandparents, my brand new nephew and my mother and father, to whom this work is dedicated.

Abstract

The evolution of adaptations, some of impressive complexity, help organisms to survive in a variety of environments. However, evolutionary innovations are often restricted to certain taxonomic groups with evolutionary precursors. Evolutionary novelties usually arise through the co-option of pre-existing genes into new functions and by comparing organisms with and without the trait of interest, the ancestral state of the co-opted elements can be inferred. This allows us to identify the properties that facilitated their co-option, and therefore increase the evolvability of the complex trait itself. Such comparative studies can gain their power by comparing convergent phenotypes, which provide natural replicates. One example of a convergent complex trait is C_4 photosynthesis, an adaptation that results from the coordinated action of multiple enzymes to boost primary productivity in tropical conditions. It evolved more than 62 times in flowering plants, with at least 22 independent origins within the grass family alone. I studied the molecular evolution of coding genes and differential expression patterns in C_4 and non-C_4 grasses and my research showed: I) that ancestrally highly expressed genes copies had been preferentially co-opted for the C_4 pathway; II) that the emergence of a C_4 pathway in some taxa required the increased expression of just a few key C_4 genes. In addition, III) some C_4 genes were transferred across species boundaries and that IV) reticulate evolution punctuated the history of grasses, potentially promoting the spread of adaptive features. Together, these results lead to a new model of C_4 evolution, where some components are first accumulated for reasons unrelated to C_4, but serve as preadaptations that then allow the transition to a rudimentary C_4 pathway via few changes. This event creates strong selective pressure for improvements of the C_4 trait over long evolutionary times, and involves important re-programming of gene expression patterns, and impressive parallel adaptation of the translated proteins. Biochemical predisposition likely explains the recurrent C_4 origins, as seen in grasses. The recurrent transfer of C_4-adaptive loci across species might further have contributed to the observed bursts of C_4 lineages in some parts of the grass phylogeny.

General introduction

Origins of complex traits

During the history of life, organisms came to fill almost all existing environments on Earth, which in many cases necessitated specific characters to survive and prosper in particular conditions (Huston 1985; Terborgh 1992; Rosenzweig 1995; Norris 2009). For the past 150 years, research in evolutionary biology has tried to understand how species acquired these characters to then diversify, adapt or expand their ecological niche (e.g. Darwin 1859; Haldane 1915; Wright 1920; Fisher 1930; Dawkins 1976; Lenski 2017). It is widely accepted that novel characters are acquired via the action of natural selection, which leads to increases in frequency of those heritable modifications that confer an advantage, while preferentially removing those that decrease the life-time reproductive success, or fitness (Darwin & Wallce 1958). Over time, accumulation of distinct mutations in different lineages leads to different characters appearing in distinct species, and therefore the functional diversification of groups of organisms (Palumbi 1994; Dieckmann & Doebeli 1999; Rundle & Nosil 2005). These concepts of species diversification were initially formulated based on the observation of homologous morphological traits shared across species that helped biologists reconstruct the evolutionary steps between trait states linked to different environmental requirements (Darwin & Wallce 1958; Ruff et al. 1994; Zwieniecki & Newton 1995; Brischoux & Shine 2011). While not initially built in an evolutionary framework, the same concepts were implicitly used by taxonomy, which grouped those species sharing more similarities into hierarchical entities (Linnaeus 1735), and these groups were later recognized as those derived from a single ancestor in a branching genealogical tree (Darwin 1859; Cronquist 1968; Boudreaux 1979; Artyukhin 2006). The presence of different versions of equivalent functional traits across species helped Darwin and others after him to infer the gradual evolution of anatomical traits to reach new functions, which was the foundation of the field of evolution. However, many questions remain about the origins of adaptive traits of surprising complexity.

As Darwin proposed in his theory of evolution, the evolutionary recycling of traits is the main driver of adaptive innovation, and in practice, each extant organism is

a modified version of its ancestors (Gilbert 1986; Orgel 1994). In simple cases, for example a trait varying in length across species, the evolutionary history of the trait can be interpreted as the emergence of individuals exhibiting intermediate lengths of the trait, allowing subsequent changes in different directions. If certain lengths are adaptive in distinct environments, the emergence of different lengths in different groups can then be linked to divergent natural selection under different environmental pressures (Lovejoy 1988; Xiao 2014; Jogesh et al. 2016). As an example, the neck of the giraffe is longer than that of its relatives and provides access to leaves at the top of the trees to avoid competition with shorter herbivores (Cameron & du Toit 2006). It is likely that slightly longer necks appeared by chance, but because individuals bearing them could reach leaves higher in the trees, they were more likely to survive and passed their traits to more offspring, so that the frequency of longer necks increased. After repeating this process over numerous generations, impressively long necks had appeared in one species, but not in its relatives that experienced different selective pressures or a distinct set of random changes (Cameron & du Toit 2007). The origin of many traits and functions can similarly be explained by this process of continuous directional selection, such as the evolution of hearing in early terrestrial vertebrates from rudimentary auditive systems in aquatic ancestors (Christensen et al. 2015; Knight 2015). In contrast, when traits do not represent simple quantitative variation, understanding their origins is more challenging as intermediate evolutionary steps can remain elusive (Lange et al 2000). Indeed, some traits acquire their function only when multiple underlying components are modified. This is classically the case of the camera eye (Gehring & Ikeo 1999), the ability to fly (Heers et al. 2014), or complex biochemical cycles (Lange et al 2000). If a trait only gains its function when it is fully assembled and functional, how can natural selection gradually assemble the different components? Over the years, different scenarios have been proposed and supported by various lines of evidence. In the first scenario, successive modifications of each of the components can sequentially improve the function of the trait, providing directional selection toward the more complex version of the adaptation (Gillespie 1994; Ohta 1992). This is famously the case of the camera eye, which is thought to have emerged from light-sensitive cells via gradual addition of new components (Gehring & Ikeo 1999). Simple modifications led to different stages of complexity, which can still be observed in some organisms

(Nilsson 2013). Each of these successively more complicated stages improved light capture compared to the previous one, incrementally improving vision. Therefore, there exists a path composed of single steps that leads from organisms without any visual organ to species with their complex camera eye through stable stages that are successively advantageous. However, such paths do not necessarily exist, and some traits represent massive functional leaps compared to their ancestral stages as flowers (Albert et al. 2002) or the shell of turtles (Cebra-Thomas et al. 2005). In this second scenario, the underlying components are accumulated for unrelated reasons under a variety of selective pressures, and once enough of these components are present the complex phenotype can emerge (Kimura 1983; Barret & Schluter 2008). This is the case of the ability to fly, which in birds requires limbs, light bones, and feathers among others (Heers et al. 2014). All of these evolved independently, and for reasons unrelated to flight, but when combined enabled some primitive birds to glide, providing an entry point into a selectively driven evolutionary trajectory toward complex flight (discussed in Gauthier and Padian 1989; Heers et al. 2014). In such a case, feathers and light bones are traits that were used, or co-opted, to produce the flight apparatus. They therefore represent preadaptations or exaptations *sensu* Gould & Vrba (1982). In reality, the processes of continuous directional selection and exaptations are often combined, and the case of flight origins involved directional selection after multiple components were co-opted.

Establishing the evolutionary trajectories underlying the origins of complex traits is crucial to understand the forces that shape organism functional diversification and how this allows species to cope with changing environments. In many cases however, the order of the successive changes toward complex traits and the stage at which an emerging trait provides a selective advantage remain unknown. Ancestors could be investigated to study each step in detail. This can be achieved using a few microorganisms that can be rapidly propagated in laboratory conditions, in what is called experimental evolution (Cooper et al. 2003; Lenski 2017). However, for larger organisms, evolutionary processes can only be inferred based on the comparative study of extant species and the fossil record, using historical approaches (Albert et al. 2002; Cebra-Thomas et al. 2005; Heers et al. 2014). In addition to comparing anatomical and biochemical traits, it is now possible to study in depth DNA, which encodes instructions

for each component of any trait. Keys to the origins of novel adaptations lie in the genomes, shaped by heritable random changes in DNA, forming the substrate of natural selection. These can be deciphered via genome analyses, which have revealed complex evolutionary processes but also opened new avenues for inferring the evolutionary origins of novel functions.

Molecular origins of novel adaptations

From the origins of the theory of evolution via natural selection, a key point is that changes have to be heritable to allow for natural selection (Darwin 1859). Hence only the variation arising from modifications of the heritable material, the DNA or genome, will play a role in the origin of new functions. The genome carries instructions to produce a large variety of transcripts, many of which can then be translated into proteins (Raven et al. 2006). The amount of transcripts per gene varies among genes and through space within the organism (i.e. among organs and cells) and through time (i.e. during the development and once the organism is adult). In addition, the amount of translation of transcripts into proteins varies, providing an extra filter on the transcripts and proteins that will be present in every cell (Orphanides 2002; Nolis et al. 2009). The information about transcription and translation are also contained in the genome and can be influenced by the environment (Tobin & Suttie 1980; Tobin & Silverthorne 1985; Gallie, 1993; Floris et al. 2009). The genome therefore contains the code to produce the phenotype of the organism, taking into account the environment.

The genome is duplicated during each cell division, and organismal reproduction produces a new copy of the genome, with a possible reshuffling in the case of sexual reproduction. Because the copying process is not perfect, errors will be inserted at each step, which will introduce variation in the populations (Duret & Mouchiroud 2000). Regardless of the tolerance to genetic variation that biological systems exhibit (discussed in Fares 2015) each of these changes, or mutations, will be tested by natural selection and the fate of mutations over generations will depend on their effect on the phenotype and in particular the reproductive success of individual bearing them (Gillespie 1994; Perfeito 2007; Barrick et al. 2009). Mutations in regulatory sequences will affect when and where a gene is expressed, potentially determining tissue specialization and the rhythm of developmental and metabolic

processes (Wray 2007). If the mutation happens within a coding sequence, which is transcribed into messenger RNA and translated into amino acids composing proteins by the ribosomes, it might affect the biochemical properties of the protein and therefore its functional characteristics (Bowie et al. 1990; DePristo et al. 2005; Tokuriki and Tawfik 2009). Because the protein sequence undergoes postranslational regulation, mutations can also affect the rate of accumulation of the protein and impacting largely in the phenotype, which can contribute to changes in cell and tissues structures, homoeostasis or coloration (Scroggins & Neckers 2007; Pennuto et al 2009; Janke & Bulinski 2011). Natural selection will preferentially remove those mutations that decrease survival and/or reproduction and keep the more beneficial ones (Muller 1932; Gerrish & Lenski 1998; Perfeito et al. 2007). At the extreme of the spectrum of potential effects, deadly mutations will never be passed on to future generations. The fate of neutral mutations will depend on the population dynamics, and specific conditions (e.g. small population sizes) can lead to their fixation by chance, a process known as genetic drift (Remold & Lenski 2001; Barrick et al. 2009). Finally, the few mutations that will provide an advantage will generally increase in frequency as individuals bearing them will tend to produce more descendants (Perfeito et al. 2007) if they escape genetic drift. Over long evolutionary times, the accumulation of mutations in regulatory regions and those encoding proteins is responsible for the functional diversification of organisms, with natural selection acting as a filter that can be powerful and responsible for the emergence of adaptive traits.

Different types of errors, or mutations, can occur during the process of replication of the genome. For instance, point mutations are substitutions of one nucleotide for another. If occurring in regulatory regions, these can affect the DNA-binding specificity of transcription factors, and therefore the patterns of gene expression (Sayou et al. 2014; for a review of the topic see Gregory 2007 and Jarvela & Hinman 2015). Among substitutions occurring within protein-coding regions, some will change the encoded amino acids. This can drastically change the properties of the protein, including its activity, stability, and interactions with cofactors (Daugaard et al. 2007). In addition, some substitutions will introduce stop codons, rendering the protein non-functional. Other mutations will introduce or remove some bases from the nucleotide sequences. These insertions/deletions, known as indels, can alter the regulatory

sequences, but will have the strongest effect within coding sequences. Indeed, indels are likely to alter the reading frame, leading to completely different, and in most cases non functional, proteins (Hu & Ng 2012). While indels represent short losses or gains, large regions of the genomes can be lost or duplicated, resulting in losses of entire genes or gains of duplicates, which can be important for long term evolutionary dynamics (see below). Finally, mutations include rearrangements that move a DNA fragment to a different position in the genome (Bennetzen 2000). These can create chimeras between genes, but are more likely to alter the expression patterns by placing the gene in a different chromosomal context (Coen et al 1986; van de Lagemaat 2003) (see Pal & Papp 2017 for a comprehensive discussion about the role of mutations in the evolution of complex traits). It is widely accepted that beneficial mutations are rare. In the case of amino acid substitutions, it is estimated that 70% are detrimental in *Drosophila*, and the rest neutral, with only a few slightly beneficial (Sawyer et al. 2007; reviewed in Barrick & Lenski 2013). Therefore, most new mutations will quickly disappear from the populations. The frequency of mutations during reproduction and their rate of fixation are consequently disconnected. The link between the two will moreover vary among phases of population dynamics, with periods of adaptive evolution seeing an overaccumulation of advantageous mutations (Desai & Fisher 2007).

Over large evolutionary scales, the creation of new genetic material is of the utmost importance for the origins of novel adaptations. Some genes can appear *ex nihilo* in a genome, but most protein-coding genes evolved a long time ago in bacteria fuelling early evolution (Duboule & Wilkins 1998; True & Carroll 2002; Bergthorsson et al. 2007; Kaessmann 2010; Carvunis et al. 2012). These genes were continuously copied under a birth-death dynamics where genes get duplicated and duplicates are lost (Brunet et al. 2006; Kondrashov & Kondrashov 2006; Katju & Bergthorsoon 2013). These duplications can happen during whole genome duplications, which occurred episodically during the history of life (Dehal et al 2005; Aury et al 2006; Tank et al. 2015; Soltis et al. 2015). Alternatively, single chromosomal regions or single genes can be duplicated, via illegitimate recombination, the action of transposable elements, or the reinsertion of messenger RNA (retrotransposition) (Zhang 2003). Independently of the mechanisms, gene duplications are seen as important processes for phenotypic diversification (reviewed in Kaessmann 2010). On the short term, the presence of

duplicates can lead to dosage effects, and duplications placing one of the copies in a different genomic location can have drastic effects on the expression patterns (Force et al. 1999; He & Zhang 2005). On the long term, duplications are important for the evolution of new functions. Indeed, the presence of two copies leads to functional redundancy, so that one of the copies can acquire new properties while the other maintains the ancestral function (Zhang 2003). This process is termed neo-functionalization, and has been described as key to functional diversification of organisms (Ohno 1970). Alternatively, the two copies can each maintain parts of the ancestral function, in a process of subfunctionalization (Lynch & Force 2000). In both cases, if retained, the two copies will accumulate different mutations, and diverge under either neutral or adaptive evolution. After divergence time, this will create a pool of genes encoding similar proteins, but with differences in expression patterns and/or catalytic properties of the encoded enzymes. Fascinating examples of increases in copies of genes associated with traits under selection include olfatory receptors in honeybees, visual proteins in dragonflies and heat-shock response proteins in intertidal oysters (reviewed in Holland et al. 2017) or floral pigmentation in plants (Brockington et al. 2015). Gene duplication then has been widely linked to the origin of specific complex traits, but the mechanisms underlying this process remains debated.

Accessibility to novel phenotypes

The evolutionary origins of a given trait cannot be understood without considering the ancestral state, as the number of changes needed to access a new phenotype will vary among groups and will determine their accessibility to the trait (Marazzi et al 2012; Edwards & Donoghue 2013). Indeed, most traits evolved by recycling structures and genes that already exist, so that the properties of an organism will affect its capacity or accessibility to evolutionary innovations, sometimes defined as evolvability (Wagner & Altenberg 1996; Kirschner & Gerhart 1998; Edwards & Donoghue 2013). Globally, the adaptive potential of populations defines their capacity to respond to selective pressures, and depends on the effective population sizes, mutation rates, and generation times, among others (Wagner & Altenberg 1996). From a genomic point of view, the adaptive potential is likely to be influenced by the number of gene copies (Lenormand et al. 1998; Flagel & Wendel 2009; Tank et al. 2015; Soltis et al. 2015; Wendel 2015) and

genome dynamics (Cooper et al. 2003; Crozat et al. 2005), with, for example, whole genome duplications seen as promoting functional diversification (Tank et al. 2015; Soltis et al. 2015). However, when a specific phenotype is considered, the evolutionary distance between the ancestral and derived states is likely to be important, whether considered at the phenotypic or genomic levels (Edwards & Donoghue 2013).

The likelihood to develop specific adaptations depends on the ecology, as a complex trait requiring multiple changes will not emerge in conditions where it is not advantageous. In addition, the potential of adaptation, or the degree of similarity between the ancestral and derived phenotypes, is likely to ease or prevent the evolution of certain functions (Marazzi et al. 2012). As an obvious example, light, feathered dinosaurs were more likely to evolve the ability to fly than large-bodied animals or those lacking limbs. The features present in some groups that evolved for a different reason but increase the accessibility to given phenotypes are commonly referred to as preconditions or pre-adaptations that act as evolutionary facilitators (Bock 1959). For instance, forelimbs enabled the evolution of gliding and/or flying in diverse groups of tetrapods, hence the mere presence of forelimbs facilitated the acquisition of the new functions. While the discussion initially focused on anatomical precursors, genetic elements can act as cryptic facilitators, if they allow the emergence of a new phenotype via a few changes (Hayden et al. 2011). For example, in a long-term experiment with *E. coli* growing on different substrates, a mutation without apparent effect was shown to be necessary for the later origin of the adaptive ability to digest different food sources (Blount et al. 2008; 2012) and that presumably increases the likelihood to evolve the function. The impact of genetic precursors on evolutionary trajectories is, however, still poorly understood, mainly because the information is difficult to access and, consequently, model.

From a theoretical viewpoint, evolutionary trajectories have often been described using the allegory of the adaptive landscape, in which the fitness depends on the state of multiple variables, creating a multi-dimensional landscape. This concept was first conceived by Wright, who plotted the different versions of a gene along a fitness gradient in a two dimensions map (Wright 1932). This idea has been largely expanded and modified through the years by adding numerous dimensions, including additional genes, protein sequences, morphological traits, and landscape modifications

(reviewed in Gerlee 2015). In all cases, populations can move across the adaptive landscape in small steps, and those increasing the fitness are more likely to be fixed in the population (Blount et al. 2008). Over time, drift and natural selection can lead populations to adaptive peaks. However, depending on the shape of the landscapes, some peaks might be out of reach of natural selection, because they require intermediate stages that are highly deleterious. In the context of complex trait evolution, the adaptive landscape can illustrate the possible evolutionary trajectories leading from a given ancestral state to the novel complex phenotype. The adaptive peaks that can be reached will depend on the starting point, so that unrelated changes that moved the populations in different parts of the landscape prior to the emergence of the complex trait affect its accessibility. In addition, the adaptive landscape will vary with environmental conditions, as well as with the current stage of the population. For instance, traits that were not advantageous *per se* will become highly beneficial once others have been acquired. A given complex trait can therefore evolve under natural selection acting on random mutations only if there exists a path in the condition-dependant adaptive landscape that connects successively advantageous or neutral intermediate stages. If fewer steps are required and those steps are beneficial, the emergence of the complex trait will be more likely (Blount et al. 2008).

While adaptive landscapes are great conceptual tools, there are only specific cases where they can actually be produced. First, simulations or other models can directly produce such landscapes, with the associated evolutionary trajectories (Bornberg-Bauer & Chan 1999). As with any model however, establishing whether the conclusions apply equally to real-life examples cannot be known with certainty. Second, in specific cases the fitness of different states can be measured and compared. For instance, different combinations of mutations along a protein sequence can be produced, and the properties of the protein can be assessed (e.g. Weinreich et al. 2006, Stiffler et al. 2015). In the case of short-lived model organisms, repeated experimental evolution trials can generate evolutionary trajectories, and directly compare the fitness of ancestral and extant populations (de Visser & Lenski 2002). However, experimental adaptive landscapes can only be produced for specific, simple study systems. For most other study systems, a comparative approach represents the best strategy, and phylogenetic

analyses can reconstruct past evolutionary trajectories, which in the case of convergent traits can produce a proxy of an adaptive landscape.

Phylogenetic approaches to evolution

Phylogenetic trees depict the relationships among groups of organisms (e.g. Hug et al 2016; Letunic & Bork 2016). In a phylogenetic tree, the tips represent individuals (often species), nodes represent divergence of two lineages, while branches indicate the amount of change that happened in between distinct divergence events. The phylogenetic relationships can be inferred using different methods that compare homologous traits across the individuals under study. Historically, these relationships were inferred by comparing morphological traits that were easily assessed, a method known as phenetics, grouping those species that are similar in a way that minimizes the amount of change inferred across the whole phylogenetic tree, under a maximum parsimony criterion (e.g. Agnarsson 2004). This was later progressively replaced by maximum likelihood, and later Bayesian methods, which identify the phylogenetic tree that best explains the distribution of characters among the tips. Relying on morphological characters, however, created problems because the same state can arise in unrelated species by chance (*i.e.* convergent evolution) (Wake 1991; Wiens et al. 2003). Therefore today these methods have been replaced by molecular phylogenies. Following the emergence of sequencing methods, protein sequences and later DNA sequences became available as data. While the problem of convergent evolution still exists, DNA sequences offer a large number of sites, each of which represents an independent character, so that the signal of convergent sites will be masked by all others. In addition, homology among DNA or protein sites can be easily established, and suitable mechanistic models of their evolutionary dynamics exist (Schraiber & Akey 2015). Molecular phylogenetics has consequently become the gold standard in the field. The sequencing process was previously laborious, so that phylogenetic studies were mainly limited to small number of markers, and the organellar genomes were primarily targeted, as these are easier to isolate and extract. With the recent advent of high-throughput sequencing and its increasing accessibility and reliability (Koboldt et al. 2013; van Dijk et al. 2014), large genomic datasets became available allowing the inference of phylogenetic trees using numerous markers spread across both the

organellar and nuclear genomes (e.g. Zeng et al 2014; Kayal et al 2013; Desiro et al 2017).

Originally phylogenetic trees were mainly used for taxonomy purposes, primarily to establish the relationships among organisms. Phylogenetic analyses later became embedded in many aspects of evolutionary studies, such as comparative analysis that allows the testing for correlation among traits while correcting for the statistical non-independence of species (e.g. Reiss 2001; Shi et al. 2005; Herron & Michod 2008); molecular dating that allows estimating the timing of divergence events (Rutschmann 2006); and diversification analyses that can identify bursts of diversification or other changes in the rates of speciation or extinction (e.g. Jetz et al. 2012). In the context of complex trait evolution, phylogenetic trees are crucial for the accurate reconstruction of the history of transitions (Brockington et al. 2011). Indeed, changes of character states can be inferred along a phylogenetic tree, either under a maximum parsimony criterion of with different models implemented in maximum likelihood or Bayesian frameworks (Edwards & Donoghue 2013). This potentially allows the reconstruction of the history of independent characters that constitute complex traits, which can establish the order of character transitions. Phylogenetic methods can also identify evolutionary precursors of adaptive innovations, providing insights into the factors that increase the accessibility of new phenotypes (Marazzi et al. 2012). Besides this importance of phylogenetic methods for inference of character changes, these tools are instrumental to the study of gene and genome evolution.While phylogenetic trees can be used to infer and study the species tree, as described above, they are primarily based on genes, and therefore represent gene trees. These gene trees are key components of the study of molecular evolution. Firstly, inferring the relationships among homologous genes for a set of species provides insight into the dynamics governing genomic diversification, via the inference and modelling of gene duplications and losses (methods reviewed in Kristensen et al. 2011). In addition, these analyses can identify past whole genome duplications (Dehal et al 2005; Aury et al 2006). Secondly, analyses of gene trees allows inferring changes in gene properties by, for example, the comparison of homologous genes with varying characteristics (Christin et al 2009). Indeed, changes in expression patterns and protein sequences can be inferred along a phylogenetic tree, in a process similar to that used for morphological

characters (e.g. Christin et al. 2013, 2015; Cohen et al 2016; Gerstein et al 2014). However, changes in protein sequences can also be analysed with models that explicitly model selective pressures across time (Yang 1998, 2007; Jones et al 2016). This can identify past episodes of adaptive evolution and more generally can quantify the effects of natural selection on the functional diversification of proteins (Yang & Bielawski 2000; Christin et al. 2007, 2008, 2009a, 2009b; Bielawski et al 2016). Finally, ancestral states can be inferred, leading to reconstructions of likely ancestral proteins (discussed in Williams et al. 2006). However, these analyses are more powerful when the transitions are replicated, and traits that evolved repeatedly therefore constitute outstanding systems to understand molecular evolution.

Convergent phenotypes as study systems

Differentiating causation and coincidence requires replicating experiments to verify that the outcome is constant. While this is easily done in experimental biology, large-scale evolutionary events cannot be directly repeated. However, some events happened recurrently in different groups, providing natural replicates. These cases of convergent evolution represent similar answers to a shared challenge by distinct groups of organisms (Conway Morris, 2003; Christin et al. 2010, Stern 2013). Famous examples include the ability to fly, which evolved in insects, bats, and birds, or the emergence of similar features to prosper in aquatic environments in different mammal groups, and turtles. The list of convergent phenotypes is long, and includes specific morphological characters, biochemical cycles, and protein functions (Storz 2016; Reed et al 2011; Protas et al 2006; Tuinenf 2001; Jones 2006; Bork et al 1993). These natural replicates are necessary to assess the statistical association of different features with the emerging trait, and are therefore instrumental to all comparative analyses, whether or not they are specifically referred to as ‘convergent’ (Pagel 1999; Freckleton et al 2002). This concerns association among morphological features, ecological factors, or diversification rates.

In the context of complex trait evolution, evolutionary convergence provides natural replicates that can be used to test macro-evolutionary hypotheses (Edwards and Still, 2008) and allows assessing the likelihood of different evolutionary trajectories indirectly. Indeed, the path that leads to a given trait in one group is not necessarily

representative of a general pattern, but if the same path was repeatedly used by different groups, it can be interpreted as representing a general rule. Therefore the study of convergent traits can indirectly give insights into the adaptive landscape, including what the most common peaks were and how they were attained (e.g. Heckmann et al. 2013; Williams et al. 2013). In addition, studies of convergent evolution can assess the constraints that act on the evolutionary potential, with regard to the order of the acquisition of components and the importance of certain evolutionary precursors (e.g. Protas et al. 2006; Weinreich et al. 2006; Christin 2013, 2015). Because convergent phenotypes represent transitions to the same solution from a different starting point, the effect of the ancestral state on the realized derived phenotype can be assessed. Therefore, complex traits that evolved recurrently represent exceptional systems to assess the factors increasing the accessibility to new phenotypes. Among eukaryotes, one of the best examples of a complex trait that evolved recurrently is C_4 photosynthesis.

C_4 photosynthesis

During the light-independent phase of oxygenic photosynthesis, the energy stored as ATP during the light-dependent phase is used to produce sugar and O_2 from CO_2 and H_2O. The first step of this pathway is the fixation by the enzyme Rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase) of CO_2 into organic compounds (Sage & Monson 1999). In most plants, referred to as C_3 plants, this CO_2 is extracted directly from the atmosphere. However, Rubisco has a tendency to fix O_2 instead of CO_2 (Tcherkez et al. 2006), which produces different metabolites that need to be recycled in an energetically costly process named photorespiration (Mallman et al. 2014). When Rubisco evolved more than 2.8 billion years, the atmosphere was rich in CO_2, with almost no oxygen so that its dual CO_2/O_2 affinity was not problematic for early photosynthetic organisms (Christin and Osborne 2013). However, as atmospheric CO_2 decreased through time and O_2 concentrations increased as a direct consequence of the success of photosynthetic organisms, CO_2:O_2 relative concentration dropped drastically, revealing the flaw of Rubisco (Christin and Osborne 2013). In the low CO_2 atmosphere that prevailed for the last 30 million years, photorespiration became especially important in warm and dry conditions (Sage et al. 1999). Photorespiration increases in warm conditions as CO_2

solubility decreases faster than O_2 solubility with increasing temperature, a condition that also lowers Rubisco CO_2:O_2 specificity (Carmo-Silva et al. 2015). Similarly, aridity triggers stomatal closure to limit water loss, which limits CO_2 input in the leaf and increases the relative concentration of O_2 (Sage et al. 2001).

The problem of Rubisco and photorespiration was recurrently solved by the evolution of C_4 photosynthesis, a complex trait that evolved more than 62 times independently in different groups of flowering plants (Sage et al. 2011). C_4 plants solve the problem of Rubisco dual affinity by fixing atmospheric CO_2 via PEPC, an enzyme without affinity for O_2. This reaction happens in the mesophyll cells, and incorporates CO_2 into organic acids, which are then transformed and transported to different types of cells (Fig. 1; Hatch 1987; Sage & Monson 1999; Sage 2004). These latter cells are separated from the atmosphere as they are usually nested deep within the leaf, and host Rubisco and the Calvin cycle in C_4 plants. CO_2 is biochemically released therein to feed Rubisco (Fig. 1; Kanai and Edwards 1999). This concentrating mechanism saturates Rubisco with CO_2, which effectively suppresses photorespiration and increases the carbon gain per light absorbed (Hatch et al. 1987; Tcherkez et al. 2006). C_4 photosynthesis consequently boosts carbon assimilation in warm, high-light conditions (Atkinson et al. 2016). In addition, because the CO_2 concentrating mechanism allows plants to function with a lower stomatal conductance, C_4 photosynthesis increases water use efficiency. C_4 photosynthesis also improves nitrogen-use efficiency because the higher CO_2 concentration achieved around Rubisco allows plants to function with less of this enzyme (Evans et al. 1994; Ghannoum 2005, 2010; von Caemmerer 2008). The advantages of C_4 result in its success in open biomes in tropical and substropical areas. Although only 3% of extant plant species are C_4 (Sage et al. 2004), C_4 photosynthesis is responsible for one fifth to one quarter of global terrestrial primary production (Ehleringer et al. 1997; Still et al. 2003).

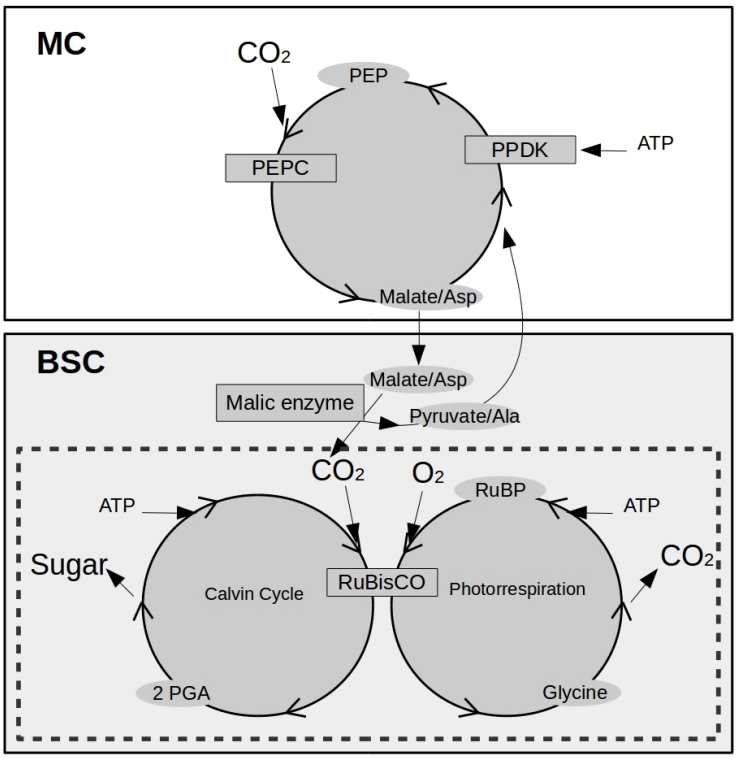


Figure 1: Simplified biochemical pathways of the light-independent phase of photosynthesis in a C_4 plant.
The main biochemical reactions of the light-independent photosynthesis phase are presented. The two cellular compartments are indicated (MS: mesophyll cell; BSC Bundle sheath cell). C_3 photosynthesis is delimited inside a dashed rectangle. In C_4 plants, the C_3 cycle is shifted to BSC. The C_4 biochemical cascade concentrates CO_2 around RuBisCO and suppresses photorrespiration. Squares represent enzymes; PEPC: phosphoenolpyruvate carboxylase; PPDK: pyruvate phosphate dikinase; RuBisCO: ribulose-1,5-bisphosphate carboxylase. Reproduced from Mallman et al. (2014).

The recurrent origins of the C_4 trait despite its apparent complexity constitutes a paradox, which can be solved by studying the trajectories from C_3 ancestors to C_4 descendants. Comparative analyses have indeed shown that some C_3 groups possess gross leaf anatomies that are similar to those of C_4 plants (Lundgren et al. 2014). Because of shared ancestry, these characteristics tend to be clustered in some phylogenetic groups, and C_4 evolved recurrently within these same groups (Christin et al. 2013; Griffiths et al. 2013). These traits were therefore interpreted as evolutionary enablers, which increase the accessibility of the C_4 phenotype. Once these traits, which include large bundle sheath areas and high vein density, are in place and some organelles are present in the bundle sheath, a CO_2-recycling mechanism can emerge. This photorespiratoy pump has been reported in a number of plants species, which were

originally referred to as 'C_3-C_4' intermediates, because they are physiologically intermediate between the two types (Ku *et al.* 1983; Monson *et al.* 1984, Mallman et al. 2014; Edwards and Ku 1987, Kennedy and Laetsch 1974; Rajendrudu *et al.* 1986; Hylton *et al.* 1988; Morgan *et al.* 1993; Sage *et al.* 2012). In these plants, the last step of the photorespiratory pathway, which consists in the release of CO_2 by the enzyme glycine decarboxylase, is segregated in the bundle sheaths, so that the photorespired CO_2 can be refixed by the Rubisco present there instead of diffusing back to the atmosphere (Monson et al. 1984; Edwards and Ku 1987). The evolution of this cycle is thought to be initiated by slight modifications in the cell-specificity of glycine decarboxylase (Rawsthorne *et al.* 1988a; Engelmann *et al.* 2008; Schulze *et al.* 2013). Because this trait relies on a mesophyll/bundle sheath segregation of carbon fixation, it likely constitutes an evolutionary stable stage between C_3 and C_4 states, in which enlarged bundle sheath areas with a high concentration of organelles are favoured (reviewed in Brautigam and Gowik 2016).

The emergence of a C_4 biochemistry in plants with anatomical enablers has only recently started to be understood. The main enzymes of the C_4 biochemical pathway were identified long ago (reviewed in Hatch 1987; Kanai and Edwards 1999). All of these enzymes also exist in C_3 plants, but are responsible for non-photosynthetic, mainly anaplerotic functions (Monson 2003; Aubry et al. 2011). In evolutionary terms, the emergence of a C_4 biochemistry therefore corresponds to the co-option of several enzymes, with modifications of their levels, tissue specificity, and catalytic properties (reviewed in Brautigam and Gowik 2016). Because an abundance of only some of the C_4 enzymes will not trigger a C_4 pathway, as evidenced by knock-down of some C_4 genes in C_4 plants or introduction of some C_4 enzymes in C_3 plants (Dever *et al.* 1995; Dever *et al.* 1997; Pengelly *et al.* 2012; Fahnenstich *et al.* 2007; Hausler *et al.* 2001; Hausler *et al.* 2002), the order in which these enzymes were co-opted is intriguing. It has recently been observed than the CO_2-recycling mechanism of C_3-C_4 intermediates creates an imbalance of nitrogen among cell types, which can be circumvented by an increase of some enzymes that are also involved in the C_4 pathway (Mallmann et al. 2014). Based on several models, a weak rudimentary C_4 cycle might therefore emerge in C_3-C_4 plants for unrelated reasons (Mallmann et al. 2014). Once this pathway exists, any increase of flux through the C_4 cycle will theoretically result in fitness gain, leading the

authors to describe the C_3 to C_4 trajectory as a smooth, continuous upward trajectory (Heckmann et al. 2013). These models, however, did not consider individual genes and enzymes when describing metabolic modules. The order and amount of genetic changes underlying the emergence of a C_4 pathway therefore remain poorly understood.

Similar to the majority of plant enzymes, most enzymes of the C_4 pathway are encoded by multigene families, where gene lineages emerged via repeated gene-specific and whole genome duplications (Wang et al. 2009; Hibberd and Covshoff 2010; Christin et al. 2013). The co-option of genes into the C_4 cycle involved changes in their expression patterns, to reach very high, cell- and time-specific levels (e.g. Bräutigam et al. 2011 2014; Külahoglu et al. 2014). Recently, it was shown that this was achieved via the recruitment of pre-existing regulatory mechanisms (reviewed in Reyna-Llorens & Hibberd 2017). In addition, the catalytic properties of the encoded enzymes were modified to fit the new catalytic context and its high concentrations of substrates and products and high fluxes related to photosynthesis (Svensson et al. 2003; Besnard et al. 2009; Christin et al. 2009; Wang et al. 2009; Mallmann et al. 2014). Comparative gene studies have shown that this was achieved via adaptive changes of the amino acid sequences of proteins linked to the C_4 pathway (Christin et al. 2007 2009; Huang et al. 2016). Intriguingly, previous work in both grasses and Caryophyllales, two groups with a high concentration of C_4 origins, has shown that some of the multiple genes encoding C_4 enzymes had been co-opted for C_4 more often than expected by chance (Christin et al. 2013b, 2015). This suggests that these genes were more suitable for the C_4 function, and the observation that the co-opted genes were the most abundant in some C_3 species indicates that the predisposition might stem from their expression patterns (Christin et al. 2013b; Emms et al. 2016). However, the genetic information available to study C_4 evolution remained limited. On one hand, complete genomes were available, but only for a few distantly related C_3 and C_4 species. On the other hand, sequence data had been generated for large samples of C_3 and C_4 taxa, but only for a limited number of genes, and without associated expression data (Christin et al. 2007 2009). Consequently, there was a need to study in detail the history of genes related to C_4 photosynthesis in closely related C_3 and C_4 species.

1. 7 Grasses

The grass family (Poaceae) includes more than 10,000 species, which are found all over the world, in habitats ranging from the tropics to the Arctic circle and from the shade of rainforests to the high-light open biomes (Kellogg 2000). Grasses constitute the main component, and the main food source, in numerous ecosystems (e.g. Fig. 2) and in agriculture, as many major crops are grasses (rice, wheat, barley, maize, sorghum, millet), as are most biofuel or fodder species (Byrt & Furbank 2011). While the age of the grasses is debated (Christin et al. 2014), it is generally estimated between 70 and 100 Ma, based on molecular dating analyses that take into account fossil evidence. However, the origin of grasses was not directly followed by diversification, and the first three splits in the grass phylogeny lead to species-poor, tropical groups (Grass Phylogeny Working Group 2012). The large majority of grass species occurr in two large sister clades, named 'BEP' and 'PACMAD' clades, whose names are based on the subfamilies that compose each of them (Soreng et al 2015). These two clades are roughly similar in size. The BEP clade consists of tropical lineages (*i.e.* bamboos and rice relatives), as well as the large subfamily Pooideae that mostly dominate temperate and cold regions. The PACMAD clade mainly consists of tropical lineages, and includes all C_4 grasses, together with a number of C_3 species (Grass Phylogeny Working Group 2012).

Figure 2: The grass *Alloteropsis semialata* in a grassland in South Africa. Photo by Marjorie Lundgren.

Based on phylogenetic evidence, the C_4 pathway evolved independently 22-24 times in the PACMAD clade of grasses (*viz*. Grass Phylogeny Working Group II), making it the family with the largest number of C_4 origins (Sage et al. 2011), and by far the largest number of C_4 species (Grass Phylogeny Working Group II; Soreng et al. 2015). In addition, C_4 grasses are the most important C_4 group, as the high total productivity linked to C_4 plants mainly reflects that of C_4 grasses, which dominate savannas and other tropical biomes (Lehmann & Parr 2016). Several of our most productive crops, such as maize, sugarcane, sorghum, and elephant grass, are C_4 grasses. Understanding the details of C_4 genetics could yield knowledge useful for the improvement of C_4 crops, and also C_3 crops by the intogression of this trait to boost productivity. In particular, engineering C_4 photosynthesis into rice is predicted to lead to strong yield improvements (Hibberd et al. 2008; von Caemmerer et al. 2012). This interest means grasses are an important study system for economic reasons, but the group is also well suited for comparative analyses. The multiple origins of C_4 in the family means that closely related C_4 groups can be compared, together with close C_3 relatives that are available. In addition, a number of 'C_3-C_4' intermediates exist in the same family, and in some cases these are closely related to C_4 lineages (Christin et al. 2012). Of special interest, grasses include the only species known to have C_3, C_4 and C_3-C_4 populations, namely *Alloteropsis semialata* (Ellis 1974; Lundgren et al. 2016). The grasses as a whole are therefore well suited to study C_4 evolution in an ecologically meaningful context. When coupled with studies conducted in parallel on groups of eudicots that are ecologically marginal but genetically, physiologically, or biochemically more tractable (e.g *Flaveria* species in Mallmann et al. (2014) and in Lyu et al. (2015), or Cleomaceae species in Bräutigam et al. (2011)), comparative genetic studies of C_3 and C_4 grasses can shed new light onto the factors that increase the accessibility of novel, complex phenotypes.

Hybridization and lateral gene transfer might provide evolutionary shortcuts

Historically, macroevolutionary changes have been studied considering changes along the species tree. Indeed, for eukaryotic organisms at least, the widespread dogma was that species represent distinct entities that cannot exchange genes. This dogma has been

challenged with the accumulation of large genomic datasets, which have revealed recurrent gene flow among related species (Martin et al. 2013; discussed in Nadeau 2014), and in some cases, lateral gene transfers among distant relatives (reviewed in Andersson 2005; Zhaxybayeva & Ford Doolittle 2011; Hotopp et al. 2007). In several cases, gene flow among related species has been linked to adaptive evolution, either by spreading adaptive loci or by increasing the genetic diversity, and therefore adaptive potential of populations (Hendry et al. 2002; Ford et al. 2005; The Heliconius Genome Consortium, Dasmahapatra et al. 2012). Several recent studies have provided new examples of gene exchanges among more distant relatives (reviewed in Soucy et al. 2015). Whoever, the implications of this mechanism are still debatable.

In the context of C_4 evolution, some potential cases of transmission of C_4 genes across species boundaries have been identified. This is especially the case of some populations of *Alloteropsis semialata*, which acquired key C_4 genes from distantly related C_4 grasses, via an unknown process (Christin et al. 2012). Other candidates include the genus *Neurachne* (Christin et al. 2012b) and the sedge *Eleocharis* (Besnard et al. 2009). In each of these cases, discrepancies between phylogenetic trees for genes encoding a C_4 enzyme and the species tree suggested reticulate evolution. Because the coding regions of C_4 genes undergo adaptive changes to adapt the encoded protein to the C_4 context, receiving a gene that is already optimized for the C_4 function might offer an adaptive evolutionary shortcut (Christin et al. 2012). However, these cases were studied with few sequences and a wider evaluation of the significance of this phenomenon is required.

In addition to C_4 evolution, cases of reticulate evolution have been reported in some small groups of grasses, including hybridization and allopolyploidy (Mason-Gamer & Linder 2004; Mason-Gamer et al. 2010). However, such studies were limited by the small number of species in the studied subgroups of grasses. On a family level, discrepancies between organelle markers and a couple of nuclear sequences (Christin et al. 2009b) might suggest ancient or recurrent events of reticulate evolution, although similar patterns might arise from incomplete lineage sorting or phylogenetic errors. There is therefore a real need to evaluate the possibility that hybridization and/or lateral gene transfer contributed to the spread of adaptive loci among grasses.

Thesis plan

In this project, I study the origins of the C_4 biochemical pathway in grasses, using phylogeny-based approaches. I use transcriptome data, which are sequences corresponding to the messenger RNA present in a tissue at the time of sampling. Transcriptomes provide a sample of the coding sequences of each species, which are the only markers useful at a taxonomic scale corresponding to a 70 million years old family. In addition, they provide estimates of the transcript level of each gene, which is fundamental to understand the origins of the C_4 biochemical pathway. Different datasets have been generated using this approach, and all have been analysed using state-of-the-art phylogenetic tools or analyses developed for this project. Four distinct, complementary studies have been conducted, which represent independent research papers.

The first chapter (Chapter I) assesses the factors affecting the likelihood of different genes being co-opted for C_4 evolution. Capitalizing on the recurrence of C_3 to C_4 transitions in the grass family, I sampled species representing multiple C_4 origins as well as their close C_3 relatives. Analyses of leaf transcriptomes identified the genes used for C_4 by each of the considered origins. Based on the data from C_3 species, the abundance and tissue-specificity of these genes were then inferred for the C_3 ancestors, and modelling showed that the most highly expressed genes had been preferentially co-opted, providing a relatively easy first step toward a rudimentary C_4 pathway.

In the second chapter (Chapter II), conducted in collaboration with Dr Luke Dunning, we evaluate the number of changes that are needed to evolve the C_4 biochemical pathway once the evolutionary precursors identified in Chapter I are in place. As distant species are distinguished by a high number of features that are not related to photosynthetic differences, we addressed this question using a species that contains C_3, C_4 and C_3-C_4 populations, the grass *Alloteropsis semialata*. Based on transcriptomes capturing the diversity within and among each photosynthetic type, we demonstrate that, once enablers are present, the transition to a rudimentary C_4 biochemical pathway requires the modification of very few components, with most changes occurring later, once the plants are already in the C_4 state.

To understand the impact of reticulate evolution on adaptive evolution, I

performed a detailed study on the history of C_4 genes within the genus *Alloteropsis*. In this third chapter (Chapter III), conducted in collaboration with Dr Luke Dunning and Dr Marjorie Lundgren, we use innovative approaches to the problem of identifying C_4 origins, establishing when each of the C_4 genes acquired its C_4-specific properties, and how the gene was transmitted with respect to the species. We demonstrate that the transition to C_4 happened three times in the genus from ancestors possessing some C_4 components. Surprisingly, reticulate evolution spread key C_4 genes among the three C_4 groups, so that the number of origins varies among C_4 components.

Finally, in the fourth chapter (Chapter IV), I assess the importance of reticulate evolution across the Panicoideae subfamily of grasses, which contain most of the C_4 origins in the family. Using a spectrum of phylogenetic analyses to the transcriptome data generated in Chapter I, I present for the first time evidence that the history of the group involved multiple events of reticulate evolution, which could have contributed to the spread of adaptive novelties, including C_4 photosynthesis and other adaptations.

Overall, my work combines large-scale comparative analyses, with small-scale, detailed investigations, bridging the evolutionary processes over different time scales. My results provide insights into the processes underlying the recurrent emergence of the C_4 biochemical pathway in grasses, as well as the importance of reticulate evolution for the evolutionary diversification of plants.

Chapter I: Highly expressed genes are preferentially co-opted for C_4 photosynthesis

Jose J. Moreno-Villena[1], Luke T. Dunning[1], Colin P. Osborne[1], Pascal-Antoine Christin[1,2]

[1] Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom.

[2] Corresponding author: Pascal-Antoine Christin, p.christin@sheffield.ac.uk, +44-114-222-0027

This work has been published in 2018 in **Molecular Biology and Evolution, Volume 35, Issue 1, Pages 94–106.**

Personal contribution: I generated and analysed the data, after helping design the study. I wrote the paper with the help of my co-authors.

Abstract

Novel adaptations are generally assembled by co-opting pre-existing genetic components, but the factors dictating the suitability of genes for new functions remain poorly known. In this work, we used comparative transcriptomics to determine the attributes that increased the likelihood of some genes being co-opted for C_4 photosynthesis, a convergent complex trait that boosts productivity in tropical conditions. We show that independent lineages of grasses repeatedly co-opted the gene lineages that were the most highly expressed in non-C_4 ancestors to produce their C_4 pathway. While ancestral abundance in leaves explains which genes were used for the emergence of a C_4 pathway, the tissue specificity has surprisingly no effect. Our results suggest that levels of key genes were elevated during the early diversification of grasses and subsequently repeatedly used to trigger a weak C_4 cycle via relatively few mutations. The abundance of C_4-suitable transcripts therefore facilitated physiological innovation, but the transition to a strong C_4 pathway still involved consequent changes in expression levels, leaf specificity, and coding sequences. The direction and amount of changes required for the strong C_4 pathway depended on the identity of the genes co-opted, so that ancestral gene expression both facilitates adaptive transitions and constrains subsequent evolutionary trajectories.

Keywords: C_4 photosynthesis, evolvability, grasses, phylogenetics, transcriptomics, gene co-option

Introduction

The evolution of novel physiological adaptations occasionally requires the development of new biochemical cascades, which are generally achieved via the co-option of pre-existing genes into new functions (Duboule & Wilkins 1998; True & Carroll 2002; Monson 2003; Monteiro & Podlaha 2009). Rewiring of biochemical pathways can require both modifications of spatial and temporal gene expression patterns and alterations of the coding sequences to adapt the encoded enzymes to the new catalytic context (Duret & Mouchiroud 2000; Carroll 2008; Aubry et al. 2014). In cases where numerous modifications are needed, the novel pathways can be assembled by natural selection only if a functional version can emerge through relatively few changes, allowing subsequent selection to fix mutations that increase the efficiency of the pathway. Genomic factors that reduce the phenotypic distance between ancestral and novel physiologies, thereby enabling the emergence of novel cascades via few mutations, would consequently be expected to increase accessibility to novel phenotypes. However, in most cases these factors remain poorly understood.

The ability of given genes or genomic features to trigger evolutionary innovation can be investigated via experimental evolution (e.g. Weinreich et al. 2006; Blount et at. 2012), but such studies are restricted to short-lived organisms that do not encapsulate the existing diversity of phyla. For larger organisms with long generation times, a historical approach is the most appropriate. Indeed, phylogenetic inference allows explicit tests of how specific features affect the accessibility of new phenotypes (e.g. Marazzi et al. 2012). Conversely, genomic features that have recurrently contributed to independent origins of a given phenotype can be safely assumed to be suitable for the trait of interest, and their origin can be regarded as potentially facilitating later adaptive transitions (Huang et al. 2016b). For example, the same autosome pairs were repeatedly co-opted to evolve sex chromosomes in turtles (Montiel et al. 2017), the same gene families encoding crystallins were used to evolve camera eyes in cephalopods and vertebrates (Zinovieva et al. 1999; Yoshida et al. 2015), and homologous genes recurrently contributed to the diversification of coloration patterns in butterflies (Jiggins et al. 2017). While such evidence indicates that some genomic regions or genes preferentially contribute to specific evolutionary transitions (Tenaillon et al. 2012), multiple factors might increase the adaptive potential, and their identification requires

the comparison of the ancestral condition of genes or genomic regions that were recurrently co-opted, to those that were not.

An excellent system to study the factors that increase gene adaptive potential is C_4 photosynthesis. This novel physiology requires a biochemical cascade arising from the high activity of multiple enzymes in specific leaf compartments, and improves autotrophic carbon assimilation in tropical conditions (Pearcy and Ehleringer, 1984; Hatch 1987; Sage et al., 2012, Atkinson et al. 2016). The C_4 trait is ecologically and agronomically extremely important (Ehleringer et al., 1997; Still et al., 2003; Byrt et al., 2011). It evolved more than 60 times in independent lineages of flowering plants (Sage et al. 2011), via the co-option of multiple genes that were present in non-C_4 ancestors (Hibberd and Quick 2002; Aubry et al. 2011; Brown et al. 2011; Kajala et al. 2012). Most enzymes of the C_4 pathway are encoded by multigene families, whose members differed in their expression patterns and catalytic properties of the encoded enzymes before their involvement in C_4 photosynthesis (Wang et al., 2009; Hibberd and Covshoff, 2010; Aubry et al. 2011; Christin et al., 2013, 2015). Previous comparisons of a handful of C_4 species have shown that a subset of gene lineages were recurrently co-opted for C_4 evolution, both among grasses and among the distantly-related Caryophyllales (Christin et al. 2013, 2015). However, the co-opted genes differed between grasses and Caryophyllales, suggesting that factors predisposing some genes for a C_4 function are specific to subgroups of angiosperms (Christin et al. 2015). It has been noted that the co-opted genes appeared to be highly expressed in the non-C_4 taxa available at the time for comparison, which might have contributed to their preferential co-option (Christin et al. 2013; Emms et al. 2016). However, systematic tests of the factors underlying the observed co-option bias are still lacking.

In this study, we compare transcriptomes across ten independent C_4 origins in grasses, and their non-C_4 relatives. Through a combination of phylogeny-based analyses, we test (i) whether a bias in the gene lineages co-opted exists across the whole set of grasses. To determine the causal factors underlying the bias, we then test (ii) whether the expression level in leaves and/or (iii) whether the tissue specificity in the non-C_4 ancestors explain variation in the co-option probability among gene lineages. In addition, we analyse coding sequences to test (iv) whether adaptive changes in the coding sequences occurred during or after the emergence of the C_4 physiology.

Together, our investigations shed new light on the factors that increase the adaptive potential of some genes, focusing on a complex trait of ecological and agronomical importance.

Results

Sequencing, read mapping and transcriptome assembly

In total, 74 individually sequenced RNA libraries from 19 species generated over 550 million 100bp paired-end reads. This represents 98.87 Gb of data, with a mean of 1.34 Gb per library (SD = 0.95 Gb; Table I.S1). Over 81% of the reads were kept after removing low-quality reads and ribosomal RNA sequences. Transcriptomes were assembled with a mean of 2.23 Gb per species (SD = 1.40 Gb), resulting in a mean of 54,255 Trinity 'unigenes' (SD = 17,218.35), 79,566.12 contigs (SD = 23,038.61), and a 1560.05 bp N50 (SD = 184.95 bp).

The C_4-related gene families considered in this study constitute 5.1% (SD = 2.02%) of the reads in the leaf libraries of C_4 plants, versus 2.34% in non-C_4 plants (SD = 0.75%). On average, 1.05% of the reads from the root libraries mapped to C_4-related genes (SD = 0.48%).

Phylogenetic trees and identification of genes co-opted for C_4 photosynthesis

A total of 533 nuclear core-orthologs were used to infer the species tree, which was well resolved (Fig. I.1). The relationships among grass subfamilies mirror those retrieved previously with other datasets (GPWG II, 2012). However, relationships within the Paniceae tribe (the group most densely sampled here) differ in several aspects from those based on plastid markers (GPWG II, 2012), and were closer to previous analyses that also included nuclear markers (Vicentini et al. 2008). The placement of the different C_4 origins within the tree was largely congruent with previous studies, and their non-C_4 relatives separated them in the phylogeny as expected (Fig. I.1).

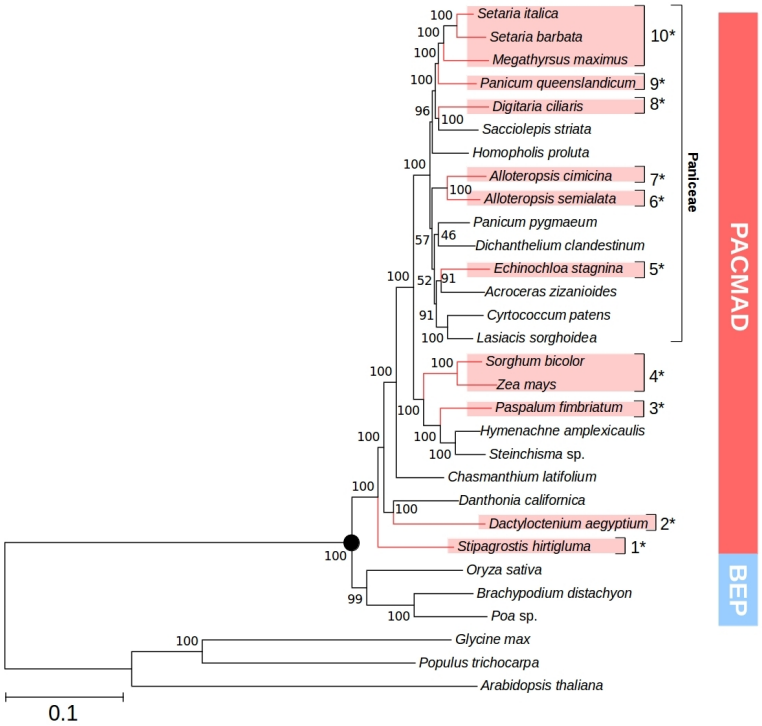


Figure I.1. Phylogenetic tree based on nuclear orthologs.
C_4 taxa are in red, and C_4 origins are numbered. One of the tribe and the two main clades of grasses are indicated on the right. The black circle highlights the node representing the common ancestor of the sampled grasses. Bootstrap values are indicated near branches

For each gene family encoding C_4-related enzymes, phylogenetic inference confirmed previous conclusions about orthology (Vilella et al. 2009). The enzyme phosphoenolpyruvate carboxykinase (PCK) and the Na^+/H^+ antiporter (NHD) are each encoded by a single gene lineage (Fig. SI.1). The number of grass co-orthologs in other families varies from two (for pyruvate, phosphate dikinase - PPDK) to eight (for triose phosphate-phosphate translocator – TPT; Fig. I.S1). Groups of co-orthologs were named as in Christin et al. (2015). Phylogenetic relationships inferred in these gene trees were mostly congruent with the species tree. Exceptions include genes for PCK, where *Echinochloa stagnina* and *Alloteropsis semialata* grouped with those of *Setaria barbata*. This pattern has previously been reported for *Alloteropsis* species and this, together with a number of other lines of evidence, was interpreted as the fingerprint of a lateral gene transfer from *Setaria* or its close relatives (Christin et al. 2012; Dunning et al. 2017). Other incongruences were observed in genes encoding PEPC, PPDK, NAD(P)-malate dehydrogenase [NAD(P)-MDH], Sodium bile acid symporter family (SBAS), TPT, and NDH (Fig. I.S1), and could stem from a combination of reticulate evolution during

grass diversification and phylogenetic bias due to adaptive evolution. Gene duplicates specific to subgroups of grasses are evident for several genes, and can in some cases be associated to recent polyploidy (e.g. in *Zea mays* genes *pck-1P1*, *ppc-1P4*, *ppdk-1P2*, *nadmdh-4P7*; Fig. I.S1). Our analytical pipeline cannot estimate the expression level individually for each of these duplicates with very similar sequences, but these duplications specific to subgroups of grasses are relatively recent and occurred after the divergence of C_3 and C_4 clades (Fig. I.S1). The inferred evolutionary changes in expression patterns and co-option events are consequently not affected.

The most highly transcribed genes encoding C_4-related proteins are those for β-carbonic anhydrase (βCA; Fig. I.2; Table I.S2), an enzyme that acts in the cytosol of mesophyll cells in C_4 plants. These genes are however equally abundant in non-C_4 species (Fig. I.2), where the enzyme plays a key role in the chloroplasts of mesophyll cells (Tetu et al. 2007). Of the 31 other gene families encoding enzymes that can be related to the C_4 pathway, 14 included gene lineages with transcript abundances above 500 rpkm in at least one C_4 species (Fig. I.3; Table I.S2). The transcript abundance of *ppa-4P4* reached 500 rpkm in some C_4 species, but similar abundance was observed in a number of non-C_4 taxa (Table I.S2), and the gene was consequently not counted as C_4 specific. For the rest of the gene lineages, such high values were not found in non-C_4 species (Table I.S2). Genes co-opted for C_4 photosynthesis were identified in each C_4 species for most core C_4 enzymes, but putative C_4 transporters and regulators were not always abundant in C_4 leaves (Table I.S2). Genes for enzymes of the photorespiration pathway were downregulated in C_4 species, as expected (Table I.S2).

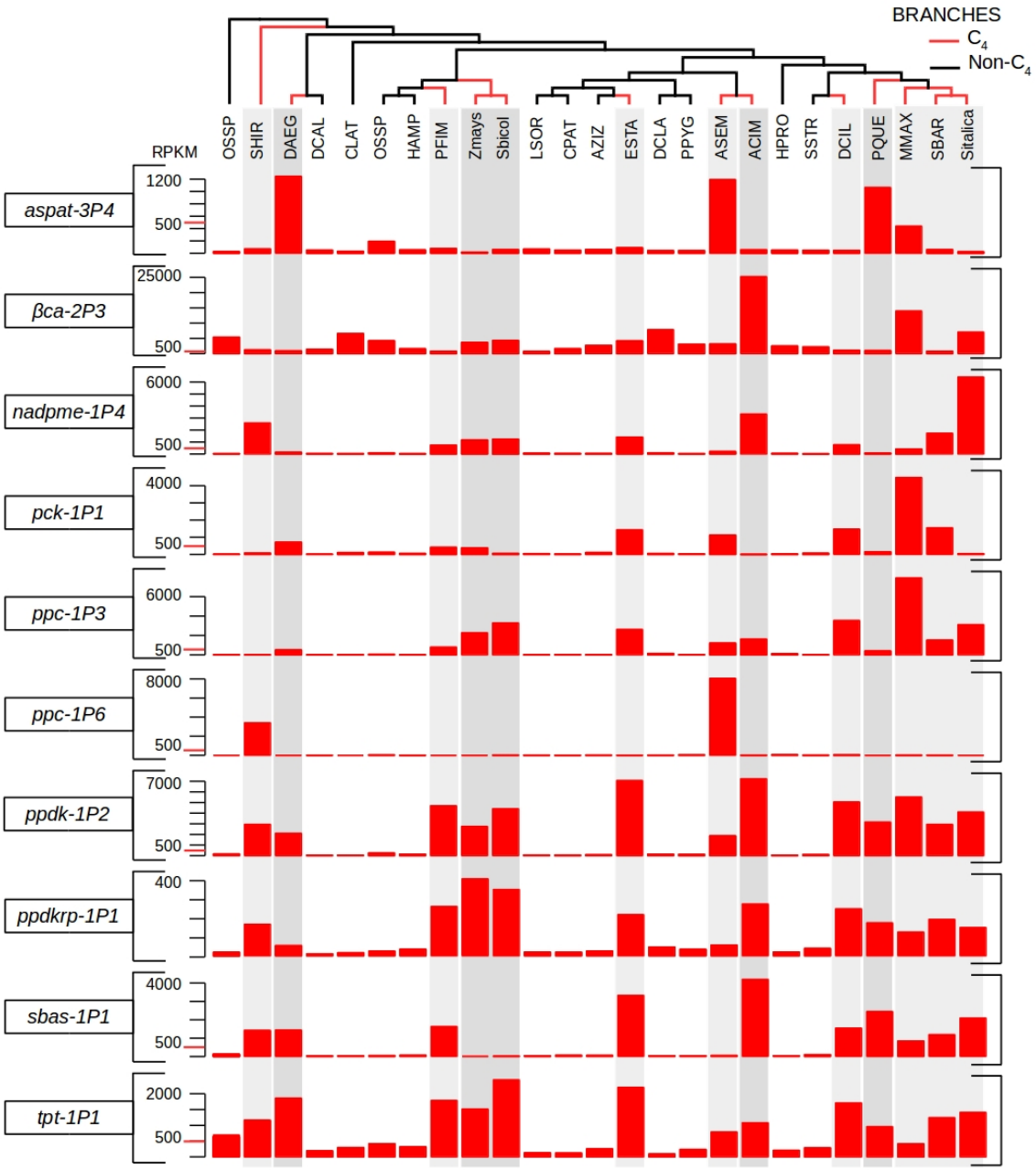


Figure I.2. Transcript abundances of the main C_4 genes in C_4 and non-C_4 species.
Barplot indicate rpkm values (reads per kilobase per million of reads) in leaves of C_4 (in red) and non-C_4 (in black species). Phylogenetic relationships among species are indicated at the top, and C_4 lineages are numbered as in Fig. I.1. Species names are abbreviated as in Tables I.S1 and I.S2.

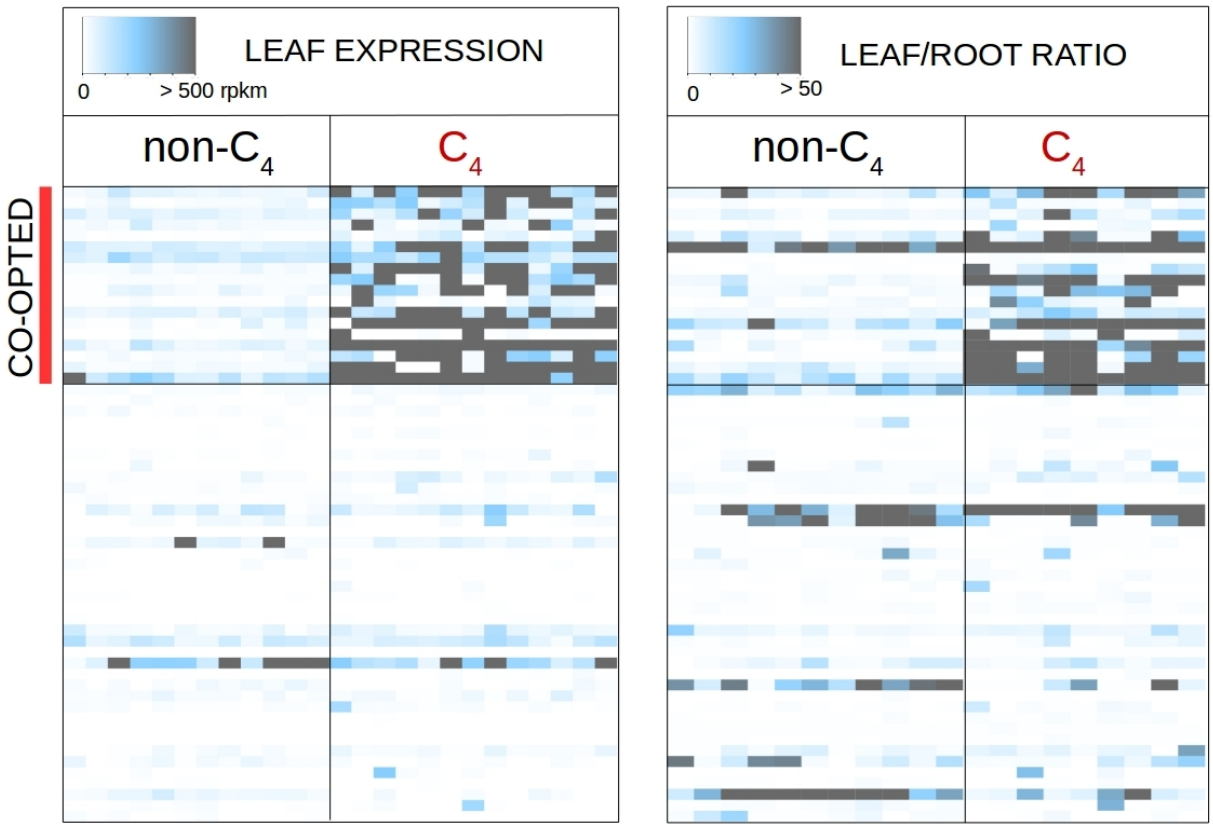


Figure I.3. Gene expression profiles of C_4-related genes in the studied taxa.
Colours indicate leaf transcript abundance and leaf/ratio abundance ratio for C_4-related genes in C_4 and non-C_4 species. Genes that have been co-opted at least once are at the top.

Factors affecting gene co-option

Out of 58 gene lineages encoding the 14 enzymes used by the C_4 species sampled here, only 18 have been co-opted at least once, and up to ten times independently for *ppdk-1P2* and *tpt-1P1* and eight for *ppc−1P3* (Table I.1). Given the size of the different gene families and the number of co-option events, fewer genes have been co-opted at least once than expected by chance (p-value < 0.00001). This confirms the existence of a co-option bias across the ten C_4 origins considered here, a result previously reported for Caryophylalles and grasses (Christin et al. 2013, 2015).

Table I.1. Number of times a gene lineage was co-opted, for genes co-opted at least once.

Gene lineage	Times co-opted	Main catalytic reaction
ak-1P1	8	AMP → ADP
alaat-1P5	3	Ala ↔ Pyruvate
aspat-2P3	3	Asp ↔ OAA
aspat-3P4	3	Asp ↔ OAA
dit-2P3	1	Dicarboxylate transporter
nadpmdh-1P1	5	Malate ↔ OAA
nadpmdh-3P4	1	Malate ↔ OAA
nadpme-1P4	7	Malate → pyruvate
nhd-1P1	5	Sodium proton antiport
pck-1P1	5	OAA → PEP
pepck-1P1	1	ATP ADP/P antiport
ppa-1P2.1	6	Pyrophosphate → phosphate
ppc-1P3	8	PEP → OAA
ppc-1P6	2	PEP → OAA
ppdk-1P2	10	Pyruvate → PEP
ppt-1P5	4	PEP phosphate antiport
sbas-1P1	8	Pyruvate sodium symport
tpt-1P1	10	3-PGA TP antiport

The ancestral state reconstructions inferred the abundance in leaves and leaf/root specificity in the last common ancestor of the sampled grasses for each C_4-related genes (Fig. I.4). This approach comes with uncertainty, especially for deeper nodes in a tree, but the confidence intervals associated with the inferred values are small compared to the difference among members of the same gene family (Fig. I.4). The inferred values are moreover tightly correlated with averages of the values among C_3 grasses (R^2 = 0.98 for the leaf abundance and R^2 = 0.91 for the leaf/root ratio), and were consequently used for modelling of gene co-option. Linear models showed that the ancestral transcript abundance in the leaf significantly affected the co-option frequency (F=13.11, df=56, p=0.0006336; R^2=0.19), and this stayed significant when the gene family was used as a co-factor (Table I.2). The effect of the ancestral leaf/root transcript abundance ratio on the co-option frequency was not significant when considered on its own (F=0.40, df=56, p=0.54), or in combination with the ancestral leaf abundance and the gene family cofactor (Table I.2). Therefore, our modelling analyses indicate that genes were co-opted for C_4 photosynthesis based on their transcription level in leaves (Fig. I.4), independently of the specificity of this expression in leaves compared with roots. The same conclusions were reached when using a threshold of 300, 1000 and 1500 rpkm for the identification of co-opted genes (see Table I.2).

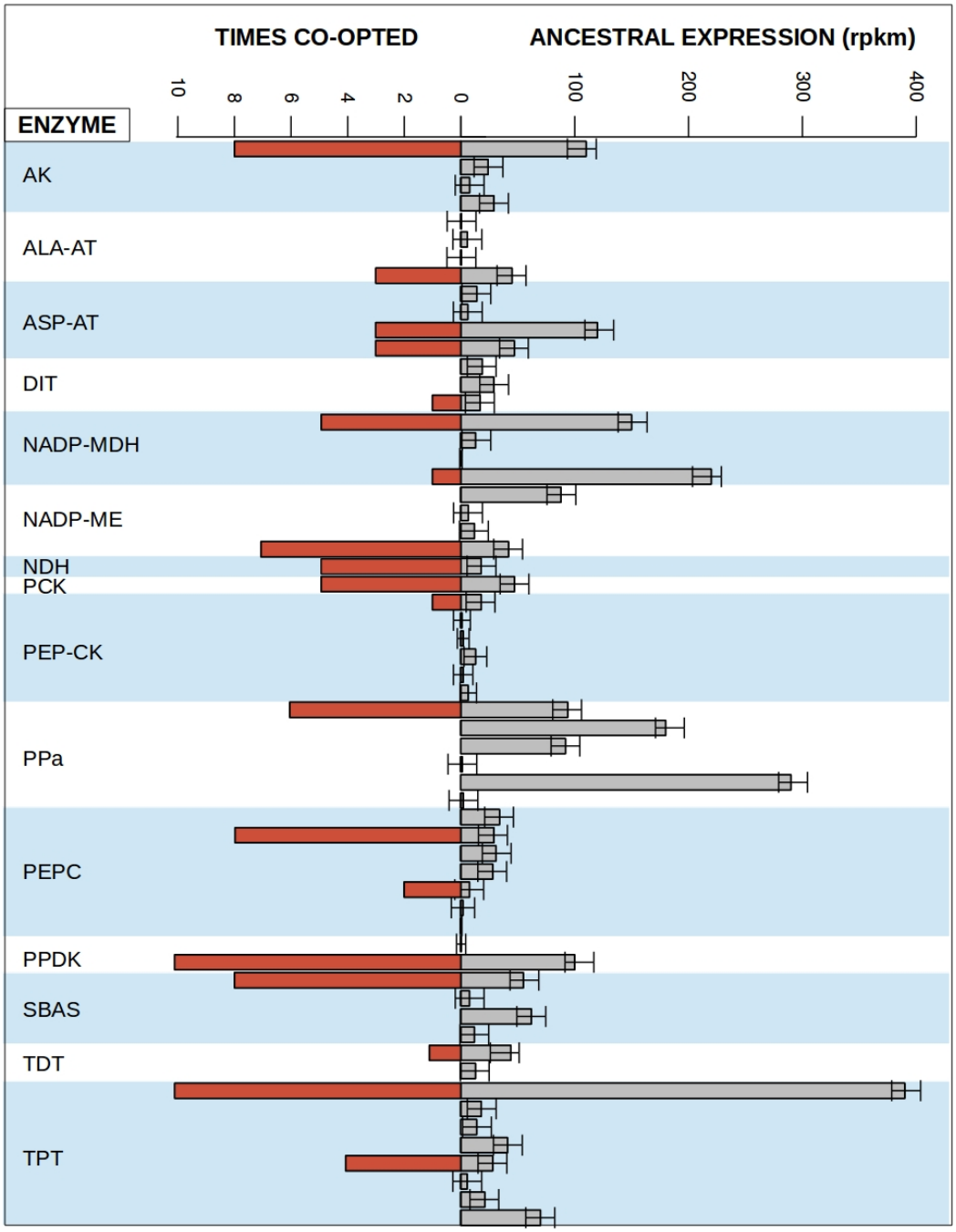


Figure I.4. Ancestral leaf transcript abundance and number of co-option events.
Barplots on the left indicate the number of times each gene was co-opted, and those on the right indicate the inferred abundance in the non-C_4 last common ancestor of grasses (see Fig. I.1), with the associated confidence intervals. Genes are sorted by enyzme, indicated on the left.

Table I.2. Results of analyses of variance on linear models of number of co-option events based on ancestral leaf abundance (ala), leaf/root ratio, and gene family identity (family), with co-opted genes identified with different rpkm thresholds.

rpkm threshold	**300**	**300**	**300**	**500**	**500**	**500**	**1000**	**1000**	**1000**	**1500**	**1500**	**1500**
Factors	**ala**	**leaf/root**	**family**	**ala**	**leaf/root**	**family**	**ala**	**leaf/root**	**family**	**ala**	**leaf/root**	**family**
p-value	0.00	0.52	0.38	0.00	0.57	0.56	0.00	0.88	0.21	0.01	0.77	0.10
df[1]	1,42	1,42	13,42	1,42	1,42	13,42	1,42	1,42	13,42	1,42	1,42	13,42
F-stat	17.07	0.78	0.95	12.65	0.32	0.90	14.46	0.21	1.37	8.29	0.0.09	1.71

[1] df = degrees of freedom. For each variable, the degrees of freedom for the residuals are given after the comma.

Transcriptome datasets for clades containing C_3 and C_4 species other than grasses are focused on small taxonomic groups, so that ancient evolutionary events cannot be inferred yet outside from grasses. A test using published transcriptomes for one C_3 and C_4 species within the eudicot family Cleomaceae failed to detect any effect of expression levels, on the identity of genes co-opted for C_4 (Tables I.S3 and I.S4), but the availability of a single C_4 origin and only one C_3 relative likely decreased statistical power. Although the same statistical limitations applied to the *Flaveria* dataset, our preliminary investigation suggested that the effect of leaf abundance on the co-option probability might apply to multiple C_4 origins across the angiosperms. Indeed, there was a significant effect of the leaf abundance in the close relatives on the co-option probability for *Flaveria* (Table I.S4).

Marked differences in transcript abundance and coding sequences

While the ancestral transcript abundance significantly affects the probability of a gene being co-opted, the evolution of C_4 photosynthesis is accompanied by major increases in transcript abundance. The transcripts of genes encoding C_4 enzymes increase by a fold change of up to 480 for *ppc-1P6* in *Alloteropsis semialata* compared to related non-C_4 taxa (Fig. I.2). In addition, their leaf specificity increases, to reach leaf/root ratios of up to 6204 after their co-option into C_4 photosynthesis, compared to a maximum of 257 in non-C_4 taxa (Fig. I.3).

Besides these changes in transcript abundance, tests for positive selection revealed adaptive evolution in the coding sequences of a number of genes during or slightly after their co-option into C_4 photosynthesis. After correction for multiple testing, the test for a shift of selective pressures along C_4 branches (A1 vs. M1a comparison) was significant

for nine genes out of 19 (Table I.S5). The test specifically testing for a shift to positive selection as opposed to a relaxation of selection (A1 vs. A comparison) was also significant for four of these nine genes; *ppc-1P3*, *ppdk-1P2*, *sbas-1P1*, and *tpt-1P1* (Table I.S5). The sites identified by the Bayes Empirical Bayes analysis as being under positive selection along C_4 branches showed widespread cases of parallel amino acid replacements (Fig. I.5).

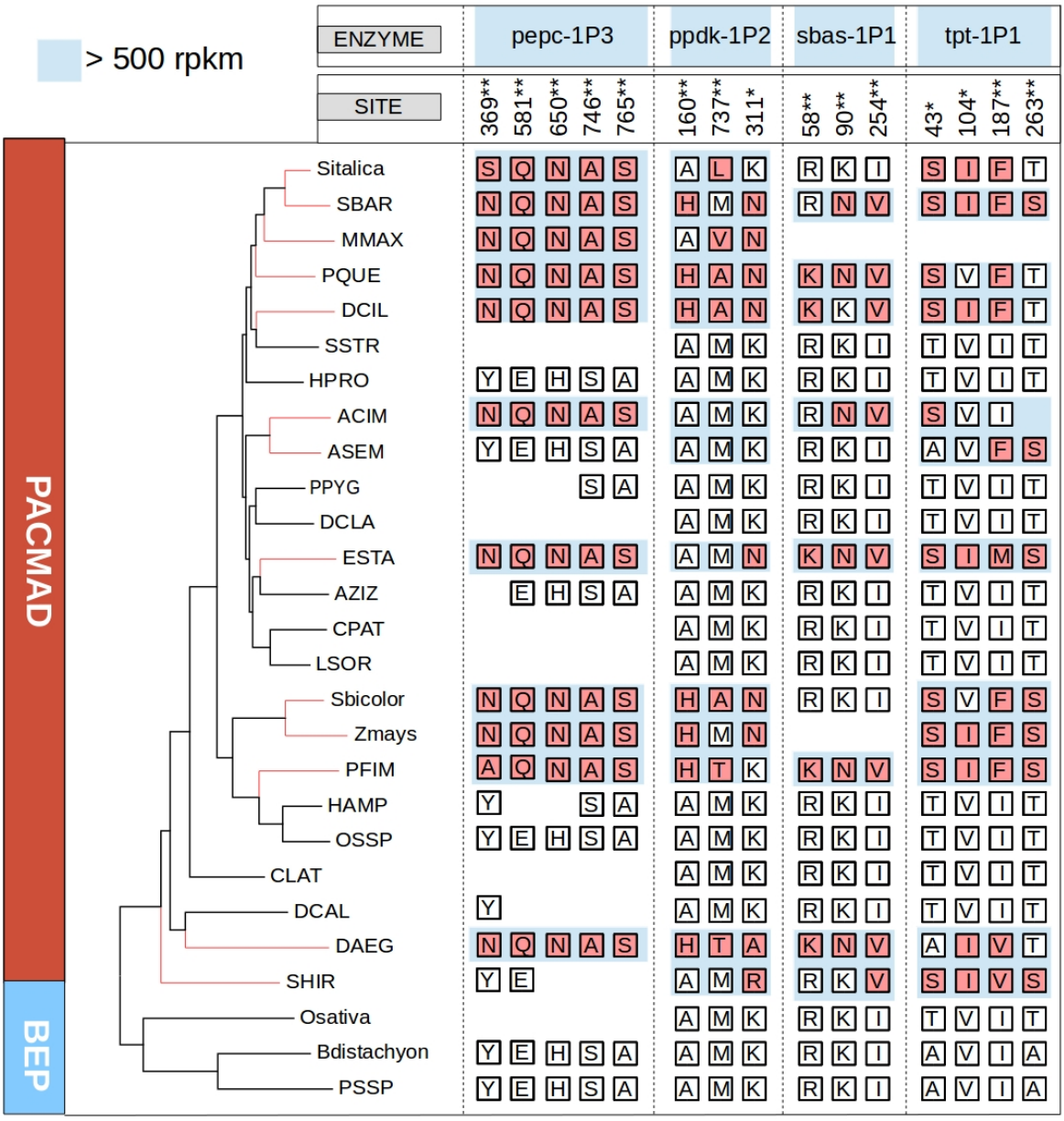


Figure I.5. Patterns of convergent adaptive amino acid replacements.
The phylogeny of the sampled species is indicated on the left, with species names abbreviated as in Table I.S1. Branches leading to C_4 species in red. Amino acids at sites under positive selection ($p<0.05$*; $p<0.01$**) are indicated on the right. Residues of co-opted genes are highlighted with a blue background.

Discussion

Expression patterns determined which genes were co-opted for C_4

In this study, we analyzed root and leaf transcriptomes from grass species representing ten independent origins of C_4 photosynthesis as well as the close non-C_4 relatives to each of them (Fig. I.1). As previously suggested based on smaller species samples (Christin et al. 2013, 2015), the co-option of genes for the C_4 pathway has been a non-random process. Indeed, despite multiple gene lineages existing for most C_4-related enzymes, a few of them were co-opted more frequently than expected by chance, while most were never used in the ten C_4 lineages evaluated here (Table I.1; Fig. I.3 and 4). A number of factors could explain the preferential co-option of some genes for a novel function, including their availability via genomic redundancy, the suitability of their kinetic properties, the fit of their expression patterns, and their evolvability (Aharoni et al. 2005; Landry et al. 2007; Christin et al. 2010, 2015; Stiffler et al. 2015; Huang et al. 2016b). Our approach was specifically designed to test for the effects on co-option probability of two dimensions of the expression patterns inferred for non-C_4 ancestors; the transcript abundance in leaves and the leaf versus root specificity. Thanks to the evolutionary-informed sampling (Fig. I.1), we were able to unambiguously show that the likelihood of gene co-option into C_4 photosynthesis was determined in a large part by their transcript abundance in leaves prior to C_4 evolution (Fig. I.4), with no apparent effect of the leaf to root specificity (Table I.2).

The C_4 biochemical pathway, like any complex pathway, is assumed to result from many rounds of fixation of adaptive mutations (Sage et al. 2012; Heckmann et al. 2013; Dunning et al. 2017). However, natural selection cannot gradually improve a pathway before it exists, even in a rudimentary stage (Huang et al. 2016b). It is likely that a primitive, weak C_4 cycle initially emerged in some species via a slight upregulation of few genes, as observed in intermediate plants accumulating only part of their CO_2 via the C_4 cycle (Mallmann et al. 2014; Dunning et al. 2017). We show here for the first time that some genes were already moderately abundant in leaves of non-C_4 plants (Fig. I.4), a pattern that likely evolved for a number of reasons not related to C_4 photosynthesis, but eased its later evolution. This facilitator effect would have been even stronger if C_4-related genes were upregulated in the low-CO_2 conditions that prevailed until the Industrial Revolution, as has been suggested for the distantly-related

Arabidopsis (Li et al. 2014). The encoded enzymes, present in the leaves of the non-C_4 ancestors, constituted the building blocks needed to generate a weak, yet functional, C_4 pathway following key mutations. These could have included further upregulation of key C_4 enzymes or alterations of the leaf structural arrangements, pushing the system beyond a tipping point where the C_4 pathway could emerge. Models predict that, once a C_4 pathway is in place, any increase in the rate of the C_4 pathway will increase productivity in warm conditions (Heckmann et al. 2013; Mallmann et al. 2014). Any rudimentary C_4 pathway based on ancestrally abundant enzymes would therefore have created the selective impetus for upregulation of enzymes, generating the striking patterns observed in derived C_4 plants (Fig. I.2 and I.3).

Besides elevated abundance of numerous enzymes, the C_4 trait is characterized by a precise compartmentalization of the biochemical reactions in different parts of the leaves (Hatch and Osmond 1976; Hatch 1987; John et al. 2014). Interestingly, transcript abundance in non-photosynthetic tissues, such as roots, did however not prevent the co-option of a gene lineage for C_4 photosynthesis (Table I.2; Fig. I.3), and previous pairwise comparisons have established that orthologs to C_4 genes have a diversity of expression patterns in non-C_4 species (Külahoglu et al. 2014). We conclude that being abundant in leaves was a sufficient condition for the C_4 function, independently of the presence in other tissues. Cellular and subcellular localization, which was not captured by our whole-leaf transcriptomes, probably still contributed to determining which genes were co-opted for C_4. For instance, only one of the four gene lineages for NADP-ME present in grasses encodes a chloroplast-specific isoform, and this gene lineage has been recurrently co-opted for C_4 despite an ancestral abundance of a second gene (Fig. I.4; Christin et al. 2009). Similarly, the product of *ppc-1P2*, the most highly expressed gene for PEPC in non-C_4 plants (Fig. I.4), is chloroplast-specific (Masumoto et al. 2010), which very likely prevented a function in C_4 photosynthesis, since this enzyme is cytosolic in the C_4 pathway. Independently of these specific cases, the mere moderate abundance in leaves explains a large fraction of the co-option probability.

Despite genetic enablers, C_4 evolution required massive changes

Our study is the first to scan the transcriptomes of a number of non-C_4 grasses closely related to C_4 species, and showed that genes co-opted for C_4 tended to already be

abundant in non-C_4 ancestors (Fig. I.3 and I.4). Although transcriptomes in other groups are not available for multiple C_4 origins and their C_3 relatives, our reanalysis of eudicot datasets suggested that the preferential co-option of the most abundant genes might underly C_4 origins in groups other than grasses (Table I.S4). This suggests that the abundance of some enzymes able to fulfil a C_4 function facilitated the emergence of a C_4 pathway. However, massive changes in gene expression are still observed between non-C_4 and C_4 relatives (e.g. Bräutigam et al. 2011, 2014; Külahoglu et al. 2014). Indeed, genes encoding C_4 enzymes are orders of magnitude more abundant in C_4 leaves, and leaf specificity strongly increased after the co-option of genes for C_4 (Fig. I.2 and I.3). In addition, evidence for widespread adaptive evolution of coding sequences for the C_4 context, obtained here and in other studies (Fig. I.5; Besnard et al. 2009; Christin et al. 2009; Wang et al. 2009; Huang et al. 2016a), suggests important modifications of the kinetic properties, shown for some enzymes (Bläsing et al. 2000; Tausta et al. 2002). Instead of being involved in the initial emergence of a C_4 cycle, we propose that these massive changes were involved in the transition from a weak to a strong C_4 pathway able to match the high rates of the Calvin cycle, as suggested for specific study systems (Svensson et al. 2003; Mallmann et al. 2014; Dunning et al. 2017).

Since the major requirement for a C_4 function was sufficient abundance in leaves, the co-opted genes were not necessarily the best suited for the C_4 function, in terms of the tissue specificity or kinetic properties of the encoded enzyme. The ancestral abundance might therefore have constrained the initial emergence of a weak C_4 cycle based on specific sets of genes, forcing natural selection to later adapt their properties to those required for a high-flux strong C_4 cycle. The recurrent co-option of the same co-orthologs would have increased the likelihood of adaptation via similar changes, explaining the observed parallel amino acid replacements among C_4 origins in grasses (Fig. I.5; Christin et al. 2007). It has been shown that C_4 lineages belonging to distant groups of angiosperms in some cases co-opted distinct genes (Christin et al. 2015; Table I.S4). Because of the large evolutionary distances separating these groups, which are further increased when different co-orthologs are co-opted (Table I.S4), the encoded enzymes likely varied in their kinetic properties in addition to their leaf and cell specificities. The amount of optimizing adaptive changes might have varied among major C_4 groups as a consequence, explaining that the frequency and identity of

selection-driven amino acid replacements shows high convergence among closely related C_4 lineages (Fig. I.5), but varies between C_4 origins in grasses and those in the distantly related sedges and eudicots (Besnard et al. 2009).

Conclusions

In this study, we sequenced the transcriptomes of species from the main C_4 grass lineages as well as their close non-C_4 relatives, and used models to show that the identity of genes co-opted for C_4 photosynthesis was largely explained by transcript abundance before C_4 evolution. The co-option, likely dictated by the mere presence of each protein in leaves, was followed by massive upregulation and widespread adaptation of coding sequences. Both of these processes likely accelerated and optimized a C_4 pathway that initially emerged from the combined action of enzymes already present in leaves. It is currently unknown why some gene lineages came to be more expressed than others in non-C_4 plants but, despite variation among species, the increased abundance of these genes seems to date back to at least the last common ancestor of grasses. Comparison among distant groups of angiosperms indicates that the preferential co-option of the most abundant gene lineages might be a recurrent pattern, but the sampling is not yet dense enough across angiosperms to precisely determine when increased transcript abundance first happened, among the ancestors of grasses and other groups that recurrently evolved C_4 photosynthesis. When this information is available, we might be able to test whether gene abundance combined with anatomical variation determined which plant lineages were more likely to evolve C_4 photosynthesis, once environmental changes created the selective pressure for this physiological novelty.

Material and Methods

Species sampling

Grass species were selected for analyses based on their photosynthetic type to include multiple C_4 origins and their non-C_4 relatives, based on previous phylogenetic analyses (GPWG II 2012). We sequenced eight C_4 species and eleven non-C_4 species, which separate them in the phylogenetic tree of grasses (GPWG II 2012, Fig. I.1). Most of

these belong to the PACMAD clade (subfamilies Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae and Danthonioideae), which contains all C_4 origins in grasses, and one non-C_4 Pooideae species was added as an outgroup for comparisons.

The selected species were grown from seeds, using the material from Atkinson et al. (2016) and Lundgren et al. (2015). Plants were maintained in controlled environment growth chambers (Conviron BDR16; Manitoba, Canada), with 60% relative humidity, 500 µmol m^{-2} s^{-1} photosynthetic photon flux density (PPFD), and 25/20°C day/night temperatures, with a 14-hour photoperiod. John Innes No. 2 potting compost (John Innes Manufacturers Association, Reading, England) was used. Plants were watered three times a week to keep the soil damp, and were fertilised every two weeks with Scotts Evergreen Lawn Food (The Scotts Company, Surrey, England). After a minimum of 30 days in these controlled conditions, two young roots and the most photosynthetically active distal half of fully expanded leaves were sampled from two individuals of each species (biological replicates) during the middle of the photoperiod, and immediately frozen in liquid nitrogen. All samples were stored at -80 °C until RNA extraction.

RNA extraction, sequencing and transcriptome assembly

Samples were homogenised in liquid nitrogen using a pestle and a mortar, and RNA was extracted using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. The isolated RNA was DNA digested on-column using the RNase-Free Dnase Set (Qiagen, Hilden, Germany) and eluted in RNAse-free water with 20 U/µL of SUPERase-IN RNase Inhibitor (Life Technologies, Carlsbad, CA). Extractions that yielded an RNA integrity number (RIN) greater than 6.5 and at least 0.5 µg of total RNA, as determined with the RNA 6000 Nano kit with an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, California), were used for upstream procedures. Individual RNA libraries were prepared using TruSeq RNA Library Preparation Kit v2 (Illumina, San Diego, CA), following the manufacturer's protocol with a target median insert length of 155 bp. A total of 24 indexed libraries were pooled per lane of flow cell and sequenced on an Illumina HiSeq 2500 platform with 100 cycles in rapid mode generating 100bp paired-end reads, at the Sheffield Diagnostic

Genetics Service.

Reads were filtered and assembled using the Agalma pipeline version 0.5.0, with default parameters (Dunn *et al.*, 2013). This pipeline removes low quality reads (Q <33), and those that are adaptor-contaminated or correspond to ribosomal RNA. The filtered reads are then used for *de novo* assembly using Trinity (version trinityrnaseq_r20140413p1; Grabherr *et al.*, 2011). One assembly was generated per species, using all the libraries available. Leaf assembly and reads in duplicates from the C_4 *Alloteropsis cimicina* were retrieved from Dunning et al. (2017), and reads for the C_4 *Megathyrsus maximus* and the non-C_4 *Dichanthelium clandestinum*, in triplicates and without replicate, respectively, were retrieved from Bräutigam et al. (2014). RNA-seq reads for C_4 grasses with a completely sequenced genome were also retrieved from the literature [*Setaria italica* without replicate from Zhang et al. (2012), *Zea mays* without replicate from Liu *et al.* (2015), and *Sorghum bicolor* in duplicates from Fracasso *et al.* (2016)]. The final RNA expression dataset included 12 non-C_4 species and 13 C_4 species of grasses.

Inference of a species tree based on core orthologs

Coding sequences (CDS) were predicted from the assembled contigs and those retrieved from the literature using the standalone version of OrfPredictor (Min et al. 2005). Protein sequences of eight publicly available genomes (*Arabidopsis thaliana*, *Brachypodium distachyon*, *Glycine max*, *Oryza sativa*, *Populus trichocarpa*, *Setaria italica, Sorghum bicolor* and *Zea mays*) were used as references to improve the identification of open reading frames by providing the program with a pre-computed BLASTX output file, using parameters suggested by the authors (Min et al. 2005). CDS from contigs with “no hit” in the BLASTX output were predicted *ab initio*. The predicted CDS were used for subsequent analyses.

CDS homologous to an *a priori* defined set of plant genes were retrieved using a Hidden Markov Model based search tool (HaMSTR v.13.2.3; Ebersberger et al. 2009). The set of genes includes 581 single copy core-orthologs from plants and is derived from the Inparanoid ortholog database (Sonnhammer and Ostlund 2014), using five high quality genomes (*Arabidopsis thaliana*, *Vitis vinifera*, *Oryza sativa*, *Sorghum bicolor* and *Ostreococcus lucimarinus*). Sequences were aligned as described in

Dunning et al. (2017); alignments shorter than 100 bp after trimming were discarded, and alignments including sequences from at least ten species were concatenated. The resulting alignment was used to infer a maximum likelihood tree with Phyml (Guindon and Gascuel 2003), using a GTR + G + I nucleotide substitution model, which was identified as the best model using the Smart Model Selection (Lefort et al. 2017). Support was evaluated by 100 bootstrap pseudoreplicates.

Identification of homologs and grass co-orthologs encoding C_4-related enzymes

For each gene family that encodes enzymes related to the C_4 pathway (identified based on the literature; Mallmann *et al.*, 2014; Li *et al.*, 2015), homologous CDS were retrieved from three publicly available genomes (*Setaria italica, Sorghum bicolor and Arabidopsis thaliana*), based on the annotation and previously inferred homology (Vilella *et al.*, 2009). The same approach was used to analyse genes of the photorespiration pathway, which are expected to be downregulated during C_4 evolution (Mallmann et al 2014). CDS from the sequenced transcriptomes or retrieved from the literature that were homologous to any sequence in each gene family were identified via BLAST searches. Positive matches with a minimal e-value of 0.01 and minimal mapping length of 500bp were retrieved and added to the datasets. Only the first transcript model was considered for complete genomes, and the longest CDS from each set of Trinity gene isoforms was used.

A new alignment was produced for each gene family ensuring high quality alignments while maintaining as many sites as possible. This approach requires manual curation, and was consequently not used for the 581 sets of core orthologs described above. A preliminary alignment was obtained for each gene family using MUSCLE (Edgar 2004). The alignment was manually inspected in MEGA version 6 (Tamura *et al.* 2013), and potential chimeras and sequences of ambiguous homology (false positives) identified through visual inspection and comparison with other sequences were removed. The remaining sequences were re-aligned as codons using ClustalW (Thompson et al. 1994), and the alignments were manually refined. For each gene family, the alignment was used to compute a maximum likelihood phylogenetic tree, using PhyML (Guindon & Gascuel, 2003), and the GTR + G + I substitution model as best-fit model identified previously for most of the gene families in this study (Christin

et al. 2015). Support values were evaluated with 100 bootstrap pseudoreplicates.

Groups of grass co-orthologs, which include all the genes that descend from a single gene in the last common ancestor of grasses through speciation and gene or genome duplications (including the ancient polyploidy in the common ancestor of grasses; Tang et al. 2010), were identified based on the phylogenetic trees inferred for each gene family. Duplicates specific to some groups of grasses, which might have emerged via gene or genome duplication (whether via auto- or allopolyploidy) after the diversification of grasses, would be grouped in the same co-orthologs, so that our orthology assessment and subsequent expression analyses are not influenced by polyploidization events. Cleaned reads were mapped back to sequences belonging to any of the gene families as single reads, using the local alignment option in Bowtie2 (Langmead & Salzberg, 2012). Our approach allows reads to map back to sequences from the same species, but also allows sequences from other closely related species to serve as the reference. The number of reads mapped to each group of co-orthologs was reported as reads per kilobase of aligned exons per million of cleaned reads (rpkm). These proxies for transcript abundances were obtained for each replicate.

Identification of co-opted genes and factors increasing co-option rates

Enzymes of the C_4 pathway are abundant in the leaves of C_4 species because high catalytic rates are needed to match the fluxes of the Calvin cycle (Furbank et al. 1997, Mallmann et al. 2014). Transcripts encoding enzymes known to act in the C_4 pathway were consequently identified as those that reached an abundance of at least 500 rpkm in leaves of a given C_4 species. Because this threshold is arbitrary, subsequent analyses were repeated with other thresholds (300, 1000, 1500 rpkm), which did not affect our conclusions (see Results). Previous investigations comparing a limited number of species have shown that, within a given taxonomic group, independent C_4 origins tend to co-opt the same gene lineages (Christin et al. 2013, 2015; Emms et al. 2016). To test this expectation across our larger species sample, the number of genes co-opted at least once in our dataset was compared to the number expected by chance given the size of the different gene lineages and the number of co-option events, following the resampling approach of Christin et al. (2015).

Once a bias in gene co-option was confirmed (see Results), we tested for factors

potentially affecting the probability of a given group of co-orthologs being co-opted for C_4. We used the values inferred for the last common ancestor of grasses as proxies for the condition before C_4 evolved, with two different dimensions of the expression patterns. First, we inferred the leaf transcript abundance. Second, we inferred the leaf/root ratio of abundances as a proxy for leaf specificity. For each group of co-orthologs, the values of these variables in the common ancestor of grasses were estimated using the phylogeny obtained with HaMSTR and the 'ace' function in the R package ‘ape’ version 3.5 (Paradis et al. 2004). The maximum likelihood method was selected, with a Brownian motion model. In this approach, the value of the continuous variable that maximizes the likelihood is calculated for each node, with the associated confidence intervals. Only non-C_4 species were included in the ancestral state analyses to avoid biases caused by high levels in C_4 taxa. Considering only the gene families co-opted at least once, linear models, as implemented in the ‘lm’ function in R version 3.3.2 (R Development Core Team 2016), were used to test independently for an effect of ancestral leaf transcription abundance and of ancestral leaf/root ratio on the number of times each group of co-orthologs has been co-opted. An analysis of variance on multiple linear models was then used to determine whether the effect of ancestral leaf abundance and/or leaf/root ratio remain when the gene family was included as a co-factor.

Transcriptome datasets available for groups of closely related C_3 and C_4 species outside of grasses were used to assess whether the observed patterns are valid across flowering plants. Data for one C_3 and one C_4 Cleomaceae were retrieved from Bräutigam et al. (2011), and the phylogenetic annotation of C_4-related genes in these datasets was deduced from the identity of orthologs from the closely-related *Arabidopsis* and the phylogenetic trees from Christin et al. (2015). For *Flaveria*, RNAseq data were retrieved for two C_3 species from Mallmann et al. (2014) and for one C_4 species from Lyu et al. (2015). The reads were annotated in the original study based on their similarity to *Arabidopsis* sequences, but the evolutionary distance between *Flaveria* and *Arabidopsis* can potentially mislead orthology assessments. We consequently performed *de novo* assemblies using the published reads, and obtained the transcript abundance for C_4-related genes using the previously published phylogenetic annotation pipeline (Christin et al. 2015). Groups of co-orthologs co-opted for C_4 by *Flaveria* or Cleomaceae were identified based on the literature (reviewed in Christin et

al. 2015) or based on leaf abundance reaching 500 rpkm in C_4 species for the genes not included in previous reviews. The effect of the abundance in the C_3 relatives on the co-option probability was modelled as for grasses, independently for Cleomaceae and *Flaveria*. Because two C_3 species are available for Flaveria, their average abundance was used. Root abundance was not available for the same species, so that the effect of leaf specificity in these groups of eudicots could not be tested.

Positive selection tests

Codon models were used to test for positive selection following the co-option of genes for C_4 photosynthesis. For each group of co-orthologs that has been co-opted at least once for C_4, the inferred alignment was truncated as needed to remove poorly aligning ends and a new phylogenetic tree was inferred with phyML, considering only 3rd positions of codons to remove potential biases due to adaptive evolution. The inferred topology was used to optimize three different codon models, using codeml as implemented in PAML (Yang 2007). These models rely on the ratio of non-synonymous mutation rate per synonymous mutation rate (ω; Yang and Nielsen 2002, 2008; Yang and Swanson 2002). In the null model M1a, codons evolve under either purifying or relaxed selection in all branches (ω smaller than and equal to one, respectively). In the branch-site models, some codons still evolve under neutral or purifying selection in all branches, but others shift from purifying or relaxed selection in background branches to relaxed (in model A) or positive (in model A1) selection in foreground branches. These foreground branches are defined *a priori*. In our case, all branches descending from each C_4 co-opted gene (identified above for the species sequenced here and from the literature for the rest of species) were set as the foreground branches. Because genes for β-carbonic anhydrase (βCA) were present at similar abundance in non-C_4 and C_4 species (see Results), but these are known to be part of the C_4 pathway (Budde et al.,1985; Hatch and Burnell, 1990), all branches leading to C_4 species in these gene families were selected as foreground branches. The fit improvement of the model assuming changes in selection pressures was evaluated using likelihood ratio tests (LRT). The model A1 was first compared to the model M1a, to test for selective shifts following the co-option event, and then to the model A to specifically test whether the shift corresponded to positive selection. P-values were corrected for multiple testing.

Acknowledgements

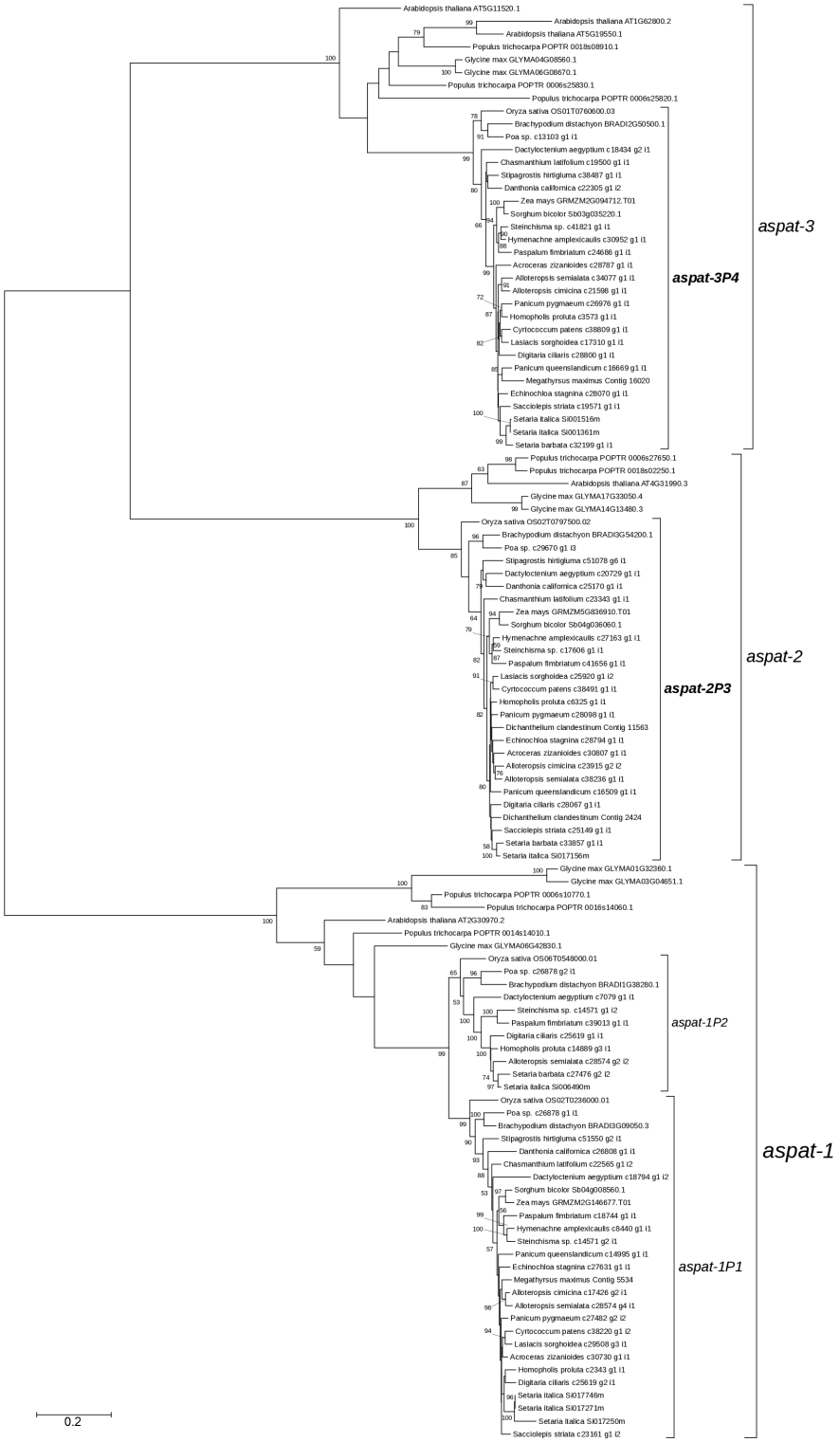
This work was supported by the Royal Society (grant numbers RG130448, URF120119), and the Natural Environment Research Council (grant number NE/M00208X/1). All sequences generated in this work have been submitted to NCBI Sequence Read Archive and Transcriptome Shotgun Assembly repository (BioProject PRJNA395007).

Chapter I: Supporting information

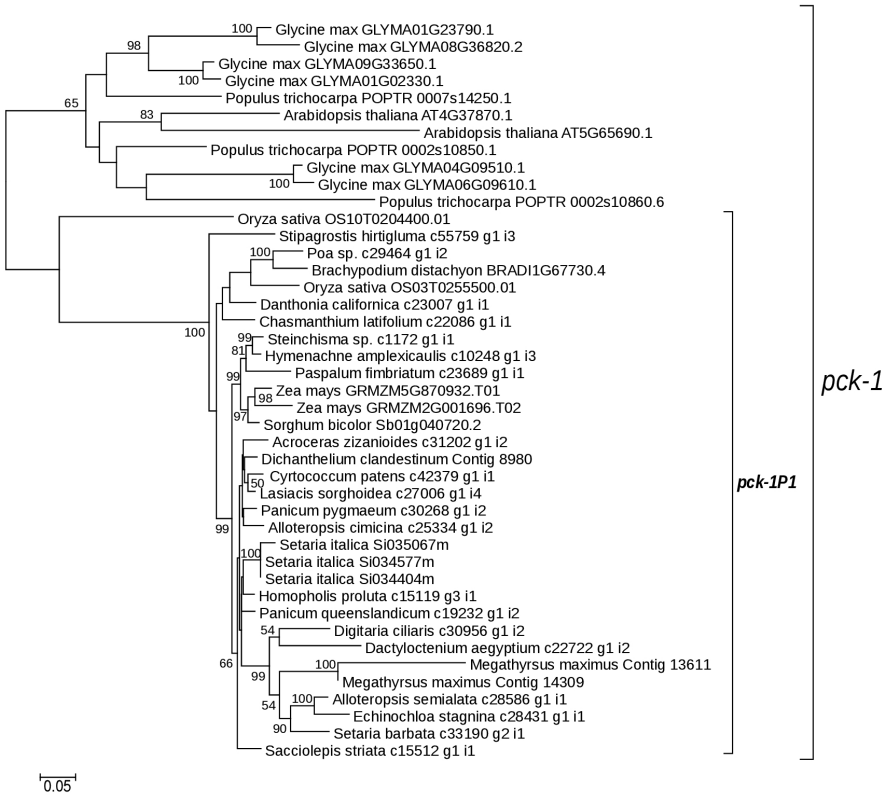
Figure I.S1: Phylogenetic trees for C_4-related gene families.

For each gene family with members potentially involved in the C_4 or photorespiratory pathways, a maximum phylogenetic tree is shown. Name of the enzyme is indicated at the top. Bootstrap support values are indicated near nodes, when higher than 50. Gene lineages are indicated on the right, with brackets.

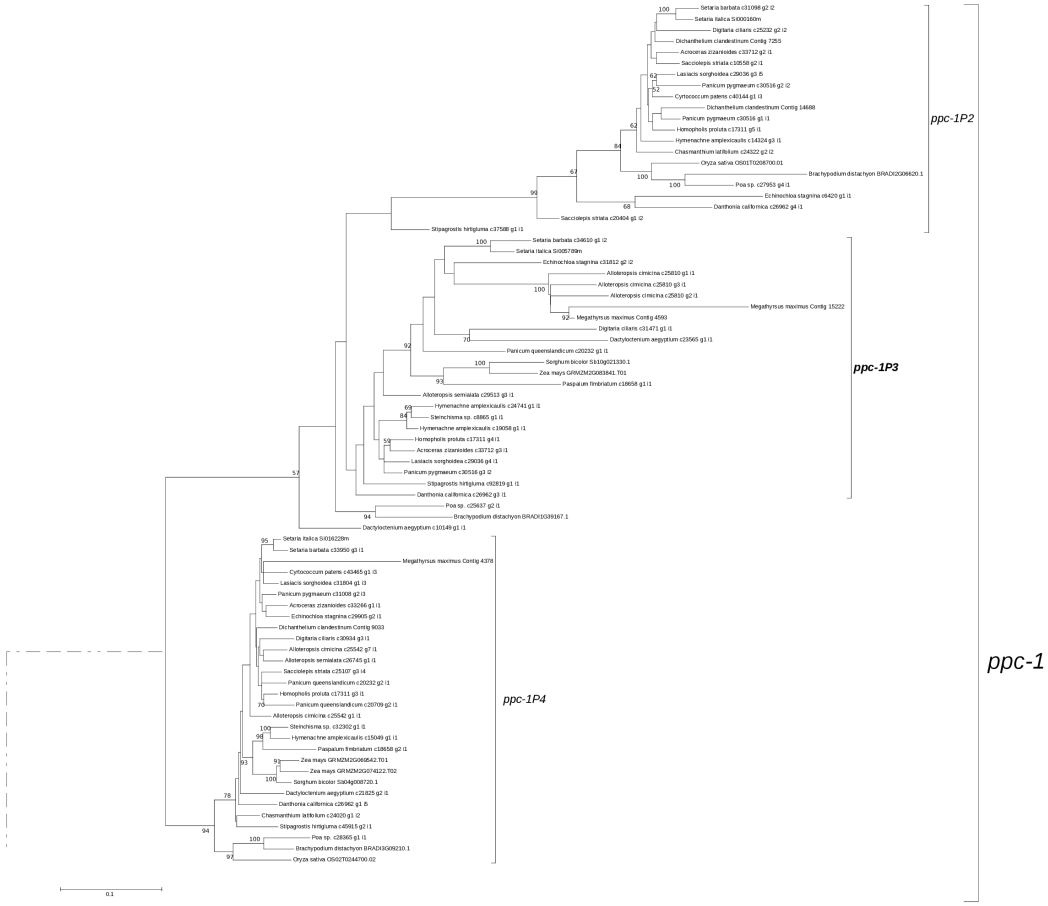
Aspartate aminotransferase (ASP-AT)



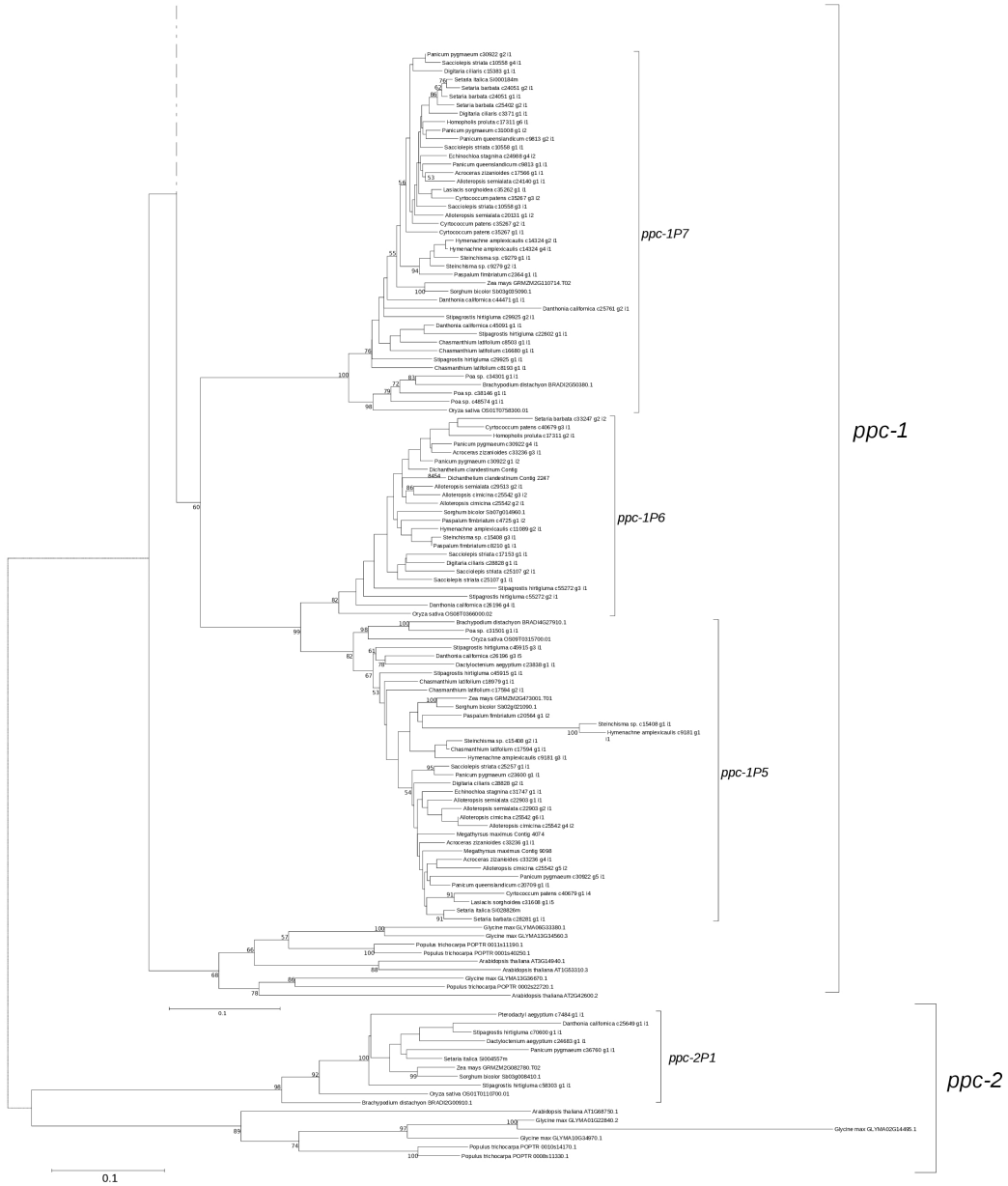
Phosphoenolpyravute carboxykinase (PCK)



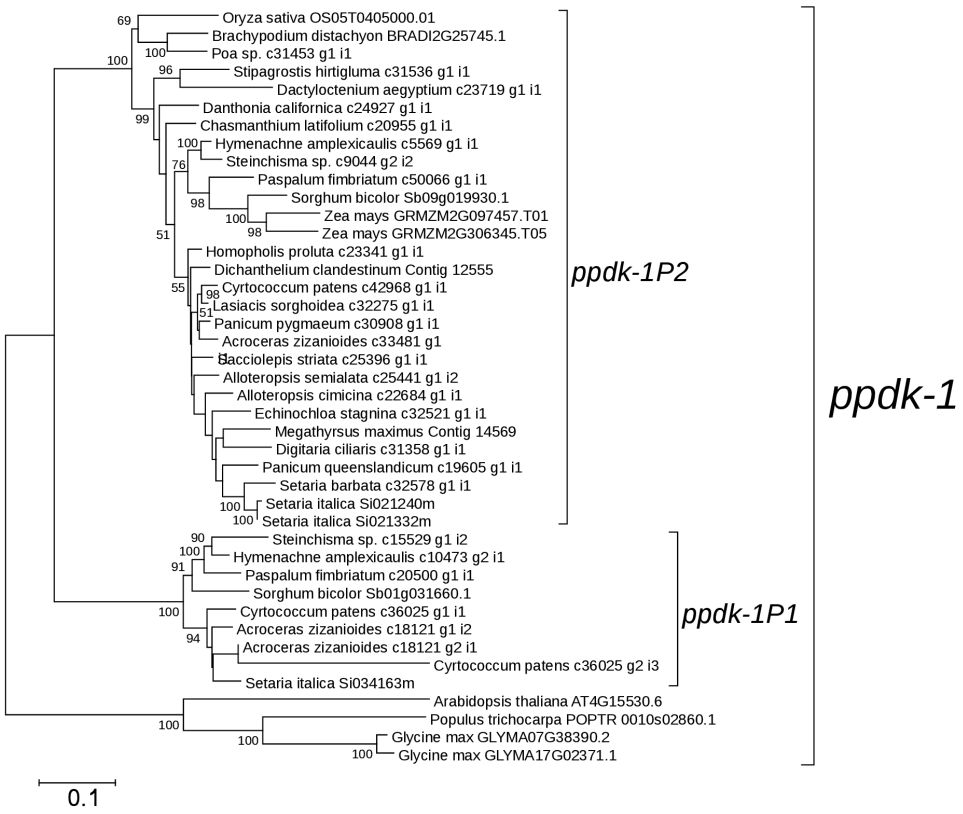
Phosphoenolpyruvate carboxylate (PEPC)



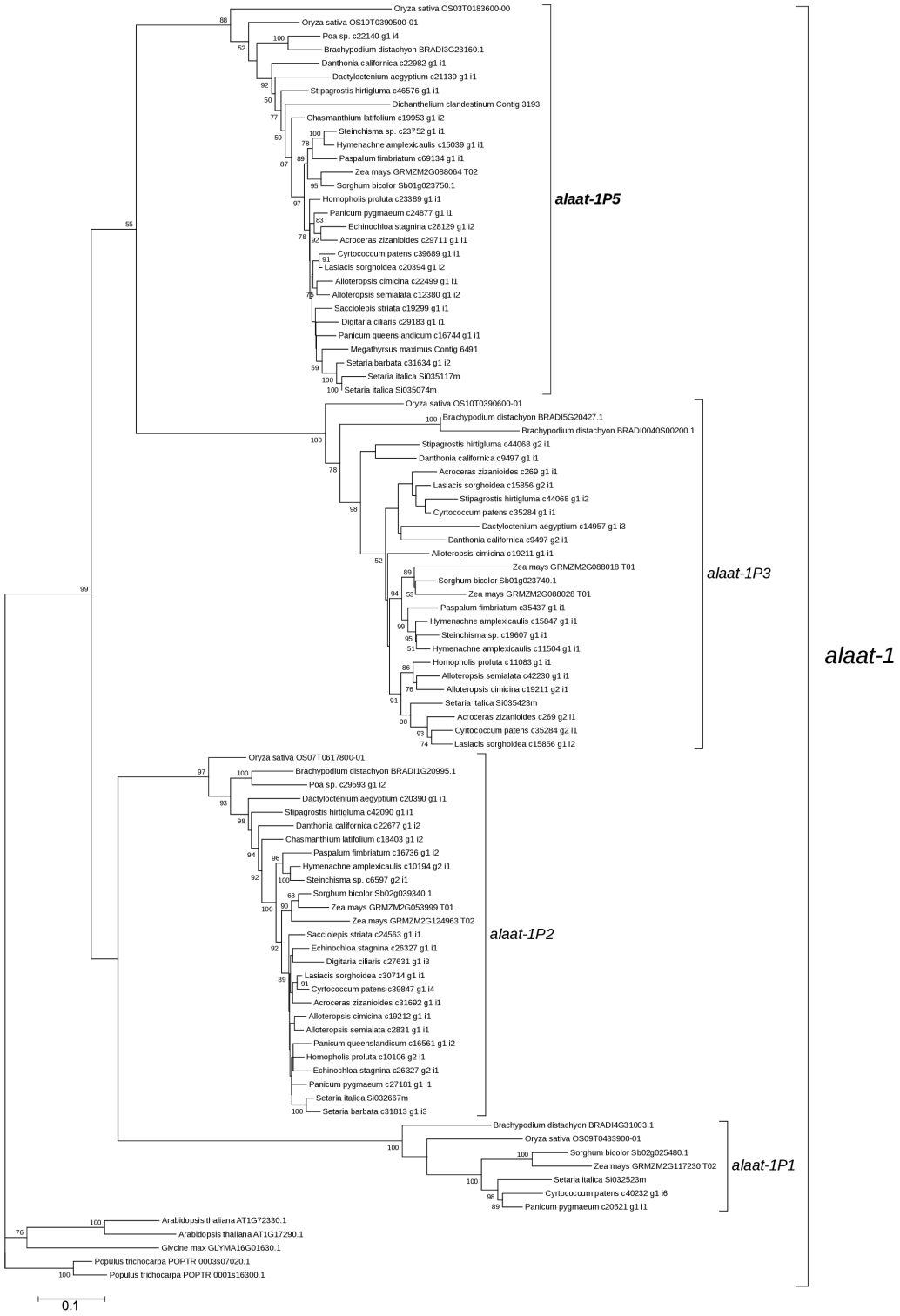
Phosphoenolpyruvate carboxylase (PEPC)



Pyruvate, phosphate dikinase (PPDK)



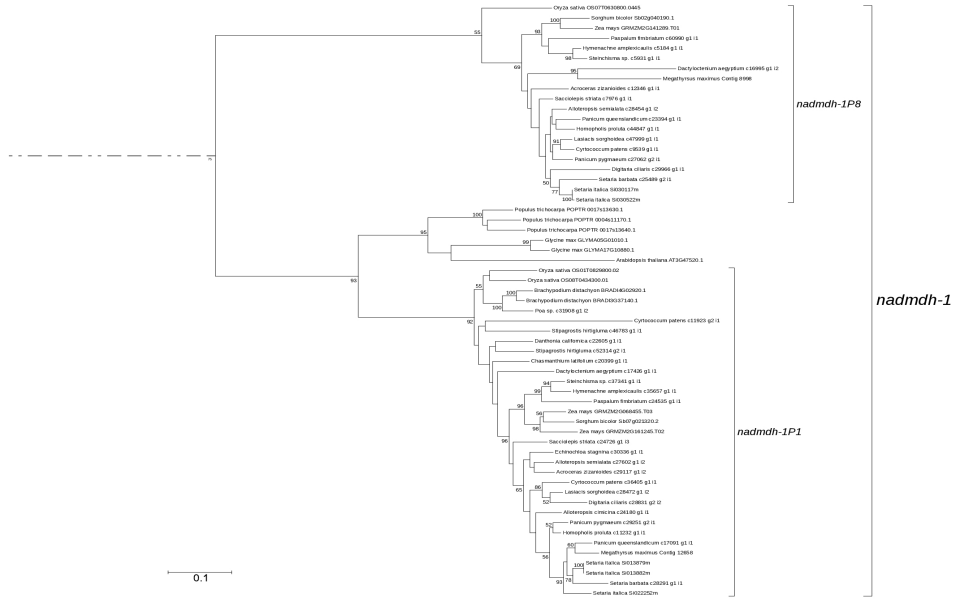
Alanine aminotransferase (ALA-AT)

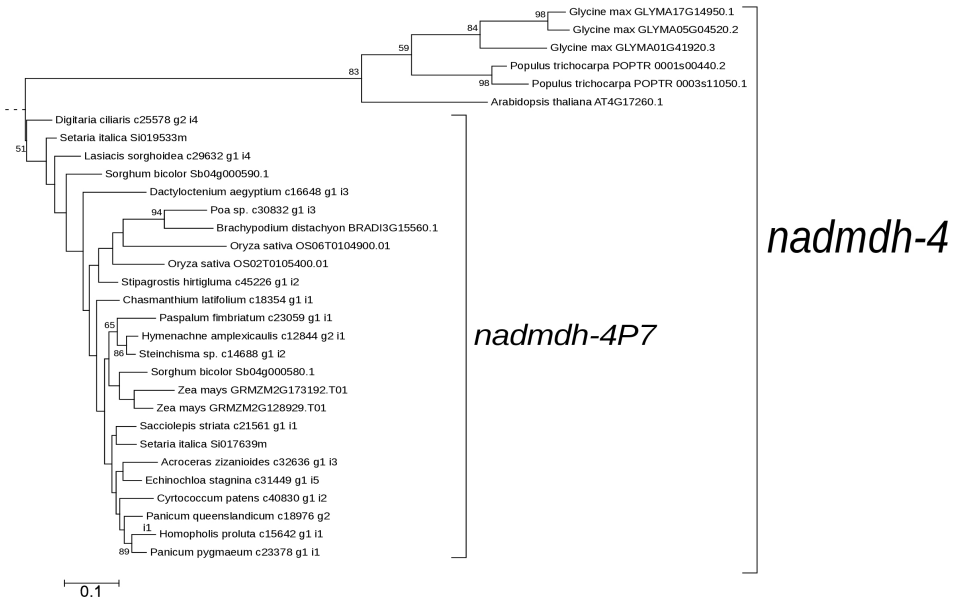


β-carbonic anhydrase (βCA)

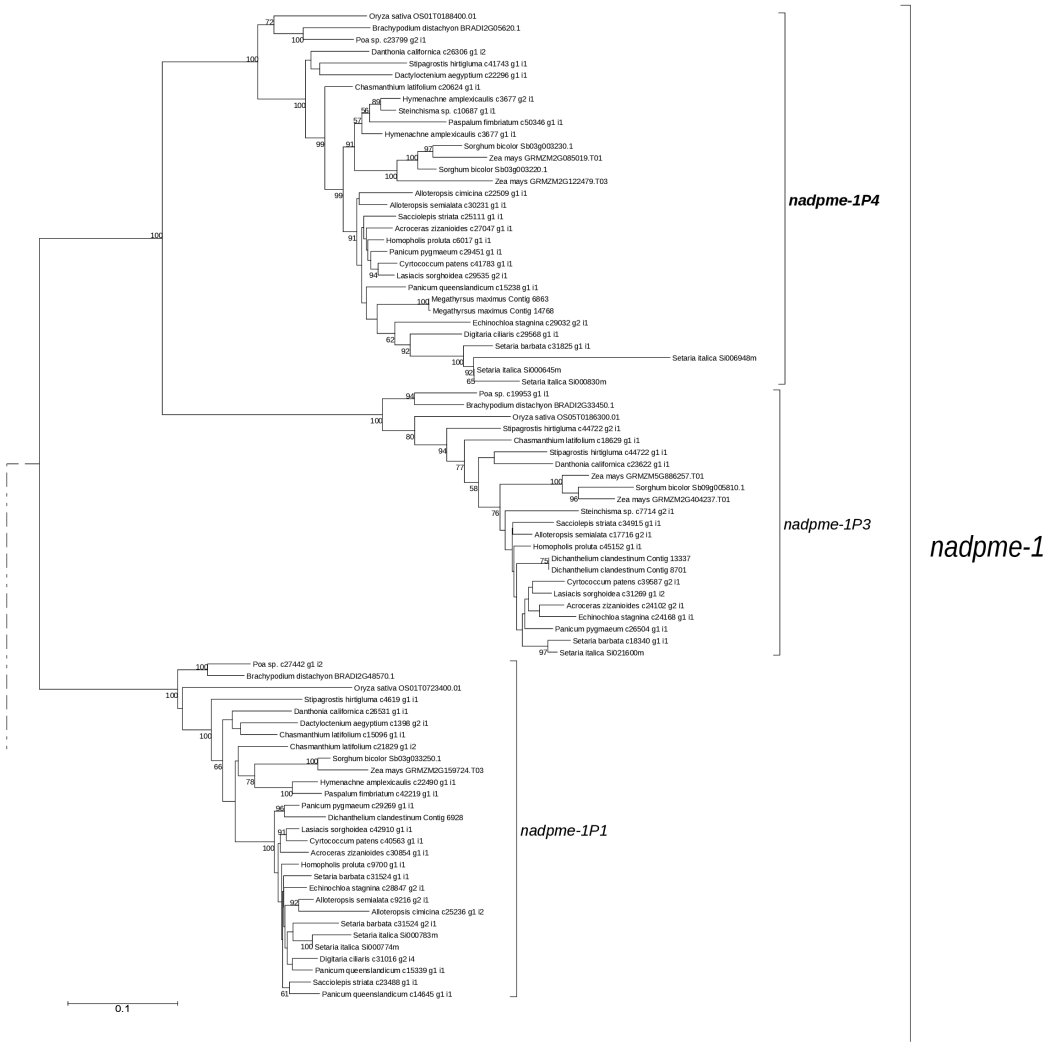
NAD-malate dehydrogenase (NAD-MDH)

NAD-malate dehydrogenase (NAD-MDH)

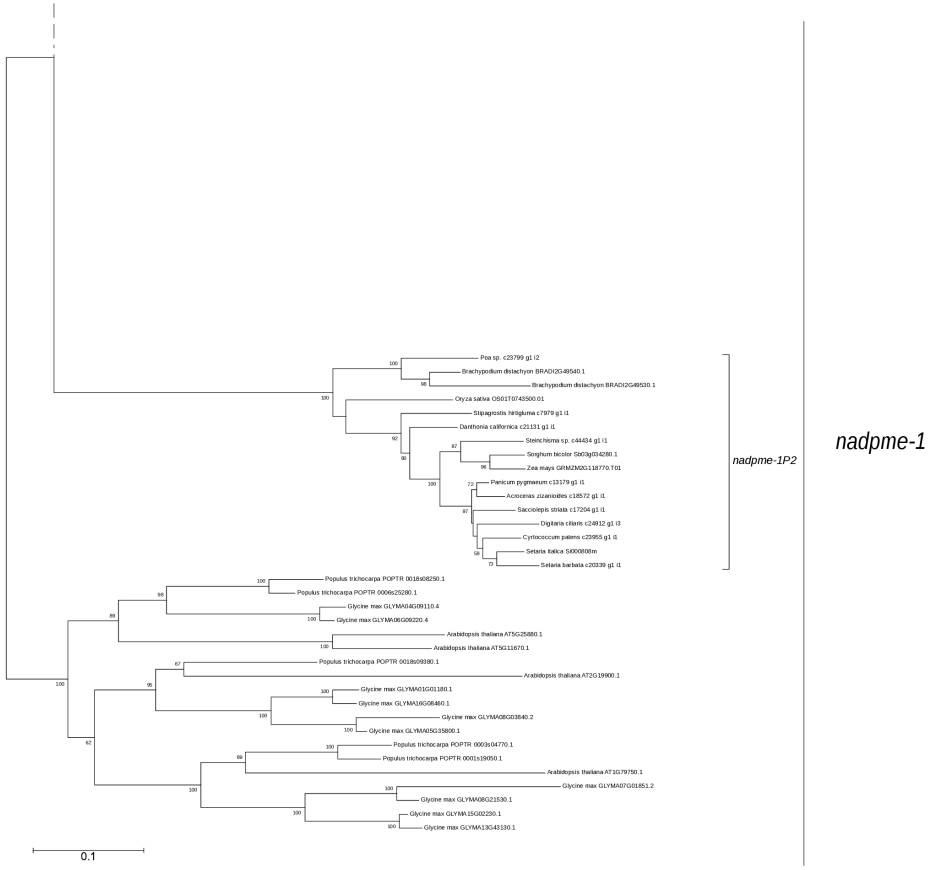




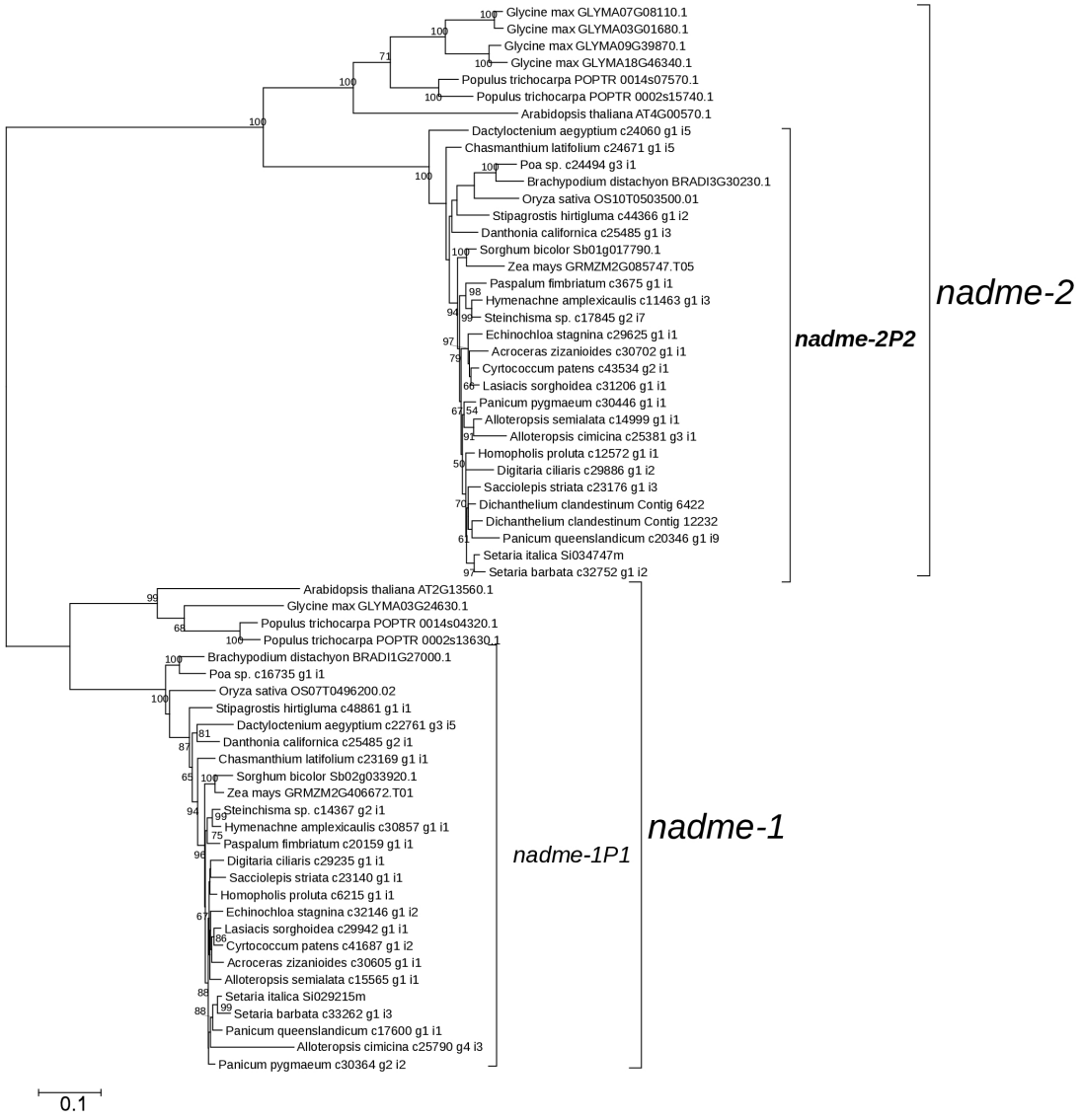
NADP-malic enzyme (NADP-ME)



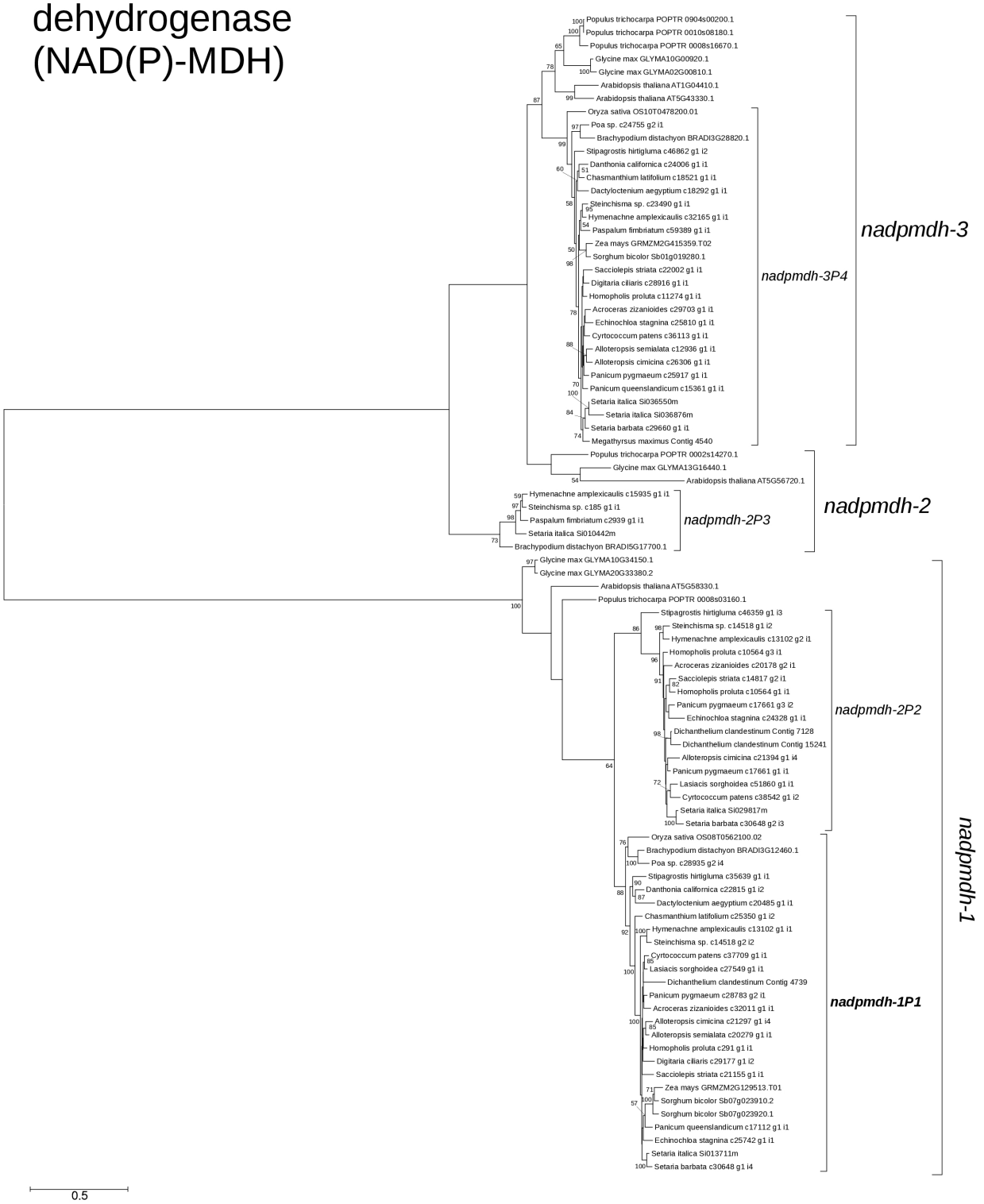
NADP-malic enzyme (NADP-ME)



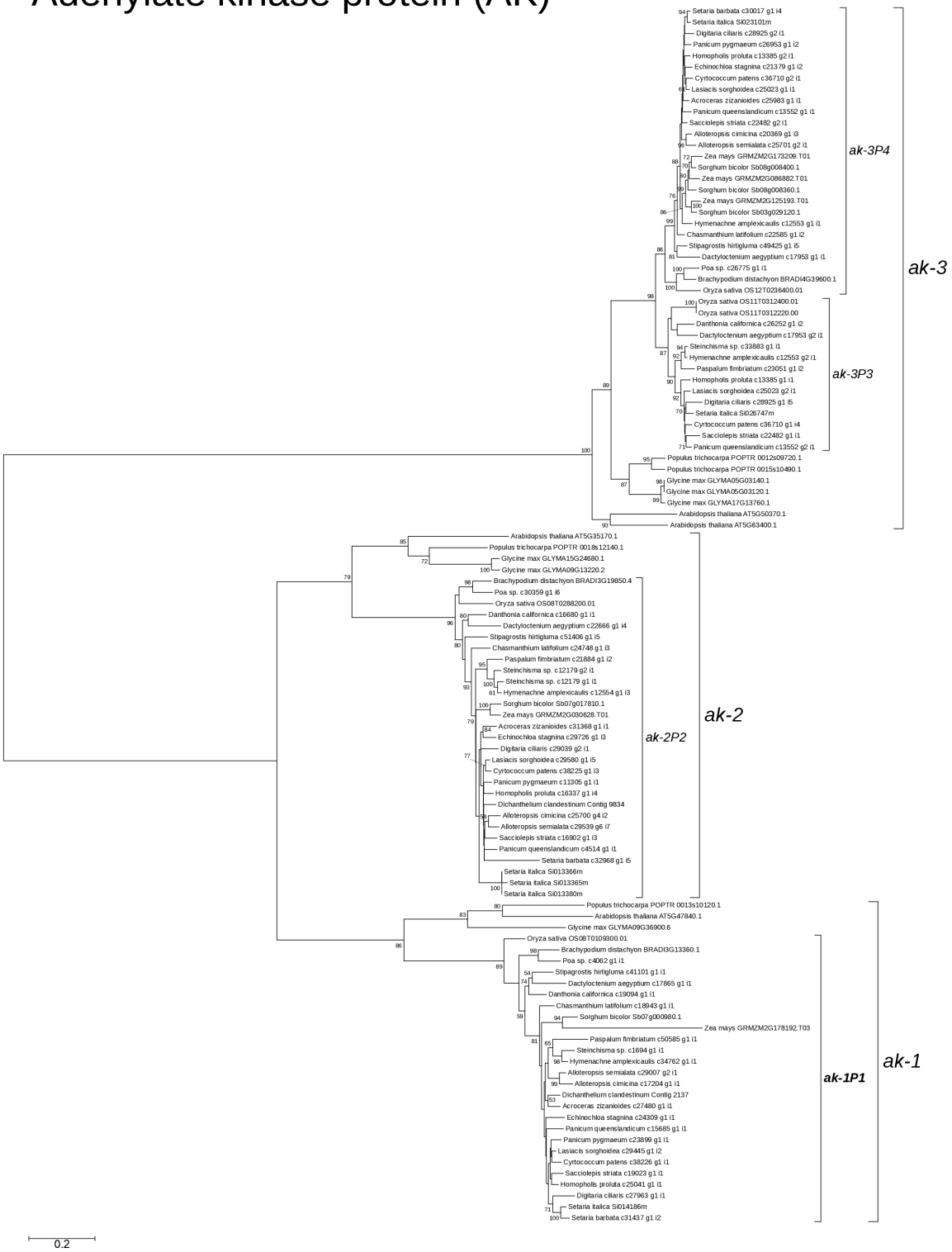
NAD-malic enzyme (NAD-ME)



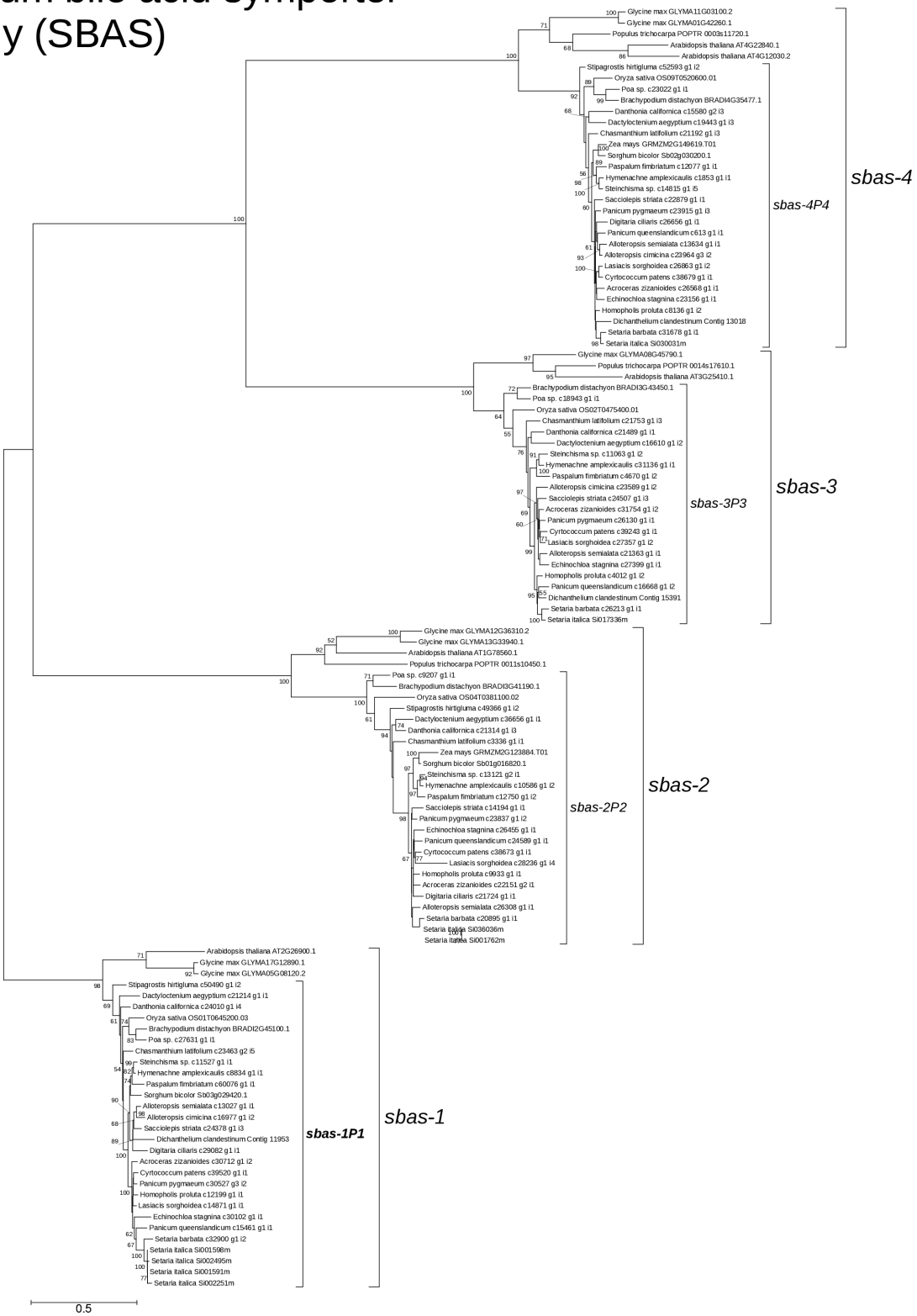
NAD(P)-malate dehydrogenase (NAD(P)-MDH)



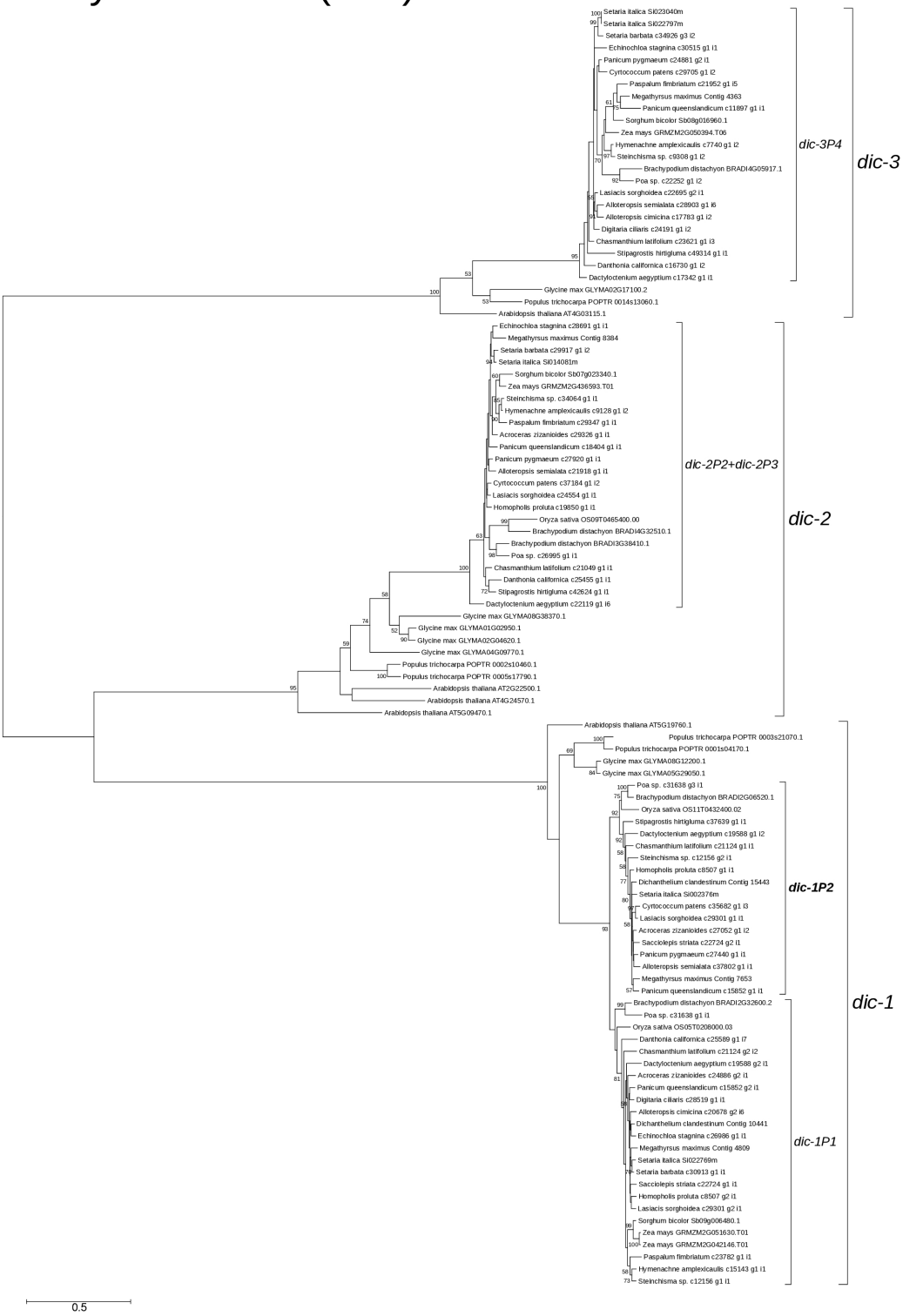
Adenylate kinase protein (AK)



Sodium bile acid symporter family (SBAS)



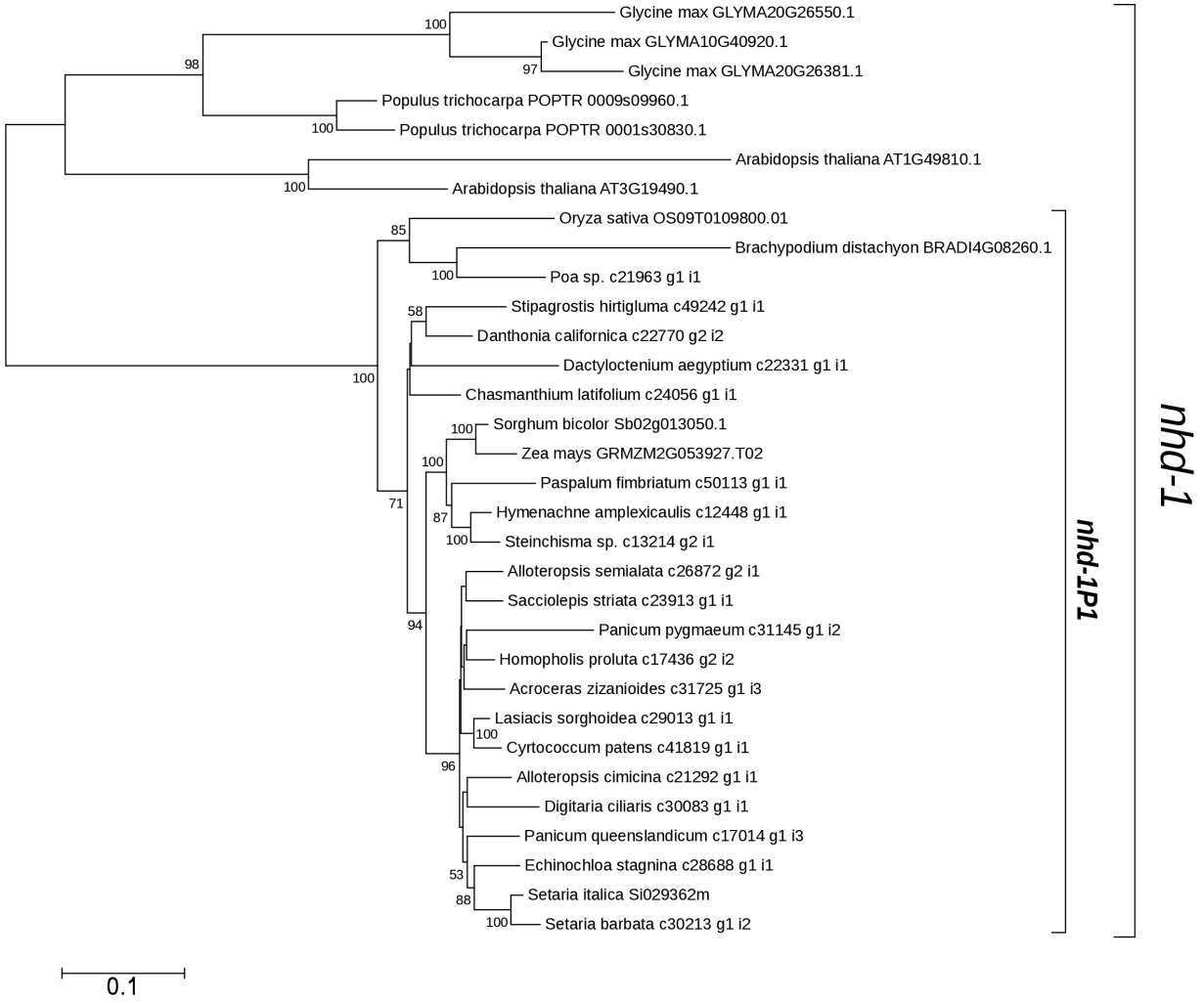
Dicarboxylate carrier (DIC)



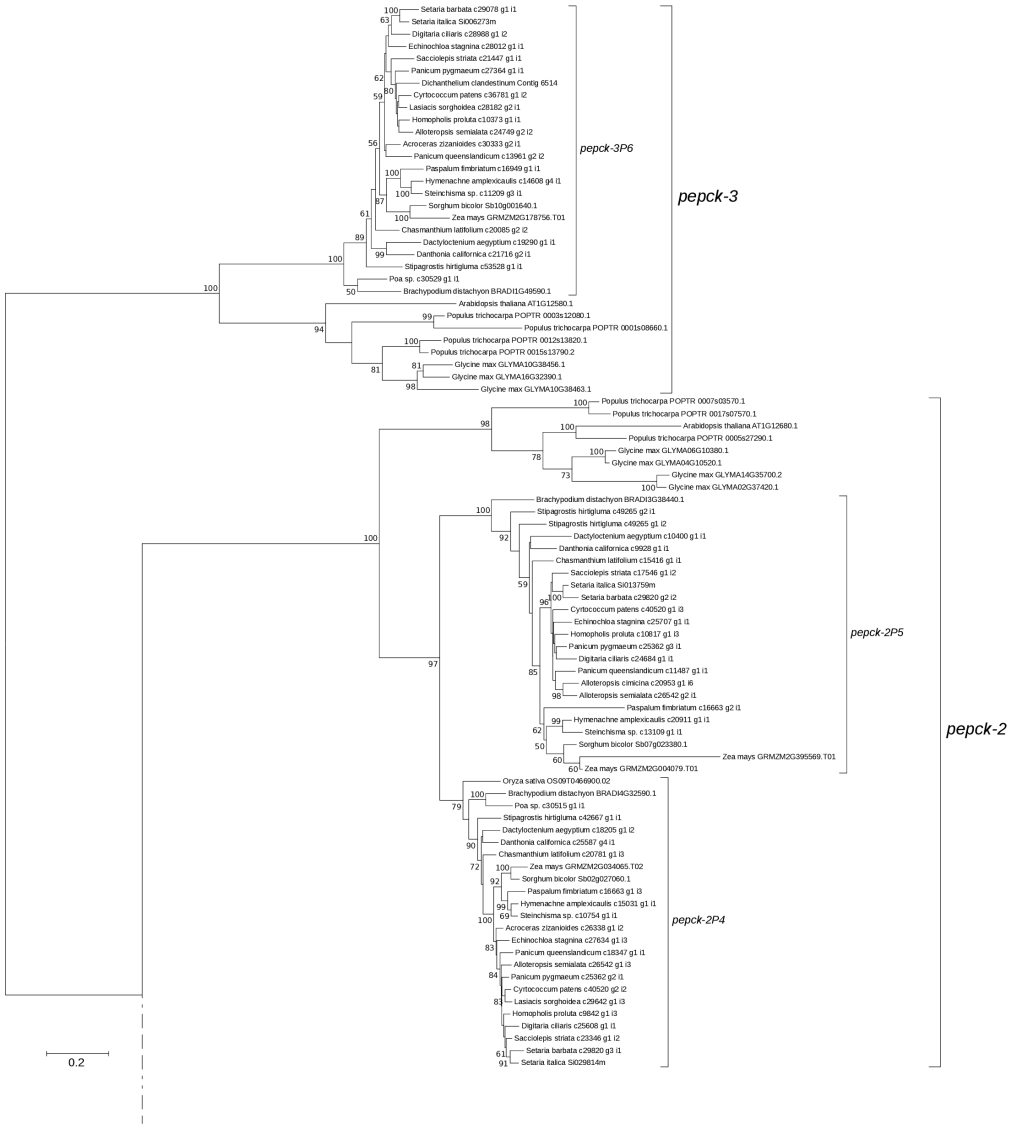
Dicarboxylate transporter (DIT)

Glyceraldehyde-3-phosphate Dehydrogenase (GAPDH)

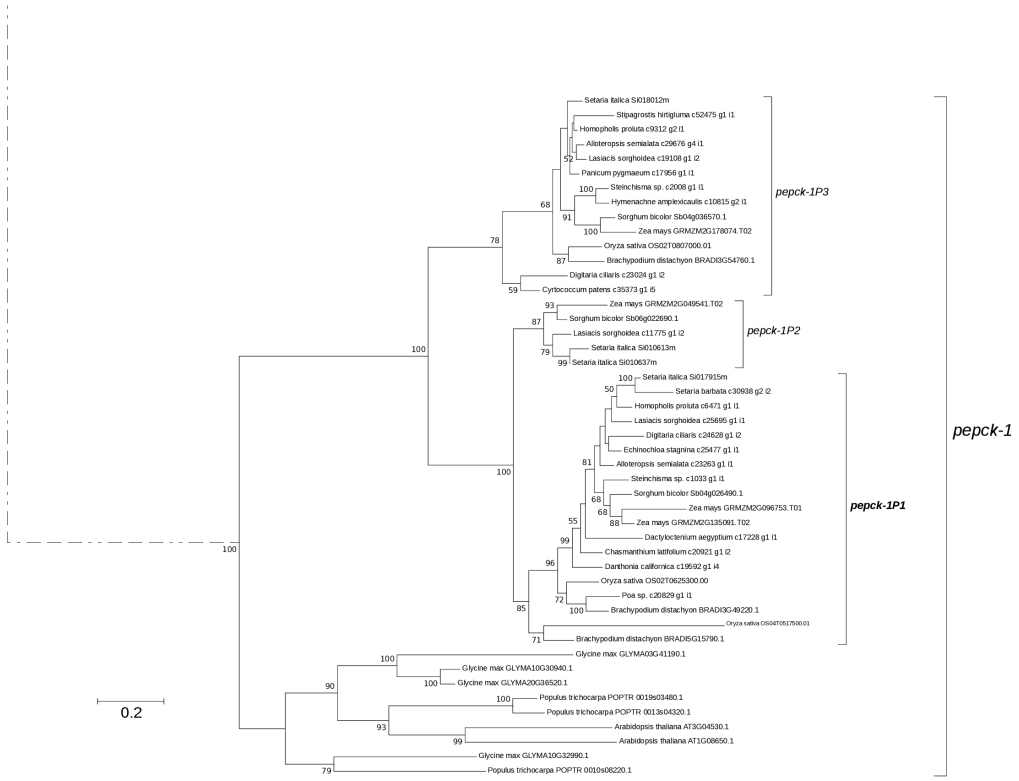
Sodium:Hydrogen antiporter (NHD)



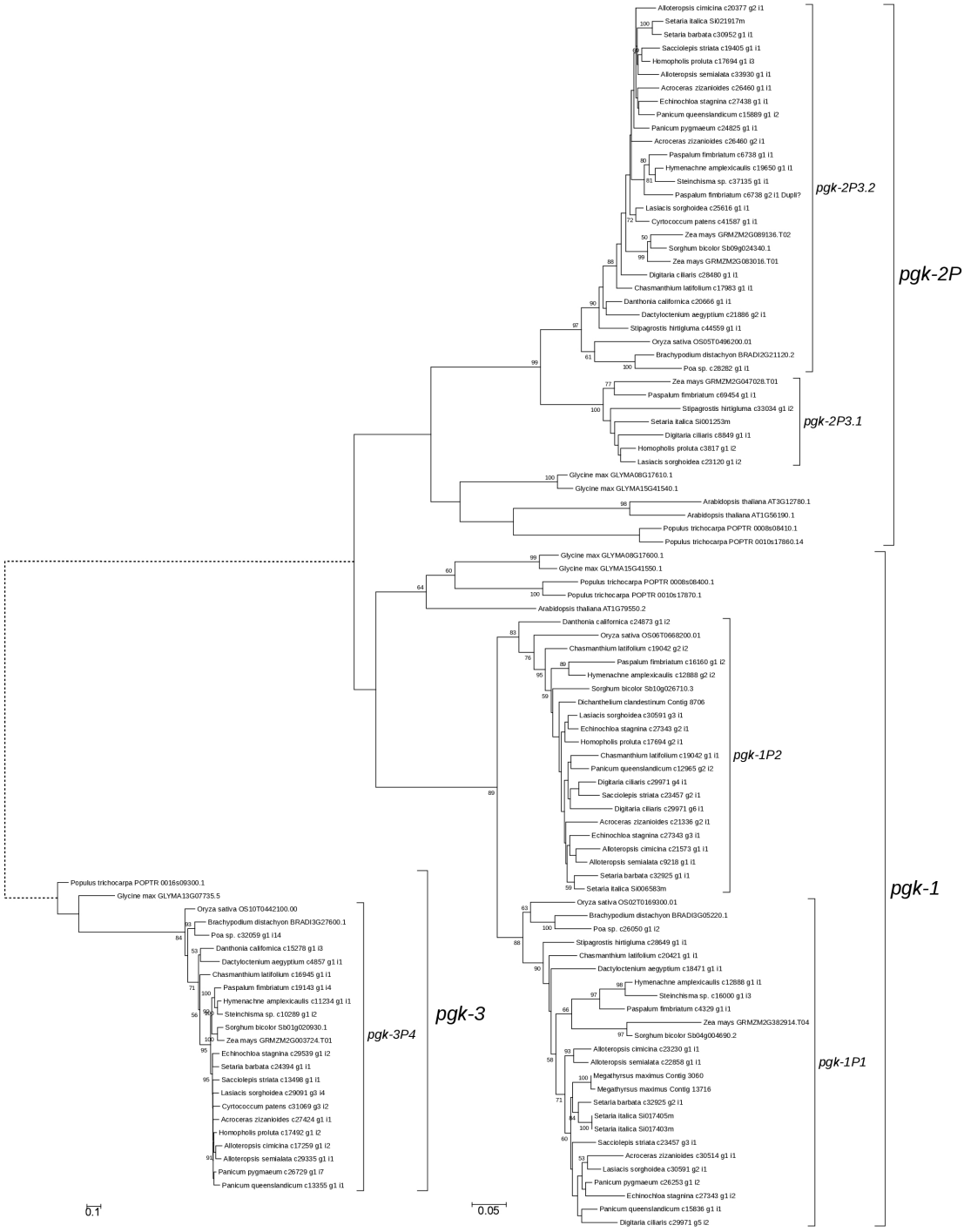
Phosphoenolpyruvate carboxylase kinase (PEPC-K)



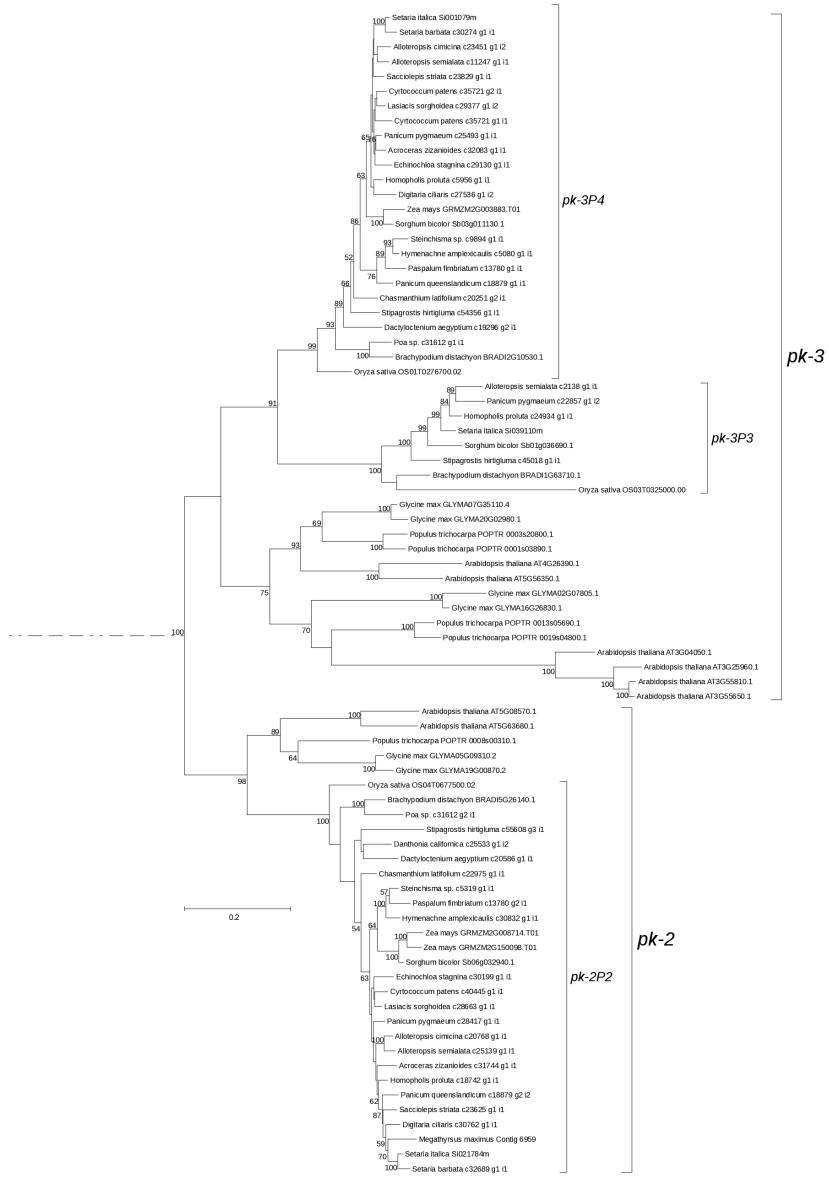
Phosphoenolpyruvate carboxylase kinase (PEPC-K)



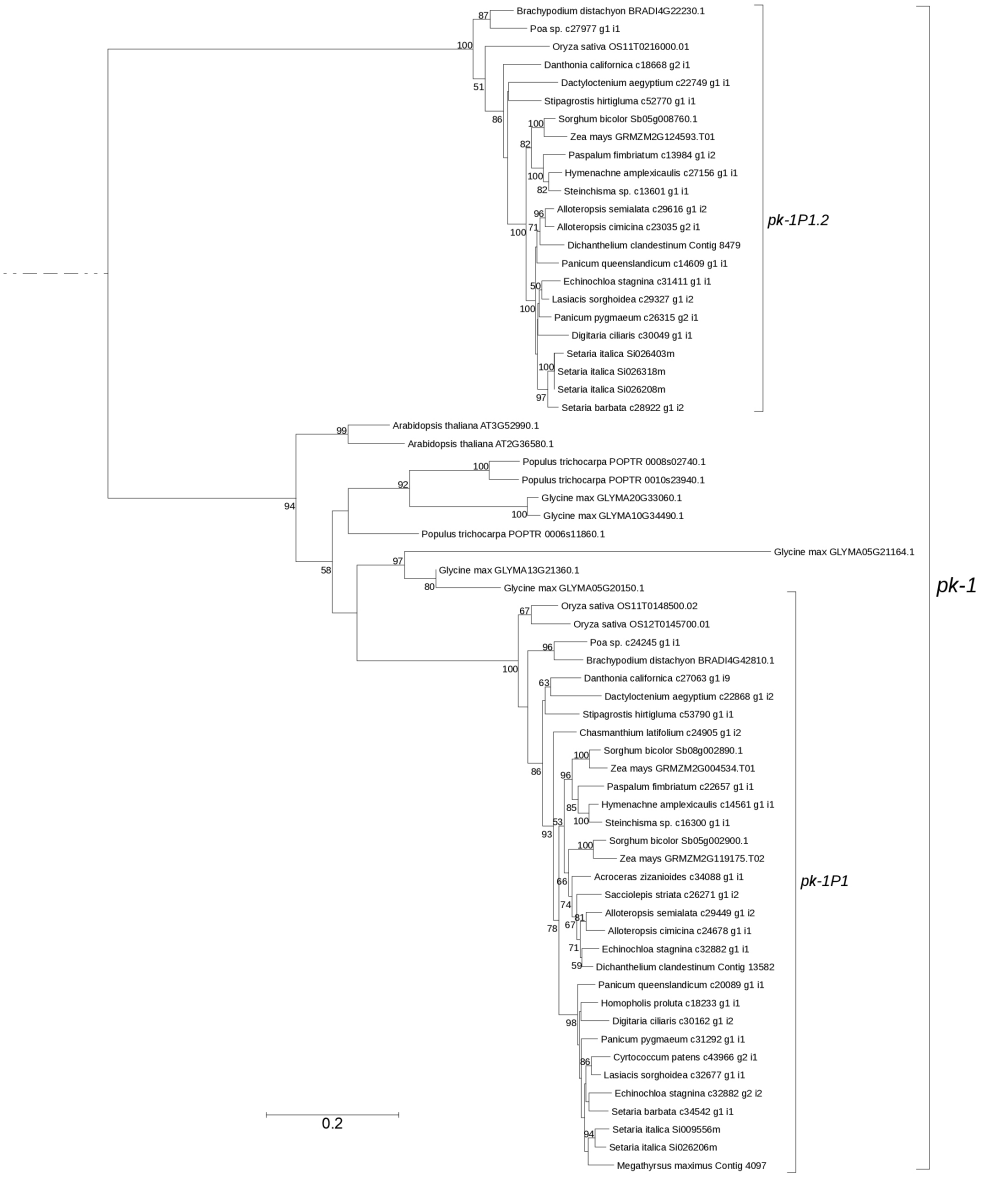
Phosphoglycerate Kinase (PGK)



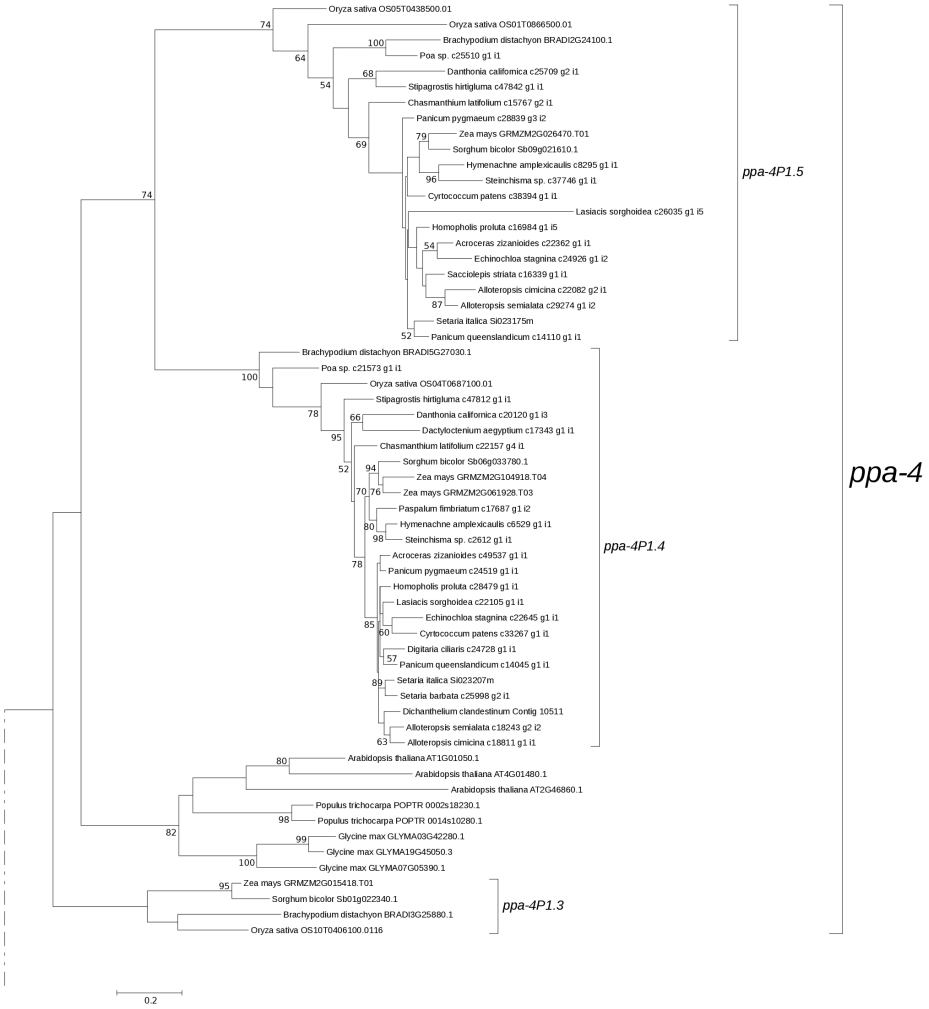
PYRUVATE KINASE (PK)



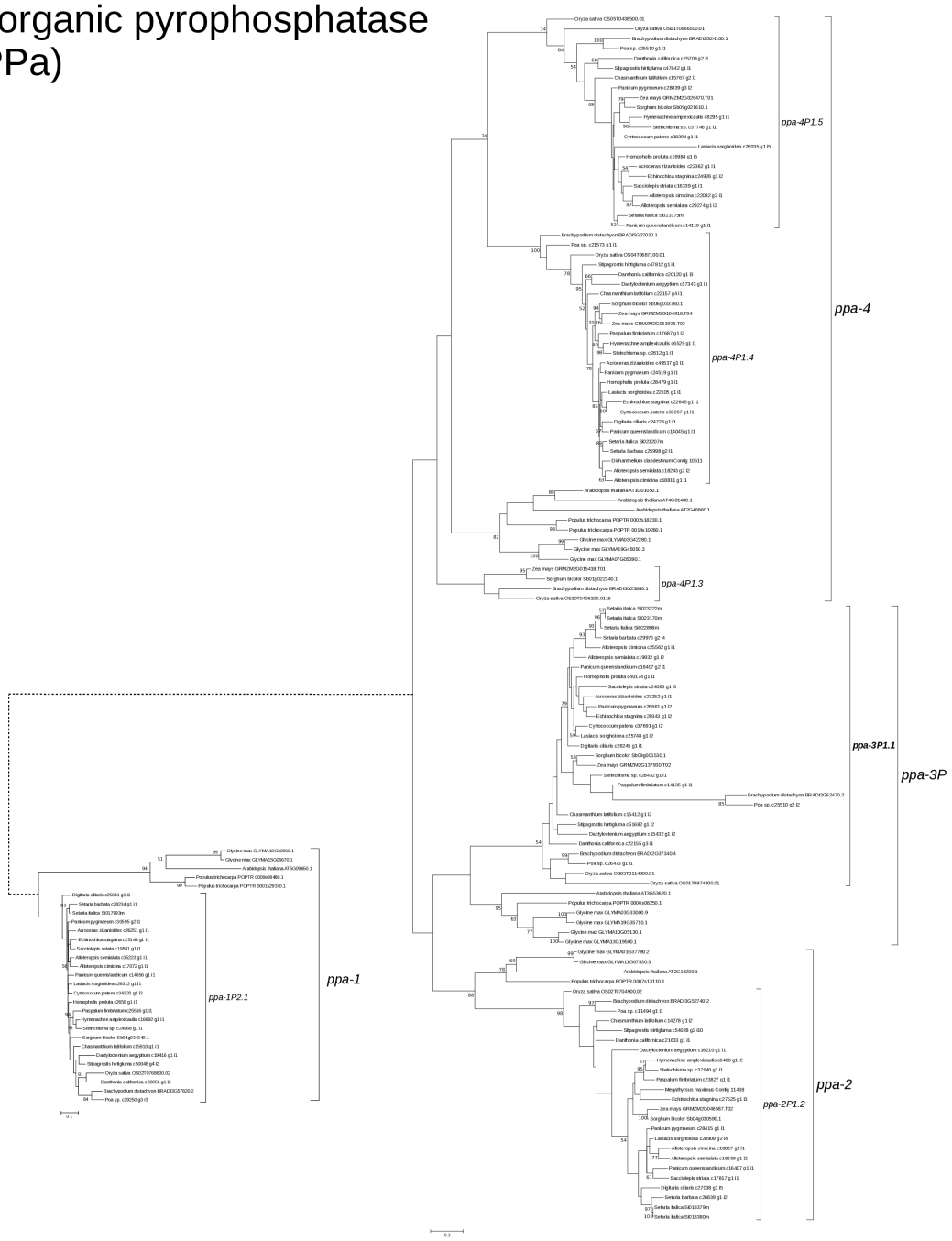
PYRUVATE KINASE (PK)



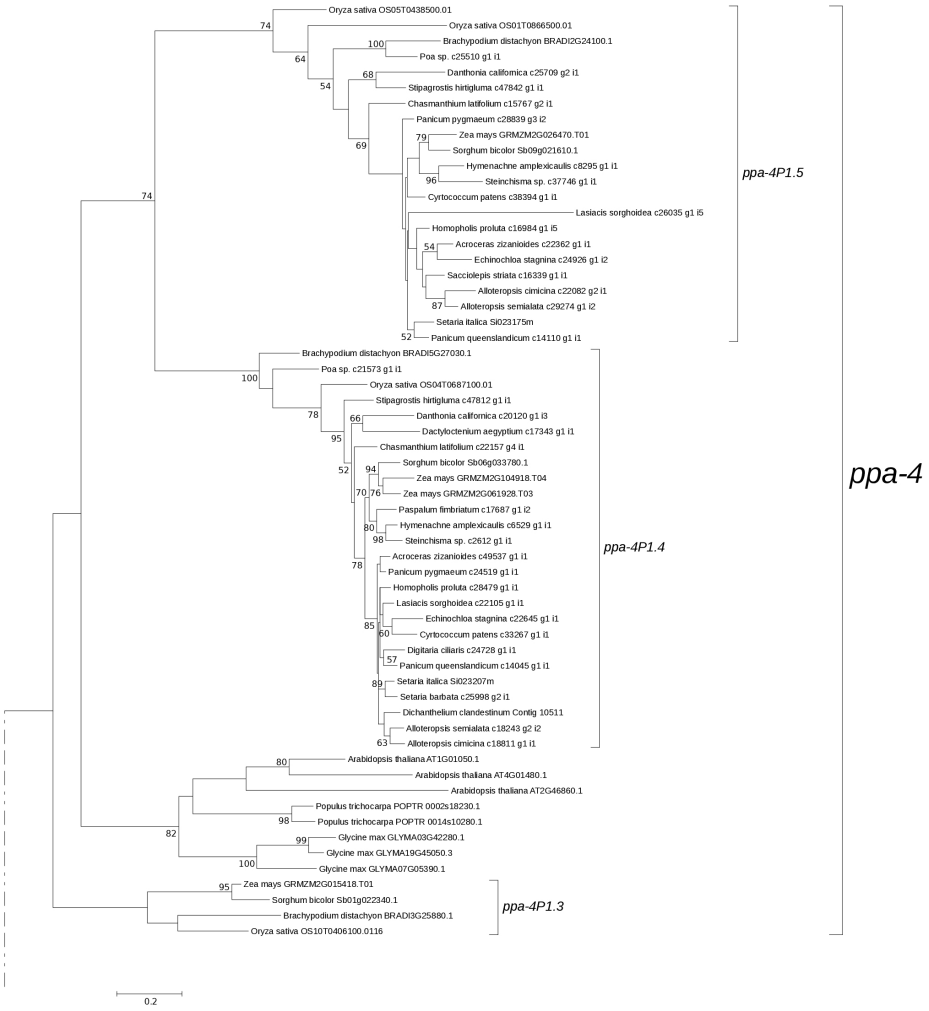
Inorganic pyrophosphatase (PPa)



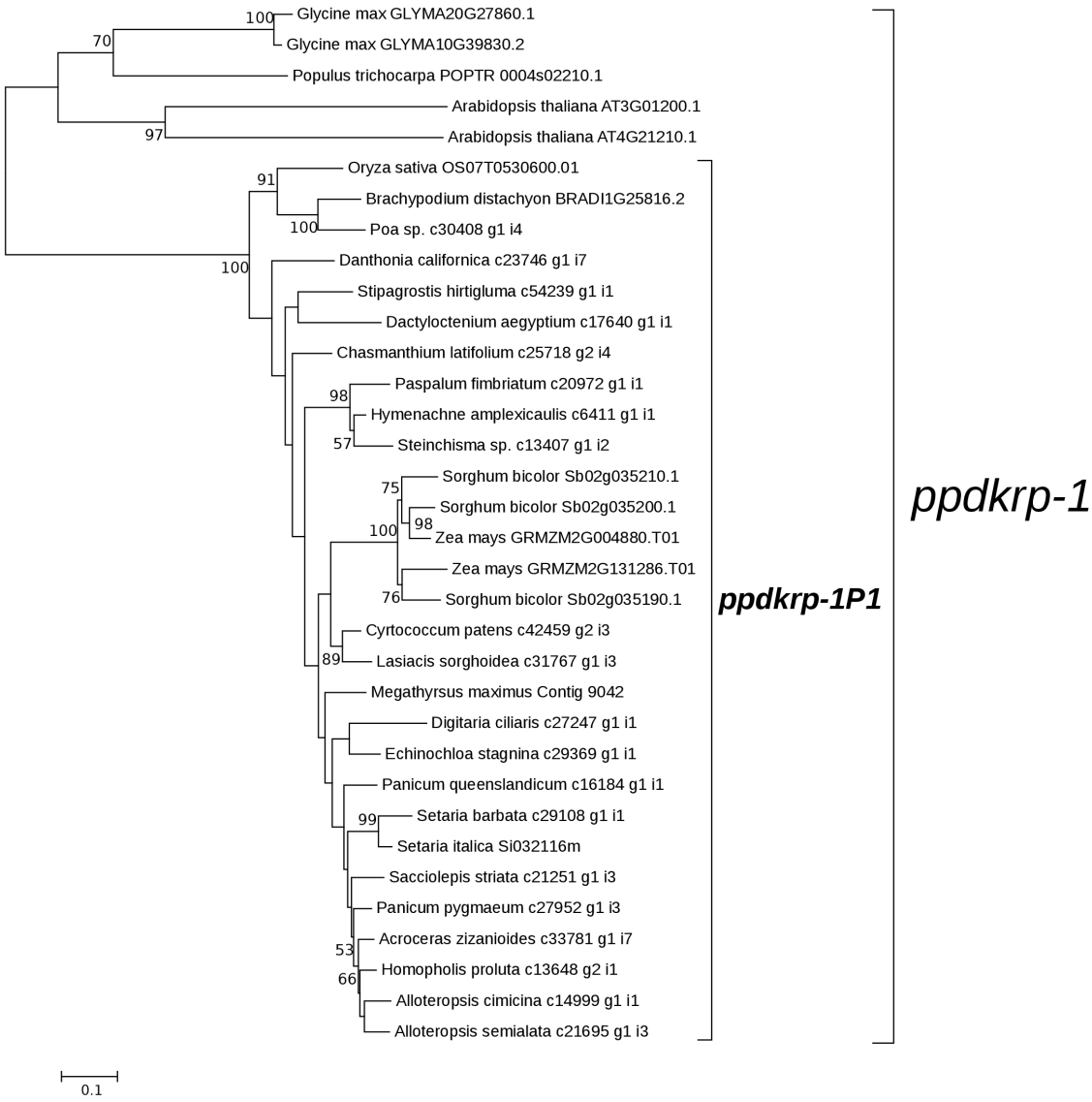
Inorganic pyrophosphatase (PPa)



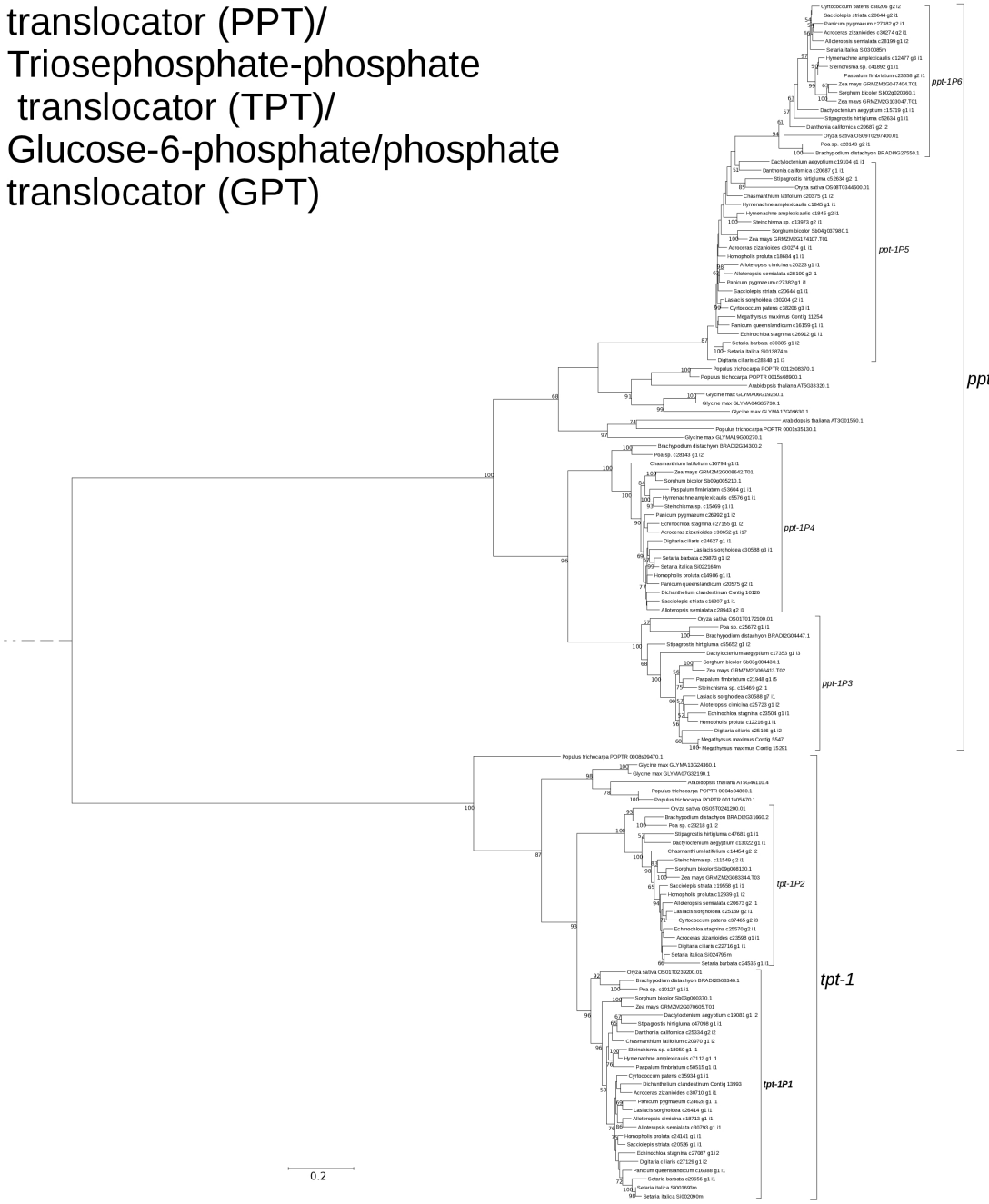
Inorganic pyrophosphatase (PPa)



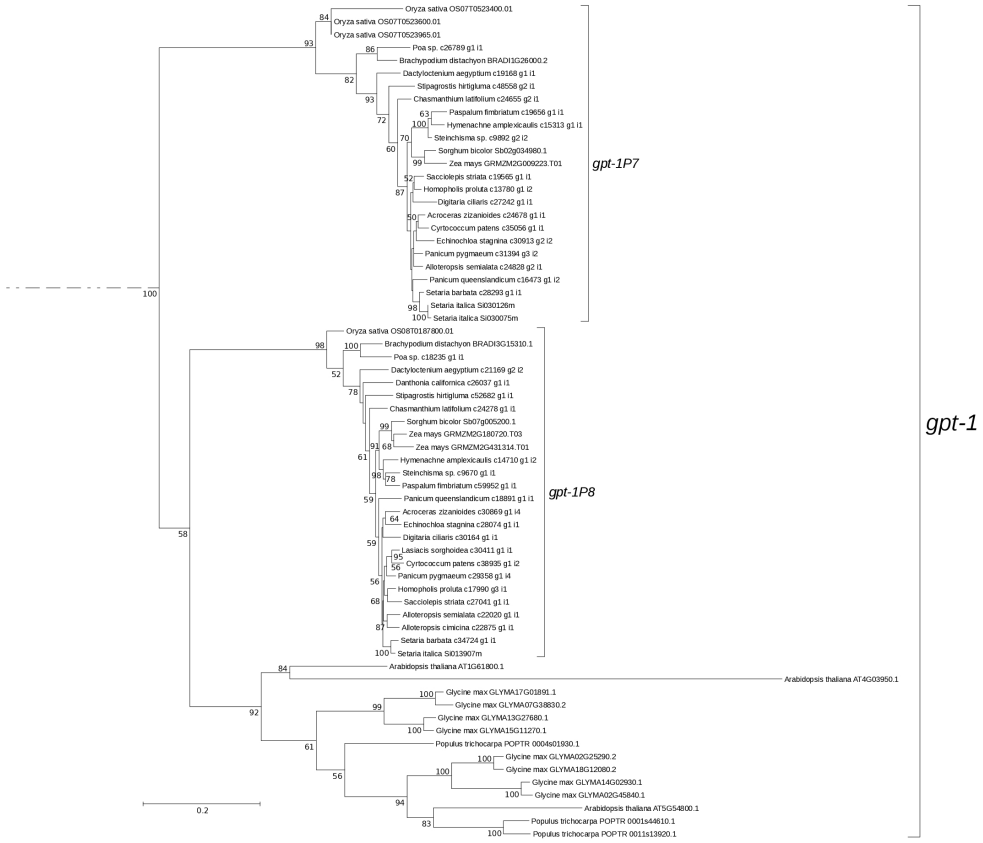
Pyruvate, phosphate dikinase regulatory protein (PPDK-RP)



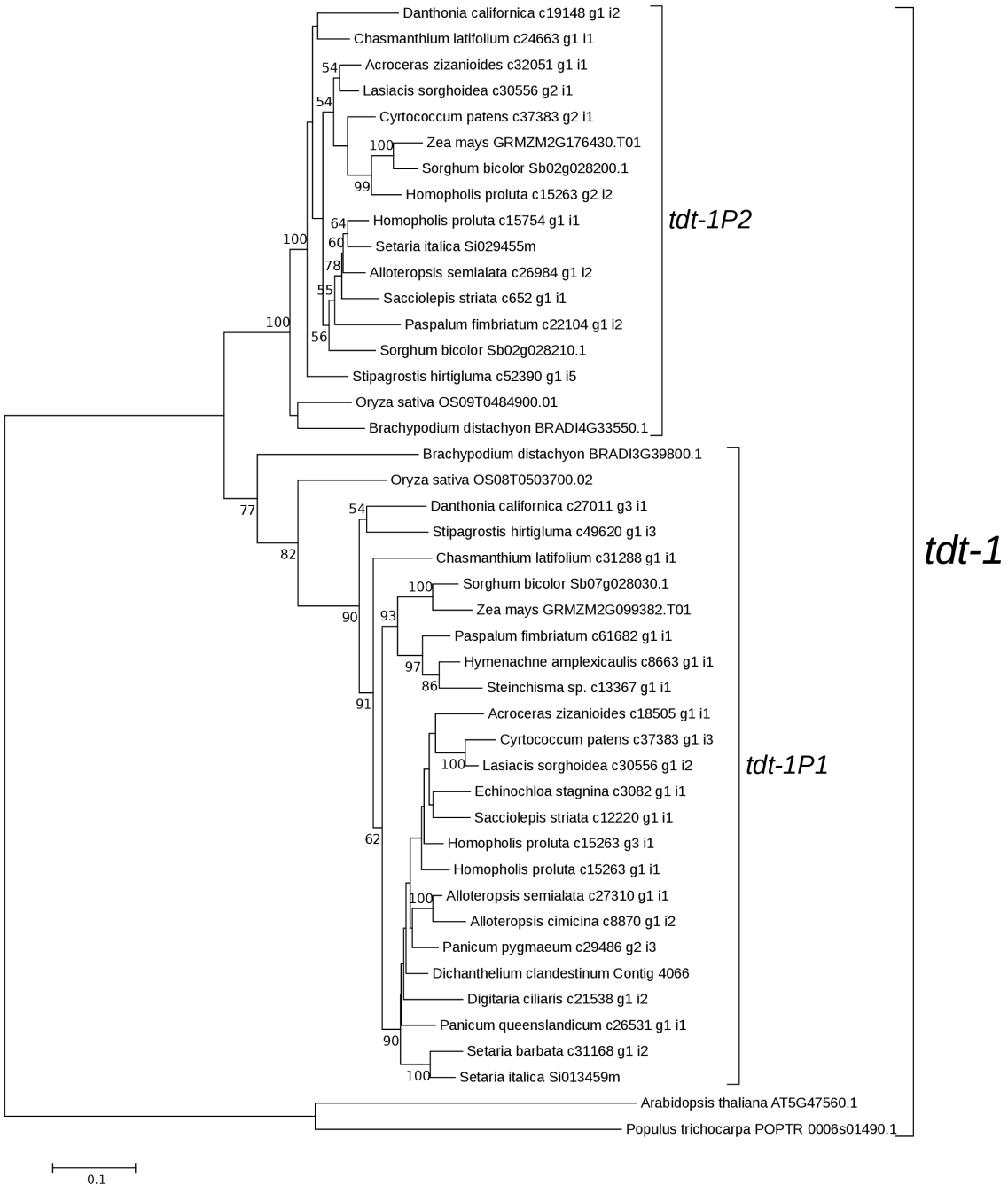
Phosphoenolpyruvate-phosphate translocator (PPT)/ Triosephosphate-phosphate translocator (TPT)/ Glucose-6-phosphate/phosphate translocator (GPT)



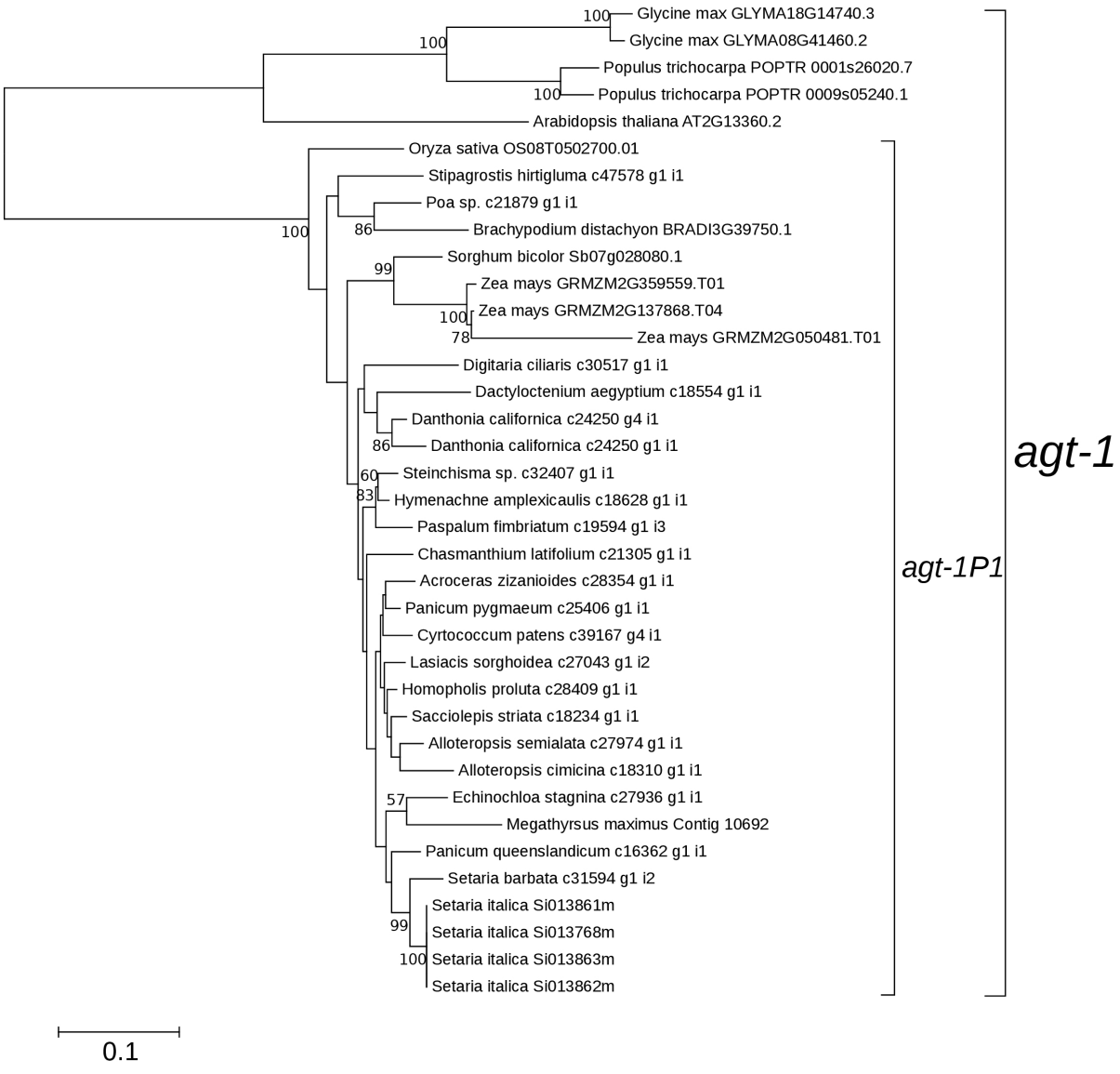
Phosphoenolpyruvate-phosphate translocator (PPT)/ Triosephosphate-phosphate translocator (TPT)/ Glucose-6-phosphate/phosphate translocator (GPT)



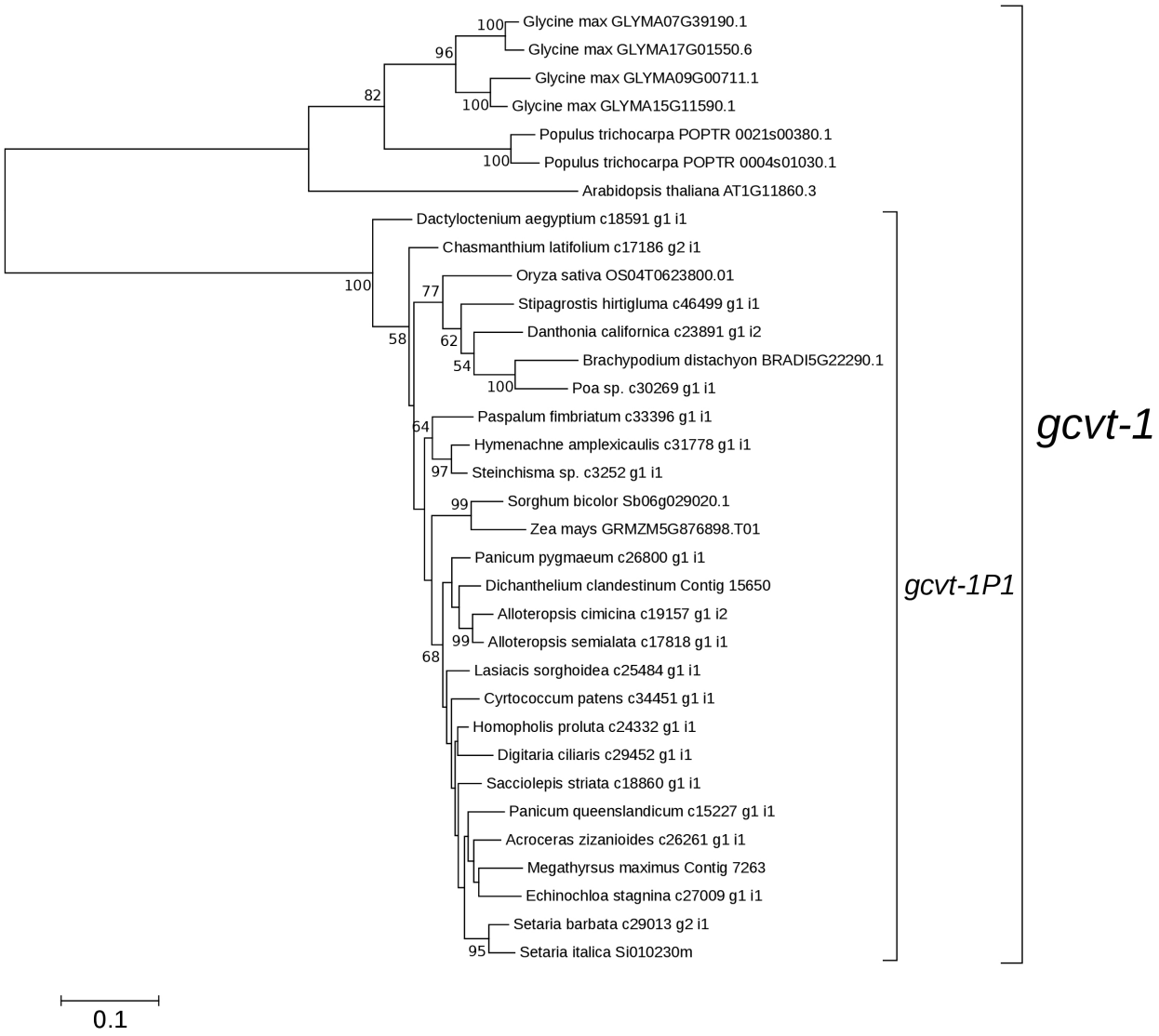
Tonoplast malate/fumarate transporter (TDT)



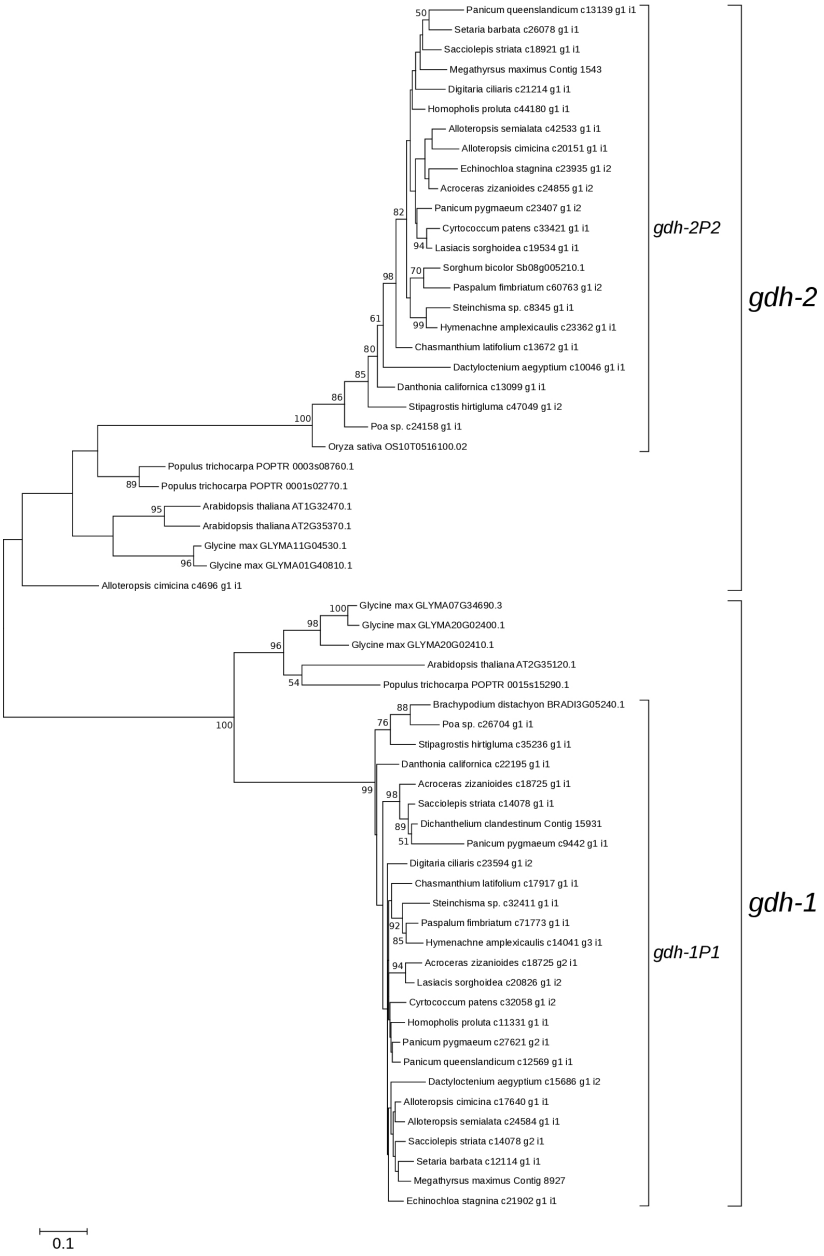
Serine--glyoxylate aminotransferase / Alanine--glyoxylate aminotransferase (AGT)



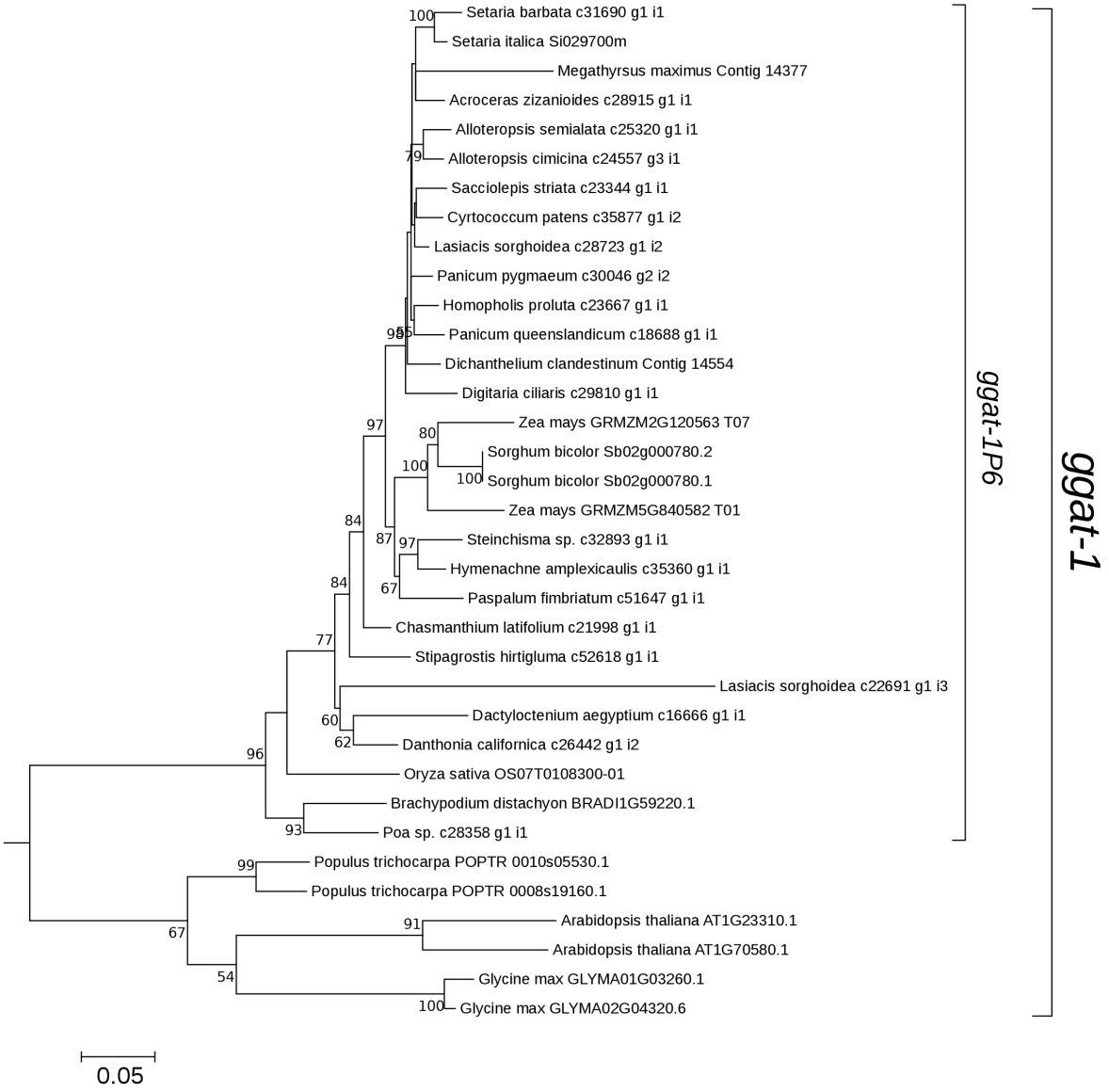
Glycine cleavage system T protein (GCVT)



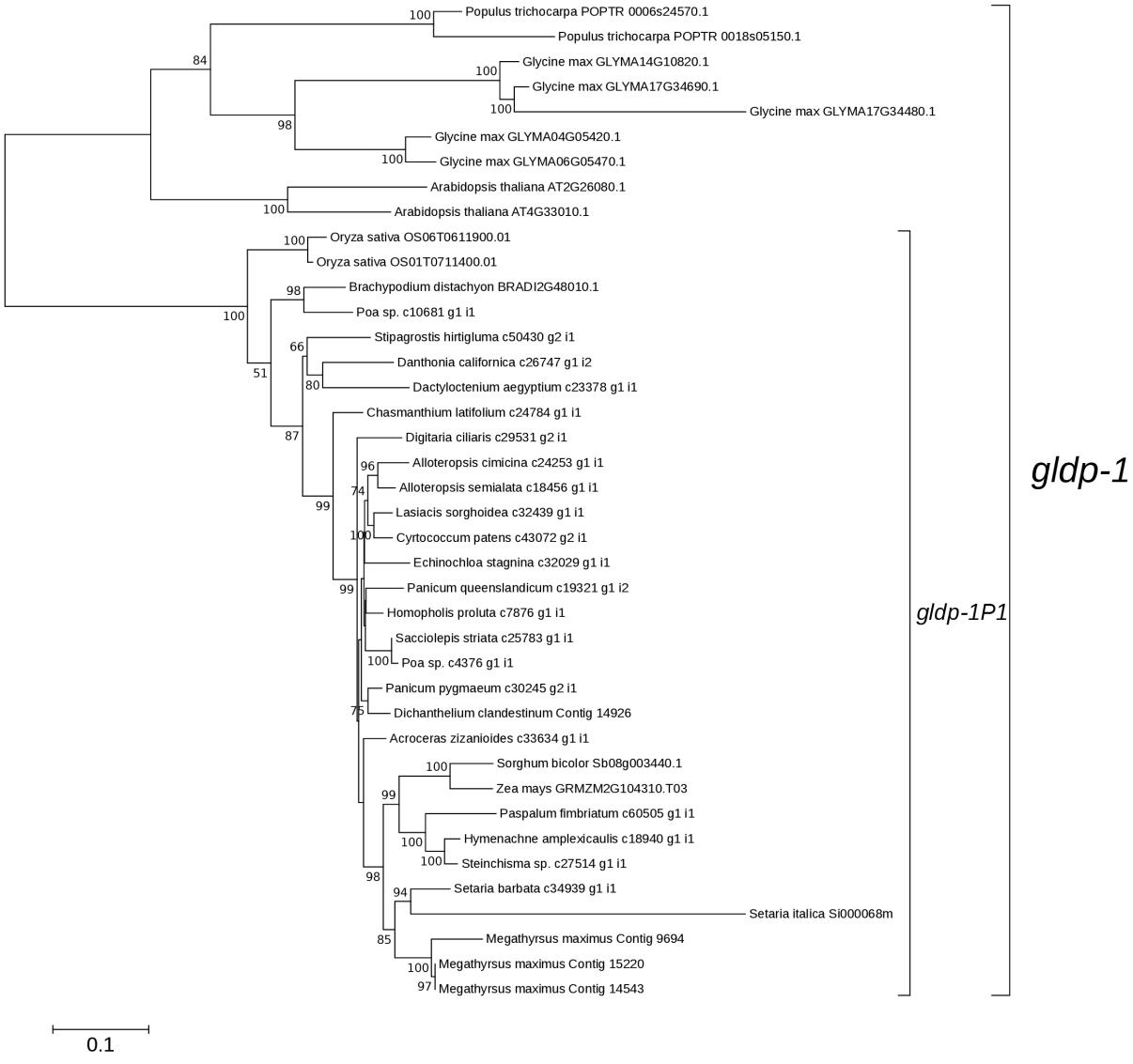
Glycine cleavage system H protein, mitochondrial (GDH)



Glutamate--glyoxylate aminotransferase (GGAT)

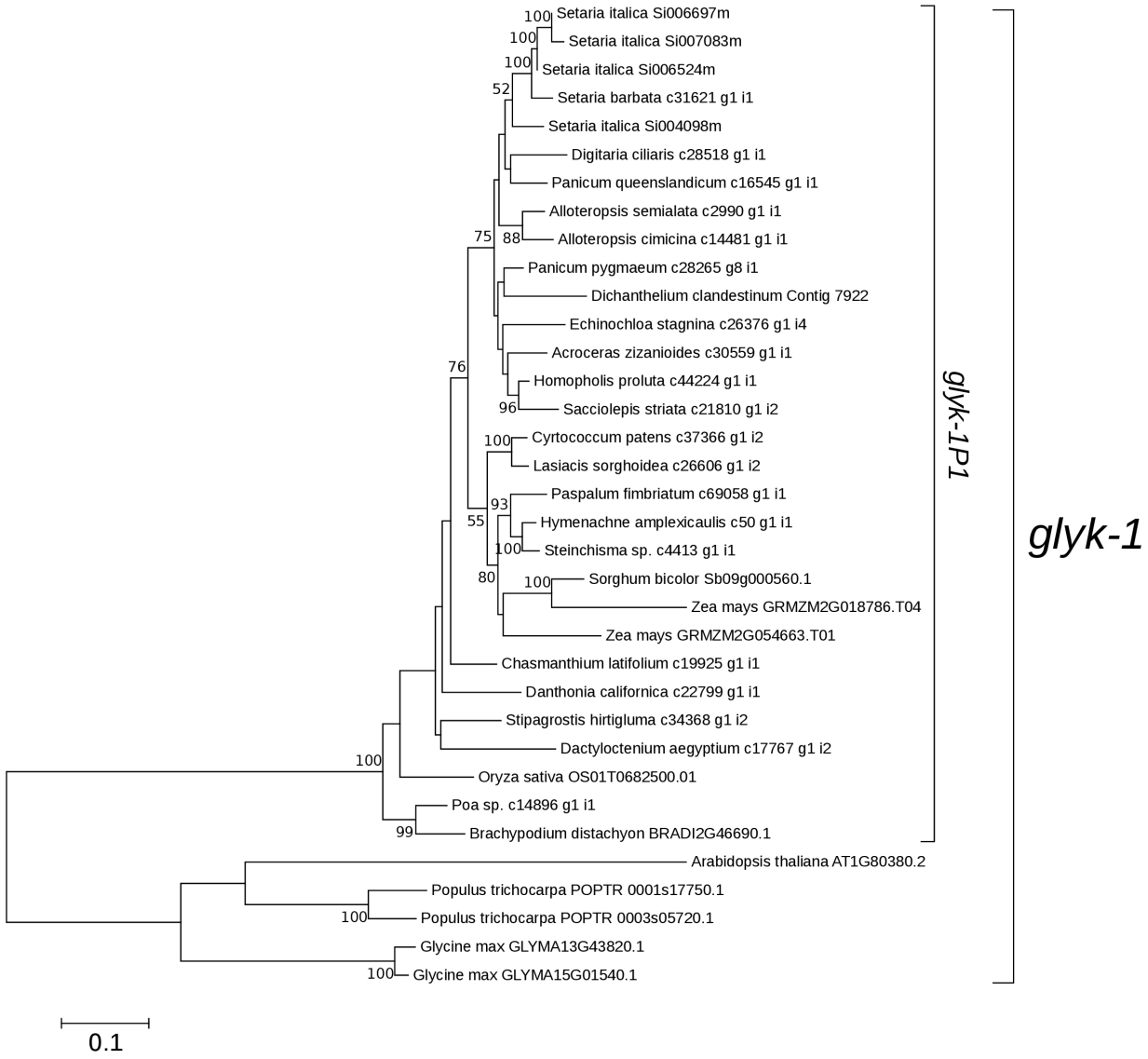


Glycine cleavage system P protein, mitochondrial (GLDP)

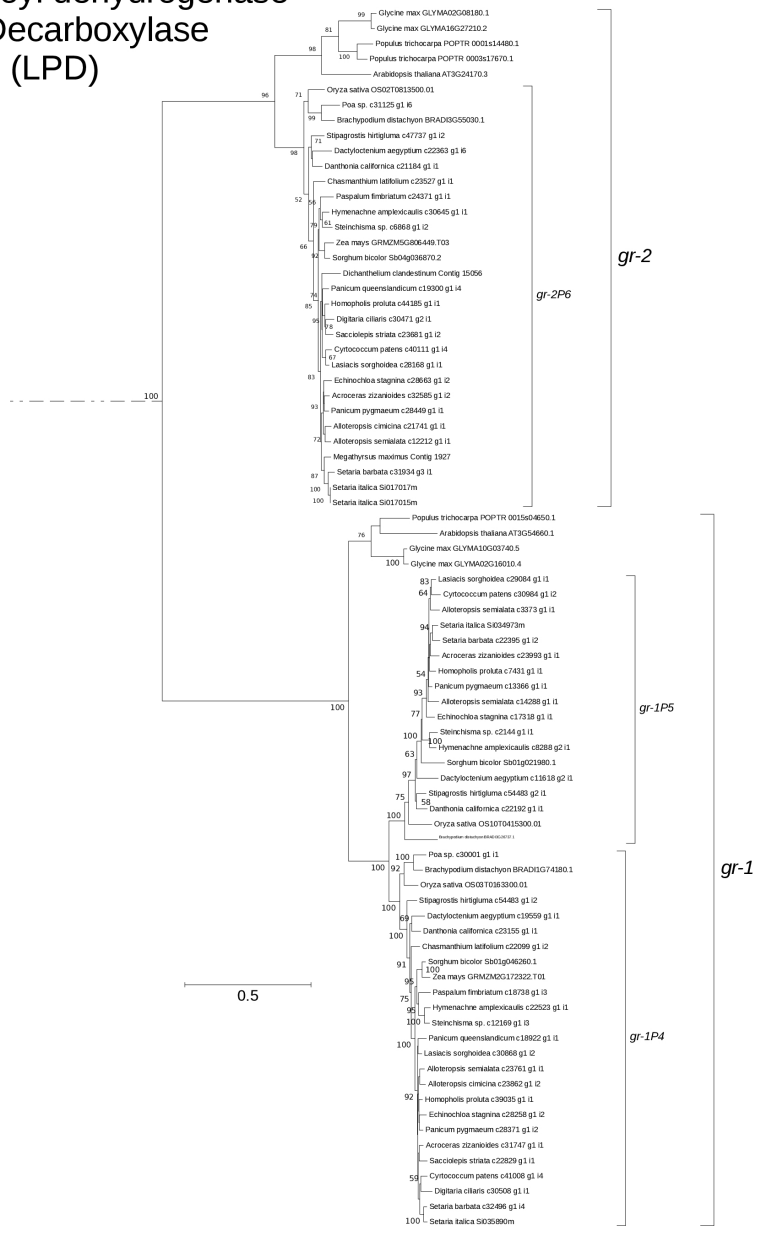


Peroxisomal (S)-2-hydroxy-acid oxidase (GLO)

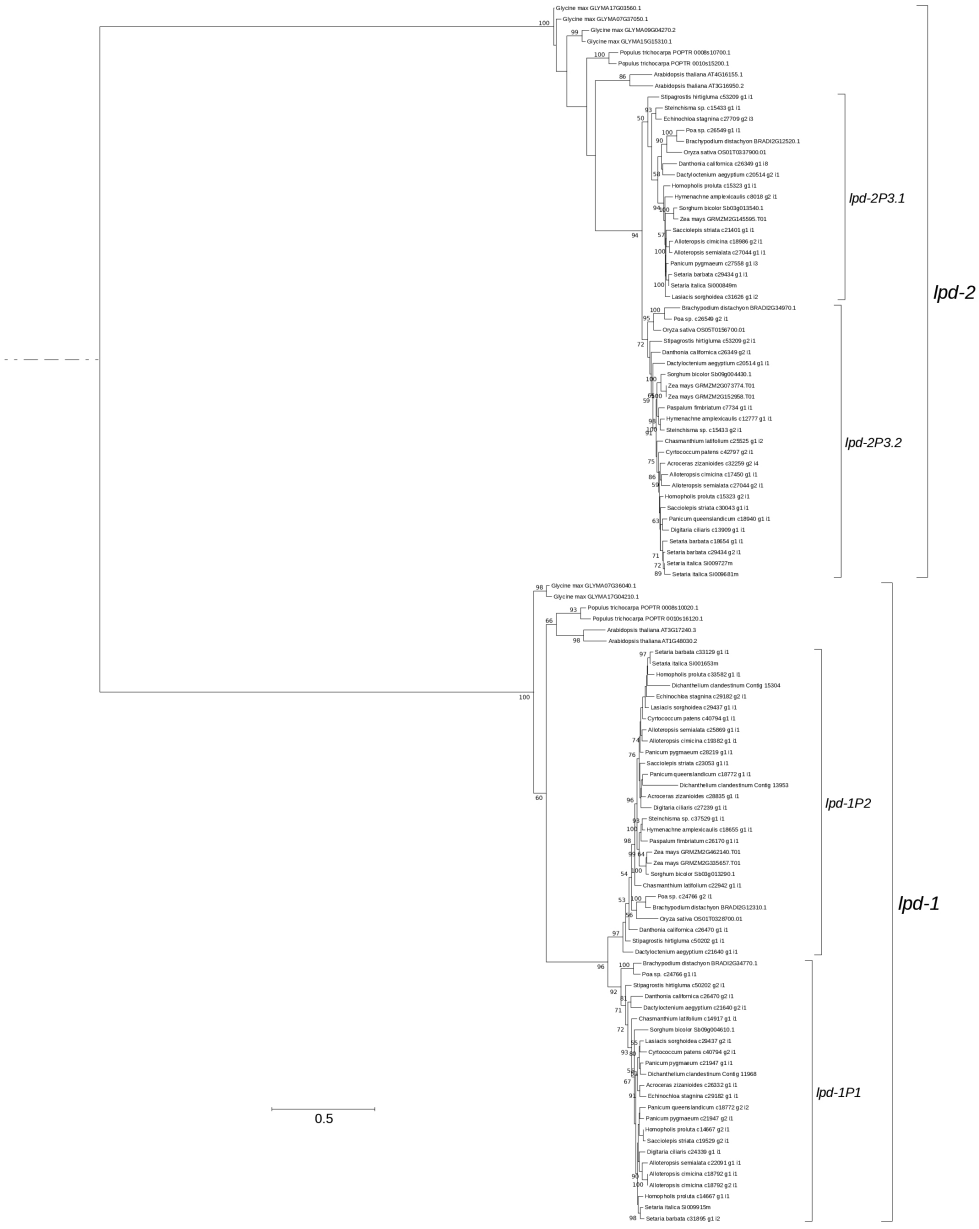
D-glycerate 3-kinase, chloroplastic (GLYK)



Glutathione reductase (GR) / Dihydrolipoyl dehydrogenase (Glycine Decarboxylase L protein) (LPD)

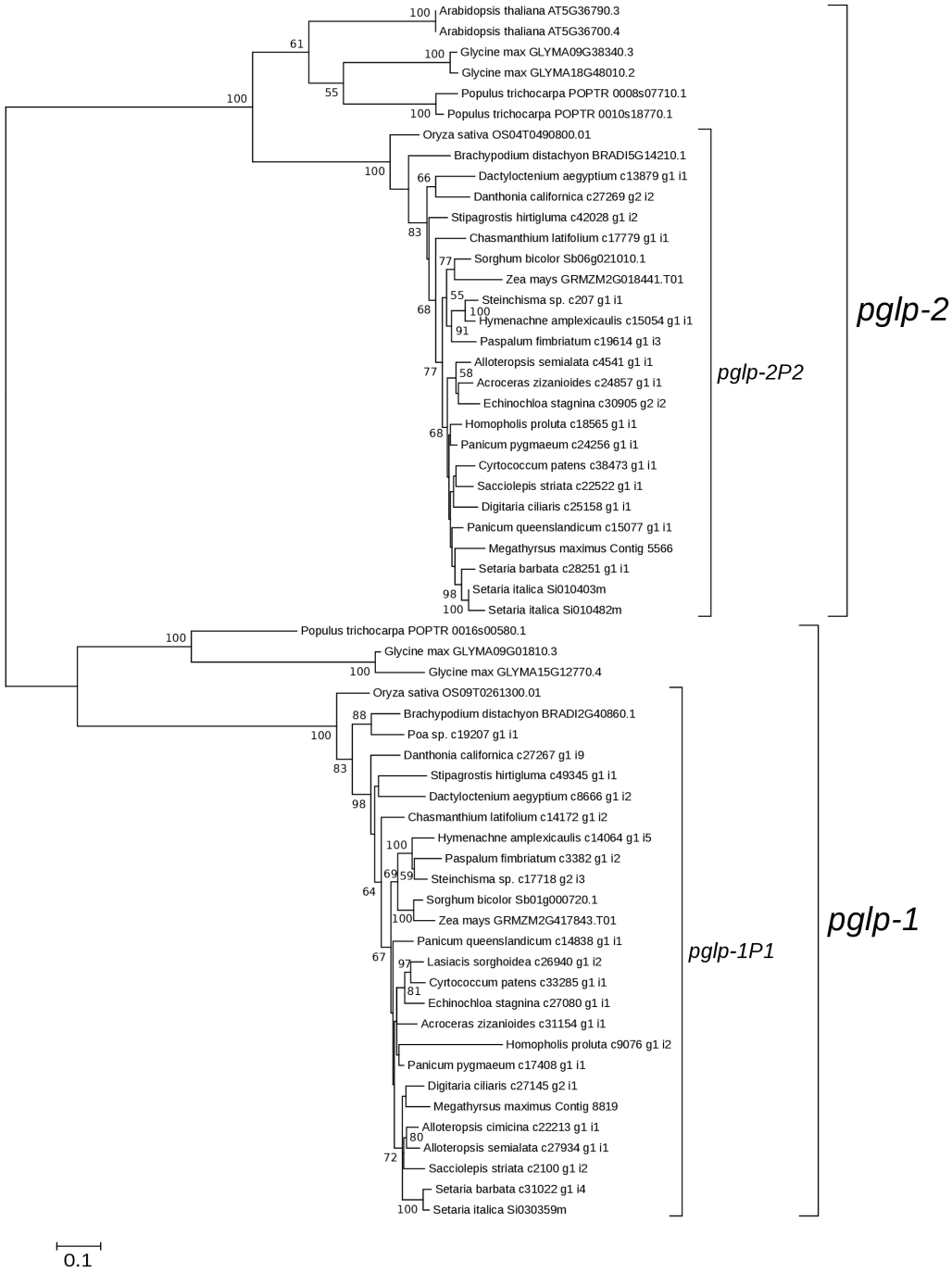


Glutathione reductase (GR) /
Dihydrolipoyl dehydrogenase
(Glycine Decarboxylase
L protein) (LPD)

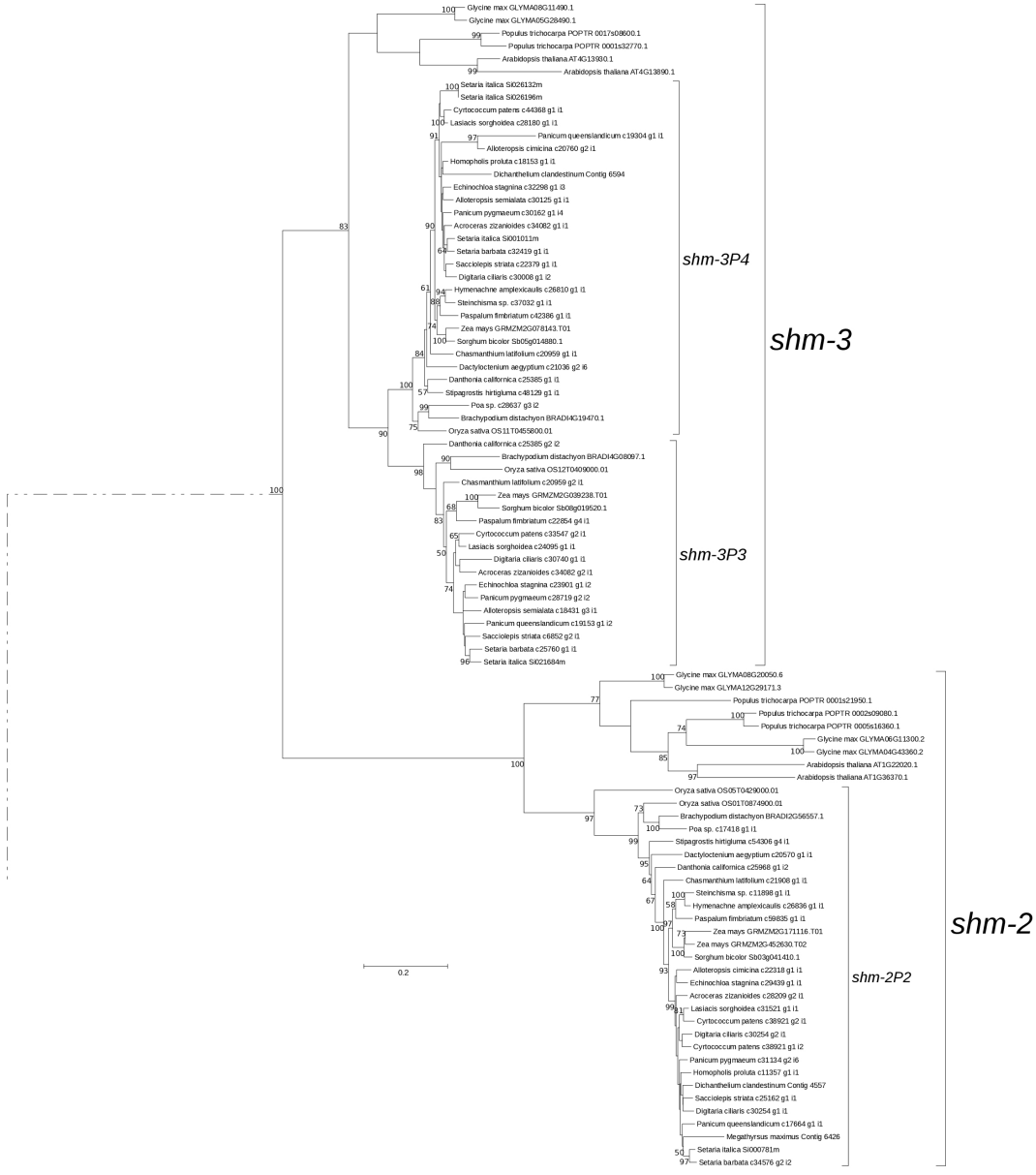


Glycerate dehydrogenase (glyoxylate/hydroxypyruvate reductase) (HPR)

Phosphoglycolate phosphatase, chloroplastic (PGLP)



Serine hydroxymethyltransferase (SHM)



Serine hydroxymethyltransferase (SHM)

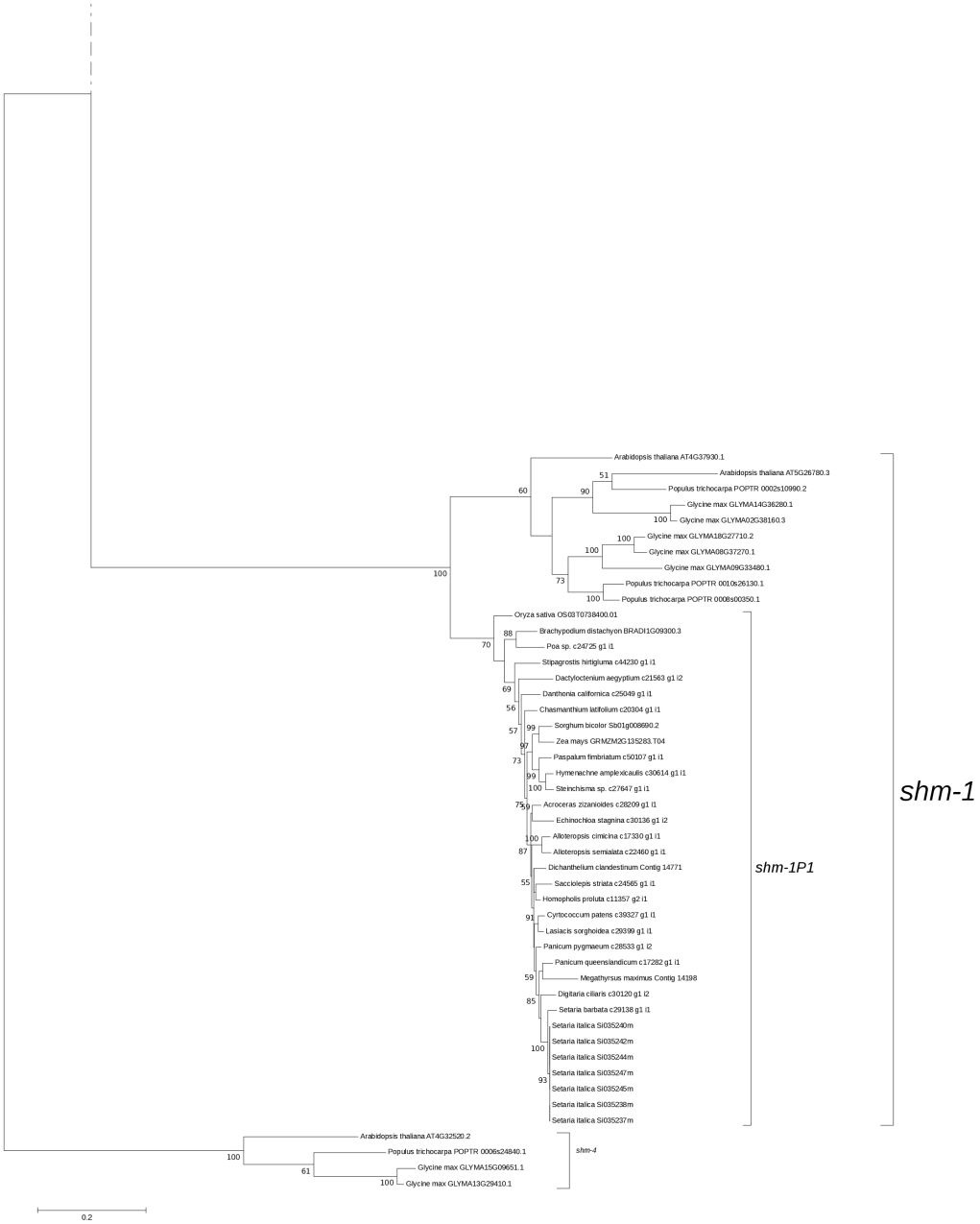


Table I.S1. Sequencing statistics.

Species	Abbre-viation	Tissue	PE Reads	Clean PE reads	Average reads	Contigs	Unigenes	N50 bp	Bowtie2 % CDS C4 genes
Acroceras zizanioides	AZIZ	Leaf	2202076	2035818	13048901	79597	53812	1514	2.66
	C_3	Leaf	3981203	3561989					2.1
		Root	5390799	5006883					1.98
		Root	2585123	2444211					1.19
Alloteropsis semialata	ASEM	Leaf	12422286	6546514	15218811	66427	47024	1591	7.4
	C_4	Leaf	1692525	1573457					4.26
		Root	3950750	3699333					1.53
		Root	3728547	3399507					1.11
Chasmanthium latifolium	CLAT	Leaf	10009259	8804513	23339126	64166	43050	1561	2.36
	C_3	Leaf	7933318	7136445					2.52
		Root	6440084	3148705					0.92
		Root	7522351	4249463					1.02
Cyrtococcum patens	CPAT	Leaf	8981967	6133295	33907966	120411	80962	1585	2.02
	C_3	Leaf	14553525	13258384					2.15
		Root	3367586	3162657					0.91
		Root	12455029	11353630					0.72
Dactyloctenium aegyptium	DAEG	Leaf	3562531	3255627	13266598	66023	44349	1661	2.58
	C_4	Leaf	8152613	7052648					3.46
		Root	3157583	2958323					0.88
Danthonia californica	DCAL	Leaf	3630245	3339890	15056375	76773	49480	1459	2.28
	C_3	Leaf	4639833	3945144					0.35
		Root	3342941	3082087					1.23
		Root	5574056	4689254					1.25
Digitaria ciliaris	DCIL	Leaf	4235388	3939771	15502945	55233	37500	1227	5.56
	C_4	Leaf	10424187	9195628					5.22
		Root	3949391	2015687					0.9
		Root	5814392	351859					0.31
Echinochloa stagnina	ESTA	Leaf	3191447	2903644	21321519	92233	66162	1458	6.99
	C_4	Leaf	11843314	10218776					6.91
		Root	1523231	1404008					1.12
		Root	7633573	6795091					0.83
Homopholis proluta	HPRO	Leaf	5729677	5134661	35973954	71089	52987	1888	3.11
	C_3	Leaf	11960661	10845052					2.18
		Root	10504783	9515317					0.91
		Root	11439945	10478924					0.89
Hymenachne amplexicaulis	HAMP	Leaf	3005381	2786048	12297065	53157	41303	1750	2.38
	C_3	Leaf	3007338	2269640					1.92
		Root	4241006	3931459					0.95
		Root	3569544	3309918					1.18
Lasiacis	LSOR	Leaf	2430756	2002336	27324623	103318	65626	1729	1.35

sorghoidea	C_3	Leaf	10391118	8957761					1.47
		Root	8533238	7936961					1.12
		Root	9610224	8427565					1.12
Panicum pygmaeum	PPYG	Leaf	4093890	3793221	17115317	72117	49086	1455	3.14
	C_3	Leaf	8925603	8087404					1.93
		Root	4106624	3482683					0.63
		Root	2026469	1752009					0.75
Panicum queenslandicum	PQUE	Leaf	3149789	2744427	11295659	54568	37682	1533	3.27
	C_4	Leaf	7280068	6630484					3.46
		Root	2590629	753302					1.27
		Root	3219178	1167446					0.11
Paspalum fimbriatum	PFIM	Leaf	3395354	3089206	21948286	96581	79239	1557	3.78
	C_4	Leaf	6499178	5782979					3.73
		Root	4306982	3919145					1.65
		Root	9895376	9156956					0.54
Poa sp.	PSSP	Leaf	4211710	3854186	23749106	83414	58393	1611	3.95
	C_3	Leaf	6384791	5587353					3.06
		Root	5060069	4639668					1.17
		Root	12291082	9667899					0.89
Sacciolepis striata	SSTR	Leaf	3148919	2867772	13833935	65538	43362	1639	2.41
	C_3	Leaf	7664481	6497248					2.59
		Root	4066769	3780174					0.98
		Root	868241	688741					0.31
Setaria barbata	SBAR	Leaf	3466394	3135585	16114099	81043	49155	1460	4.93
	C_4	Leaf	13286332	11936681					2.92
		Root	2508897	1041833					1.35
Steinchisma sp.	OSSP	Leaf	1536814	1417371	21852346	68171	50211	1906	3.29
	C_3	Leaf	8263727	7181006					2.26
		Root	5965474	5508124					0.87
		Root	8826368	7745845					0.86
Stipagrostis hirtigluma	SHIR	Leaf	14766660	14241012	72201207	141897	102685	1475	4.19
	C_4	Leaf	11712943	11296169					6.06
		Root	16975997	16292662					1.36
		Root	31560980	30371364					1.09

Table I.S2: Transcript abundance of C_4-related genes in the different samples
(Molecular Biology and Evolution link, Supplementary data)

or

https://www.dropbox.com/s/rcwjnfkluve07du/MorenoVillena_TableS2.xls?dl=0

Table I.S3: Gene expression levels and co-option in two groups of eudicots.

Gene	Family	*Tarenaya spinosa*[1,2]	*Gynandropsis gynandra*[1,3]	Cleomaceae coopted	*Flaveria pringlei*[1,2]	*Flaveria robusta*[1,2]	*Flaveria australasica*[1,3]	*Flaveria* coopted
ak-1	AK	4.2	72.1	0	32.0	92.1	350.1	0
ak-2	AK	124.3	1055.0	1	104.6	96.5	168.7	0
alaat-1	ALA-AT	135.5	3862.9	1	55.3	125.1	966.7	1
aspat-1	ASP-AT	8.4	2895.2	1	90.4	69.3	24.2	0
aspat-2	ASP-AT	57.6	42.9	0	67.7	88.0	537.1	1
aspat-3	ASP-AT	6.7	33.6	0	128.4	128.1	180.0	0
dic-1	DIC	505.3	286.3	0	341.1	218.8	70.5	0
dic-2	DIC	0.0	459.7	0	25.5	23.1	8.0	0
dit-1	DIT	189.1	187.1	0	123.1	166.3	260.1	0
dit-2	DIT	60.7	5.2	0	50.7	58.1	96.7	0
nadmdh-1	NAD-MDH	82.9	39.2	0	62.7	42.3	39.7	0
nadmdh-2	NAD-MDH	170.1	217.8	0	173.5	180.2	136.1	0
nadmdh-3	NAD-MDH	1069.0	425.4	0	894.7	953.2	275.4	0
nadme-1	NAD-ME	39.3	813.0	1	53.3	71.0	68.7	0
nadme-2	NAD-ME	22.6	602.3	1	79.7	67.3	56.8	0
nadpmdh-1	NADP-MDH	169.4	262.8	0	229.9	372.4	2397.9	1
nadpmdh-2	NADP-MDH	2.9	4.1	0	16.4	43.6	2.5	0
nadpmdh-3	NADP-MDH	1137.5	446.1	0	437.2	578.4	643.0	0
nadpme-1E1	NADP-ME	24.2	91.4	0	133.0	368.3	2690.9	1
nadpme-1E2	NADP-ME	109.7	50.4	0	1.5	19.5	0.1	0
nadpme-1E3	NADP-ME	0.0	0.0	0	196.2	158.9	513.8	0
nhd-1	NHD	39.1	621.6	1	33.1	86.6	440.4	0
pck-1	PCK	8.9	60.6	0	9.3	25.3	19.0	0
pepck-1	PEPC-K	26.9	175.1	0	6.7	55.5	183.4	0
ppa-1	Ppa	717.5	2294.6	1	407.1	343.6	591.3	1
ppa-2	Ppa	82.7	229.1	0	46.8	52.3	176.2	0
ppa-3	Ppa	64.0	74.8	0	36.9	36.5	99.0	0
ppa-4	Ppa	310.4	137.4	0	7.5	10.9	13.9	0
ppc-1E1	PEPC	36.3	2841.0	1	0.0	54.0	1.6	0
ppc-1E2	PEPC	74.0	28.1	0	157.4	755.2	8302.4	1
ppc-2	PEPC	2.9	0.0	0	9.6	9.2	2.4	0
ppdk-1	PPDK	4.4	1022.9	1	160.8	471.8	5741.5	1
ppdkrp-1	PPDK-RP	44.9	111.7	0	20.7	83.0	321.8	0
ppt-1E1	TPT	16.3	0.0	0	176.8	243.1	186.9	0
ppt-1E2	TPT	58.9	1187.0	1	76.7	119.7	600.5	1
sbas-1	SBAS	39.0	3388.2	1	171.3	389.4	3420.0	1
sbas-2	SBAS	35.1	10.7	0	23.8	26.4	21.7	0
sbas-3	SBAS	10.5	3.0	0	17.6	39.2	4.5	0
sbas-4	SBAS	0.0	0.0	0	13.2	6.9	10.5	0
tpt-1E1	TPT	NA	NA	NA	0.2	68.3	1.4	0
tpt-1E2	TPT	654.1	2846.9	0	1103.6	755.1	895.1	1

[1] Expression levels are indicated for each gene, in rpkm values; [2] C_3 species; [3] C_4 species;

Table I.S4. Statistical models of co-option events for two groups of eudicots.

C_4 lineage	**Cleomaceae**		***Flaveria***	
Factors	**la[2]**	**family[3]**	**la[2]**	**family[3]**
p-value	0.19	0.39	0.01	0.84
df[1]	1,13	9,13	1,16	8,16
F-stat	1.87	1.17	8.17	0.49

[1] df = degrees of freedom. For each variable, the degrees of freedom for the residuals are given after the comma; [2] la = leaf abundance in close C_3 relatives; [3] gene family identity.

Table I.S5. Results of codon models comparisons.

gene	M1a	A	A1	p (M1a vs A1)[1]	p (A vs A1)[1]
ak-1P1	-4987.15	-4986.06	-4986.06	0.336	1.000
alaat-1P5	-9087.17	-9086.62	-9085.65	0.217	0.162
aspat-2P3	-7876.86	-7876.86	-7876.86	1.000	1.000
aspat-3P4	-7014.93	-7007.31	-7007.31	0.000*	1.000
bca-2P3	-6077.19	-6065.60	-6065.32	0.000*	0.457
dit-2P3	-6734.09	-6733.82	-6733.82	0.766	1.000
nadpmdh-1P1	-6195.23	-6194.90	-6194.90	0.717	0.923
nadpmdh-3P4	-5526.46	-5526.46	-5526.46	1.000	1.000
nadpme-1P4	-11513.59	-11461.08	-11460.45	0.000*	0.261
nhd-1P1	-8313.08	-8312.22	-8308.42	0.010	0.006
pck-1P1	-10293.72	-10184.91	-10184.91	0.000*	1.000
pepck-1P1	-2080.55	-2079.40	-2078.91	0.193	0.321
ppa-1P2.1	-4347.99	-4347.97	-4347.97	0.978	1.000
ppc-1P3	-18420.90	-18319.22	-18312.37	0.000*	0.000*
ppc-1P6	-11066.51	-11034.80	-11034.80	0.000*	1.000
ppdk-1P2	-18091.43	-18045.71	-18032.27	0.000*	0.000*
ppt-1P5	-5425.61	-5425.46	-5425.40	0.809	0.727
sbas-1P1	-5718.98	-5697.67	-5677.22	0.000*	0.000*
tpt-1P1	-5468.10	-5445.68	-5433.08	0.000*	0.000*

[1] p-values based on likelihood ratio tests. Significant comparisons after correction for multiple testing are indicated with an asterisk.

Chapter II: Key changes in gene expression identified for different stages of C_4 evolution

Luke T. Dunning[*,1], Jose J. Moreno-Villena[*,1], Marjorie R. Lundgren[1], Jacqueline Dionora[2], Paolo Salazar[2], Claire Adams[3], Florence Nyirenda[4], Jill K. Olofsson[1], Anthony Mapaura[5], Isla M. Grundy[6], Canisius J. Kayombo[7], Lucy A. Dunning[8], Fabrice Kentatchime[9], John Thompson[10], Guillaume Besnard[11], W. Paul Quick[1,2], Andrea Bräutigam[12], Colin P. Osborne[1], Pascal-Antoine Christin[1,a,b]

[*] These authors contributed equally to this work

[1] Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom, [2] International Rice Research Institute, DAPO, Metro Manila, Philippines

[3] Botany Department, Rhodes University, 6140 Grahamstown, South Africa, [4] Department of Biological Sciences, University of Zambia, Lusaka, Zambia, [5] National Herbarium and Botanic Garden, Harare, Zimbabwe, [6] Department of Biological Sciences, University of Zimbabwe, Harare, Zimbabwe, [7] Forestry Training Institute, Olmotonyi, Tanzania, [8] Department of Social Sciences, University of Sheffield, 219 Portobello, Sheffield S1 4DP, United Kingdom, [9] CABAlliance, P.O. Box 3055 Messa, Yaoundé, Cameroon, [10] Queensland Herbarium, Department of Science, Information Technology and Innovation (DSITI), Mt Cooth-tha Botanic Gardens, Toowong, QLD 4066, Australia, [11] Laboratoire Évolution et Diversité Biologique (EDB UMR5174), Université de Toulouse, CNRS, ENSFEA, UPS, Toulouse, France, [12] Bielefeld University, Universitätsstrasse 35, 33501 Bielefeld, Germany

[a] Corresponding author: Pascal-Antoine Christin; telephone +44-(0)114-222-0027; fax +44 114 222 0002; email: p.christin@sheffield.ac.uk

Personal contribution: I generated the data, which I analysed jointly with Dr Luke Dunning. I helped design the study and co-wrote the paper with Dr Dunning, with the help of my co-authors.

Abstract

C_4 photosynthesis is a complex trait that boosts productivity in tropical conditions. When compared to C_3 species, the C_4 state seems to require numerous novelties, but species comparisons can be confounded by long divergence times. Here, we exploit the remarkable photosynthetic diversity that exists within a single species, the grass *Alloteropsis semialata*, to detect changes in gene expression associated with different phases of C_4 evolution. Comparative transcriptomics in a phylogenetic context show that the intermediates with a weak C_4 cycle are separated from the C_3 phenotype by increases in the expression of only 12 genes, including those encoding three core C_4 enzymes: ASP-AT, PCK, and PEPC. The subsequent transition to full C_4 physiology involved further increases in just five genes, including those for the C_4 enzymes ALA-AT and PPDK. These changes likely created a rudimentary C_4 trait, and isolated C_4 populations later adapted the emerging C_4 physiology, resulting in a patchwork of differential expression for other C_4-accessory genes. Our work shows how the assembly of the C_4 pathway happened in incremental steps, each requiring few alterations over the previous one. These create short bridges across the adaptive landscapes that likely facilitated the recurrent origins of C_4 photosynthesis.

Keywords: adaptation, C_4 photosynthesis, complex trait, intermediates, phylogenetics, transcriptomics

Introduction

The origins of traits composed of multiple anatomical and/or biochemical components have always intrigued evolutionary biologists (Darwin, 1859; Meléndez-Hevia *et al.*, 1996; Lenski *et al.*, 2003). If such traits gain their function only through the co-ordinated action of multiple components, their evolution via natural selection would have to cross a valley in the adaptive landscape. Despite this obstacle, complex traits have evolved repeatedly, in diverse groups of organisms. This apparent paradox is solved for most traits by the existence of intermediate stages, which act as evolutionary enablers, creating bridges over the valleys of the adaptive landscape (Jacob, 1977; Dawkins, 1986; Weinreich *et al.*, 2006; Blount *et al.*, 2012; Vopalensky *et al.*, 2012; Werner *et al.*, 2014). The accessibility of new traits likely depends on the length and complexity of such bridges, which are generally unknown. Quantifying the evolutionary gap between phenotypic states is therefore crucial to contextualise the likelihood of a novel trait evolving.

An excellent system to study the evolutionary trajectories of an adaptive trait is C_4 photosynthesis. This metabolic pathway increases CO_2 concentration at the active site of assimilation via the Calvin-Benson cycle (Hatch, 1987; Sage, 2004; Christin & Osborne, 2014). This avoids the energetically costly process of photorespiration, effectively increasing photosynthetic efficiency in warm and arid conditions (Sage *et al.*, 2012). This CO_2-concentrating mechanism relies on a set of specific leaf anatomical properties and the co-ordinated action of multiple enzymes and numerous associated proteins (Hatch, 1987; Bräutigam *et al.*, 2011; Külahoglu *et al.*, 2014; Lundgren *et al.*, 2014). Despite its apparent complexity, C_4 photosynthesis is a textbook example of convergent evolution, having independently evolved more than 60 times within flowering plants (Sage *et al.*, 2011). The origins of C_4 photosynthesis were facilitated by the presence of anatomical enablers in some groups (Christin *et al.*, 2013, Sage *et al.*, 2013), but the processes leading to a functioning C_4 biochemical pathway on top of these are less well understood. All C_4 enzymes exist in C_3 plants, but are involved in different, non-photosynthetic pathways (Aubry *et al.*, 2011). The genes ancestrally abundant in the leaves were preferentially co-opted for C_4 (Moreno-Villena et al. 2018), and changes to their expression patterns and/or kinetic properties of the encoded enzyme followed (Bläsing *et al.*, 2000; Hibberd & Covshoff, 2010; Huang et al. 2016,

Moreno-Villena et al. 2018), with cell-specific expression realized in some cases through the recruitment of pre-existing regulatory mechanisms (Brown *et al.*, 2011; Kajala *et al.*, 2012; Cao *et al.*, 2016, Reyna-Llorens & Hibberd, 2017).

The evolutionary transition between C_3 and C_4 phenotypes involves intermediate stages which only have some of the anatomical and biochemical modifications typical of C_4 plants. In particular, some C_3+C_4 plants perform a weak C_4 cycle that is responsible for only part of their carbon assimilation (these correspond to 'type II C_3-C_4 intermediates'; Ku *et al.*, 1983; Monson *et al.*, 1986; Schlüter and Weber, 2016). This weak C_4 cycle might have emerged through the upregulation of C_4-related enzymes to balance nitrogen among cellular compartments in the multiple lineages of plants that use a photorespiratory pump (Sage *et al.*, 2011, 2012; Mallmann *et al.*, 2014; Bräutigam & Gowik, 2016). Metabolic models suggest that any increase in flux of CO_2 fixed through the C_4 cycle in intermediate plants directly translates into biomass gain, leading to gradual increases in C_4 gene expression (Heckmann *et al.*, 2013; Mallmann *et al.*, 2014). The current model of C_4 evolution therefore assumes gradual, yet abundant changes in plant transcriptomes and genomes during the transition from C_3 ancestors to physiologically C_4 descendants. Indeed, comparisons of C_3 and C_4 species have typically identified tens to thousands of differentially expressed genes encoding C_4 enzymes, regulators, and accessory metabolite transporters (Bräutigam *et al.*, 2011, 2014; Gowik *et al.*, 2011; Külahoglu *et al.*, 2014; Li *et al.*, 2015). These large numbers might however partially results from the comparison of species typically separated by millions of years of divergence (Christin *et al.*, 2011), which leaves ample time for the accumulation of secondary changes linked to the C_4 trait beyond the minimal requirements, as well as variation in other unrelated traits. Furthermore, previous efforts have typically targeted very few individuals per C_4 lineage, such that the initial bout of co-option that generated a C_4 cycle cannot be distinguished from subsequent adaptation via natural selection (Christin & Osborne, 2014).

In this study, comparative transcriptomics within a phylogenetic context are used to quantify the phenotypic distance between the C_3 phenotype and weak C_4 cycle (C_3+C_4 state) independently from those responsible for the transition to the full C_4 type, and finally from those involved in the adaptation of the existing C_4 phenotype. The time elapsed between transitions, and therefore the number of changes unrelated to C_4

emergence, is reduced by focusing on a single species containing a gradient of photosynthetic types, the grass *Alloteropsis semialata*. Congeners of *A. semialata* are C_4, but previous comparative transcriptomics and leaf anatomy have shown that the C_4 biochemistry emerged multiple times in the genus, from a common ancestor with some C_4-like characters (Fig. II.1; Dunning et al. 2017). Capitalizing on the physiological diversity existing within *A. semialata*, leaf transcriptomes from multiple individuals originating from diverse populations of each photosynthetic type in this species are analysed, together with closely related C_3 and C_4 species, to detect the changes in gene expression responsible for (i) the phenotypic difference between C_3 plants and C_3+C_4 intermediates, (ii) the shift to fixing atmospheric CO_2 exclusively via the C_4 pathway in solely C_4 plants, and (iii) the adaptation of the C_4 cycle after its evolution in geographically isolated C_4 populations. This deconstruction of the genetic origins of a complex biochemical pathway sheds new light on the amount of genetic change needed to move to another part of the adaptive landscape during different stages of a stepwise physiological transition.

Material and Methods

Species sampling and growth conditions

Our sampling was designed to capture the diversity of photosynthetic types in the group and the genetic diversity within each photosynthetic type. We previously published transcriptomes for 13 individuals representing different populations of *A. semialata*, congeners and a C_3 outgroup (Table S1; Dunning et al. 2017). This dataset was complemented here with two extra populations generated with the same methodology to even the representation of photosynthetic types (Table II.S1; Fig. II.S1). These 15 unreplicated populations are used to evaluate the variation across the group, but a different set of three individuals for each of ten populations, only one of which was analysed previously, was generated specifically for analyses of differential expression (Table II.S2; Fig. II.S1). Within *Alloteropsis semialata*, these populations include two C_3 ones (RSA6 and ZIM2) that represent extremes of the C_3 geographic range (Fig. II.S1), two geographically distant C_3+C_4 populations (TAN5 and ZAM3; Fig. II.S1) that operate a weak C_4 cycle based on PEPC protein abundance (Lundgren et al. 2016), and two C_4 populations (PHI1 and TAN4) collected from different continents (Fig. II.S1),

which belong to different intraspecific genetic subgroups (Olofsson et al. 2016; Table S2). A third C_4 population (CMR1) was also sampled as preliminary investigations suggested it was distinct from all the others previously screened. The C_4 populations of *A. semialata* have decreased CO_2-compensation points, increased carboxylation efficiencies, and shifts in carbon isotopes that confirm their photosynthetic type (Lundgren et al. 2016). The C_4 leaves are characterized by increased vein density and PEPC protein abundance (Lundgren et al., 2016; Dunning et al., 2017). The C_3+C_4 *A. semialata* also show elevated levels of PEPC protein and increased concentration of chloroplasts in bundle sheaths, but no increased vein density. This results in a reduced CO_2-compensation point, but a non-C_4 carbon isotope signature (Lundgren et al. 2016; Dunning et al., 2017).

In addition to these seven *A. semialata* populations, one population of each of the C_4 congeners *A. angusta* and *A. cimicina* were added, to enable comparison of the degree of C_4-related changes in gene expression (Table II.S2). Finally, *Entolasia marginata* was included as a C_3 outgroup. Three distinct genotypes were sequenced per population except in two cases, where this was not possible. For *A. angusta*, three clones that were established more than one year before the study were used, while for *E. marginata* a genotype was sampled once together with two clones from a different genotype, again established long before the study.

Plants were collected from the field as seeds or live cuttings, and subsequently grown under controlled conditions at the University of Sheffield as previously described (Dunning et al., 2017). In brief, plants were potted in John Innes No. 2 compost (John Innes Manufacturers Association, Reading, England) and maintained under non-limiting soil moisture and nutrient conditions in controlled environment chambers (Conviron BDR16; Manitoba, Canada) set to 60% relative humidity, 500 μmol m^{-2} s^{-1} light intensity, 14h photoperiod, and day/night temperatures of 25/20°C. After a minimum of 30 days in these growth conditions, young fully expanded leaves were sampled for transcriptome analyses.

RNA extraction, sequencing, and transcriptome assembly

RNA extraction, library preparation and sequencing were performed as previously described (Dunning et al 2017). In brief, total RNA was extracted from the distal half of

fully expanded fresh leaves, sampled in the middle of the light period, using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) with an on-column DNA digestion step (RNase-Free Dnase Set; Qiagen, Hilden, Germany). Total RNA was used to generate 32 indexed RNA-seq libraries using the TruSeq RNA Library Preparation Kit v2 (Illumina, San Diego, CA). Each library was subsequently sequenced on 1/24 of a single Illumina HiSeq 2500 flow-cell (with other samples from the same or unrelated projects), which ran for 108 cycles in rapid mode at the Sheffield Diagnostic Genetics Service.

The raw RNA-Seq data were cleaned to remove low quality reads (Q<30), and sequences corresponding to ribosomal RNA or containing adaptor contamination using the Agalma pipeline v.0.5.0 (Dunn et al., 2013). *De novo* transcriptomes were subsequently assembled for each of the 12 newly sequenced populations using Trinity (version trinityrnaseq_r20140413p1; Grabherr et al., 2011). All raw data and transcriptome assemblies have been submitted to the NCBI repository (Bioproject PRJNA401220). Coding sequences (CDS) longer than 500 bp were predicted for each population using OrfPredictor (Min et al., 2005), which uses homology to a user supplied reference protein database or *ab initio* predictions if no suitable match is found. The protein database used comprised the complete coding sequences of eight model species: *Arabidopsis thaliana*, *Brachypodium distachyon*, *Glycine max*, *Oryza sativa*, *Populus trichocarpa*, *Setaria italica, Sorghum bicolor* and *Zea mays.*

Phylogenetic reconstruction using core-orthologs

Single-copy orthologs were extracted from the new and previously published transcriptome assemblies to infer phylogenetic relationships among individuals. Homologous sequences to 581 single-copy plant core-orthologs previously determined in the Inparanoid ortholog database (Sonnhammer and Ostlund, 2014) were identified. A Hidden Markov Model based search tool (HaMSTR v.13.2.3; Ebersberger et al., 2009) was used to screen the CDS of the transcriptomes. Sequences of the single copy plant core-orthologs were subsequently aligned using a previously described stringent alignment and filtering pipeline (Dunning et al., 2017). In brief, the CDS were translated into proteins and a consensus alignment was made by only retaining residues aligned in the same position by four different programs (mafft, muscle, kalign, t-coffee). The protein alignments were further parsed using the tcs residue filter (Chang et al.,

2014), only keeping the highest confidence positions. Finally, the original nucleotide sequences were threaded onto the trimmed protein alignments, before being trimmed themselves with gblocks v.0.91 (parameters: -t=c -b2=b1 -b5=h; Castresana, 2000). Sequences shorter than 100 bp after trimming were discarded. A maximum likelihood phylogeny was subsequently inferred from a concatenated alignment of the single copy plant core-orthologs that had sequence data from at least ten different populations using PhyML (Guindon and Gascuel, 2003) with a GTR+G+I model and 100 bootstrap pseudoreplicates.

Identification of gene families and co-orthologs

Quantitative comparison of transcript abundance requires classifying expressed sequences into equivalent units. To do this, transcripts were grouped into co-orthologs of Panicoideae, the subfamily of grasses that contains all the species studied here, using a previously described method (Dunning et al., 2017). Each group of co-orthologs contains all the genes descended from a single gene in the genome of the last common ancestor of Panicoideae via a combination of speciation and/or gene duplication events. CDS for eight published genomes were grouped into previously established gene families (i.e. homolog groups containing all paralogs and orthologs; Vilella et al. 2009). The eight selected reference genomes included two Panicoideae grasses (*S. italica* and *S. bicolor*), two non-Panicoideae grasses (*B. distachyon* and *O. sativa*), and four non-grass species (*Amborella trichopoda, A. thaliana*, *P. trichocarpa*, and *Selaginella moellendorffii*). To enable accurate annotation, the analysis was restricted to gene families with at least one sequence of *A. thaliana* and one of *S. bicolor* or *S. italica*. Sequences from the transcriptomes homologous to each gene family were identified using nucleotide BLAST sequence searches with a minimum matching length of 500 bp. Each gene family was then stringently aligned as described above and a phylogenetic tree was inferred using PhyML. Groups of Panicoideae co-orthologs were identified as monophyletic clades of Panicoideae sequences that contained at least one *S. italica* or *S. bicolor* sequence, and one transcriptome CDS. These groups potentially include duplicates that emerged after the diversification of Panicoideae, which might be difficult to distinguish due to low sequence divergence.

Differential expression analyses

The trimmed nucleotide CDS from the above alignments were used as the reference for read mapping, where each group of Panicoideae co-orthologs is represented by multiple sequences from different species and/or populations. Cleaned reads were mapped to the reference using the local alignment option in Bowtie2 (Langmead & Salzberg, 2012). For each read, the single best alignment was recorded, which was randomly selected if there were equally good matches (a default parameter). These raw counts were subsequently extracted using SAMtools (Li *et al.*, 2009). The reads from each sample mapped to any of the sequences of the same group of Panicoideae co-orthologs were summed. This generated an accurate estimate of transcript abundance that is comparable between samples and species. Indeed, reads can map back to sequences from the same population, but also to sequences from other closely related individuals when the population is not represented. This minimises the effect of sequence variation on mapping success and subsequent differential expression analyses.

A multivariate analysis was used to assess similarities and differences in overall transcriptome expression profiles between samples and experiments. Groupings of expression profiles based on the biological coefficient of variation (BCV) were identified with multidimensional-scaling (MDS) in edgeR v3.4.2 (Robinson et al., 2010). Subsequent differential expression analysis in edgeR was restricted to the ten populations with three biological replicates. For each pair of populations, differentially expressed genes were identified as those with an associated false discovery rate (FDR) below 0.05. The overlap between pairwise comparisons was used to identify changes associated with specific branches of the phylogenetic tree inferred from core orthologs. Changes were assigned to a branch if significant results were detected for all pairwise tests involving one member of the descending clade and one population outside the clade, and the direction of expression change was consistent. This summary of pairwise tests was done separately for each C_4 clade (*A. cimicina*, *A. angusta*, and *A. semialata*) with all C_3 populations so that convergent gene expression shifts could be detect. It was also repeated with all populations to get a general estimate of changes in gene expression levels across the phylogenetic tree. Overall, our phylogenetic inference of changes in expression level allows the identification of the changes that coincide with physiological transitions and those that precede or follow them.

While protein abundance is not a direct function of gene expression, the two are correlated (Schwanhäusser et al. 2011; Csárdi et al. 2015; Koussounadis et al. 2015). Transcriptome comparisons therefore offer a first assessment of the changes underlying adaptive transitions, allowing subsequent investigations of post-transcriptional processes.

Results

Transcriptome sequencing and assembly

Over 181 million 108-bp paired-end reads were generated in this study, including more than 167 million for the ten populations sampled in triplicates (Table II.S3). For these 30 samples used in differential expression analyses, the data comprised 36.13 Gb, with a mean of 1.20 Gb per library (SD=0.54 Gb; Table II.S3). Over 95% of reads were retained after cleaning, and a *de novo* transcriptome was assembled for each of the populations using all available reads. The transcriptomes were of comparable quality, with a mean of 85,491 trinity 'unigenes' (SD=19,595), 123,719 contigs (SD=30,991), and a 1,336 bp N50 (SD=189 bp; Table II.S4). Another 14 RNA-Seq libraries and assemblies were included (Table II.S1). CDS were predicted for each of the 24 transcriptome assembly, with a mean of 28,698 CDS longer than 500 bp per population (SD=6,767; Table II.S4).

Phylogenetic relationships based on genome-wide markers

A phylogenetic tree was inferred from a concatenated alignment of 516 'core-orthologs' extracted from the 24 transcriptomes, for a total of 682,137 bp after cleaning. Each population was represented by at least 388,935 bp (mean=549,179 bp; SD 60032 bp). The phylogenetic relationships were congruent with previous genome-wide nuclear trees (Olofsson et al., 2016), and confirmed that all the sampled C_4 populations of *A. semialata* form a monophyletic group, which is sister to the C_3+C_4 populations (Fig. II.1). These two are in turn sister to the C_3 populations, so that previously inferred nuclear clades I (C_3), II (C_3+C_4), and III and IV (both C_4) are retrieved, with the polyploid populations (RSA3 and RSA4) branching in between (Olofsson et al., 2016; Fig. II.1). The position of the population from Cameroon (CMR1) with respect to clades III and IV was poorly resolved (Fig. II.1). *A. angusta* and *A. cimicina* branched

successively outside of *A. semialata* (Fig. II.1), mirroring previous results (Lundgren et al. 2015; Olofsson et al. 2016).

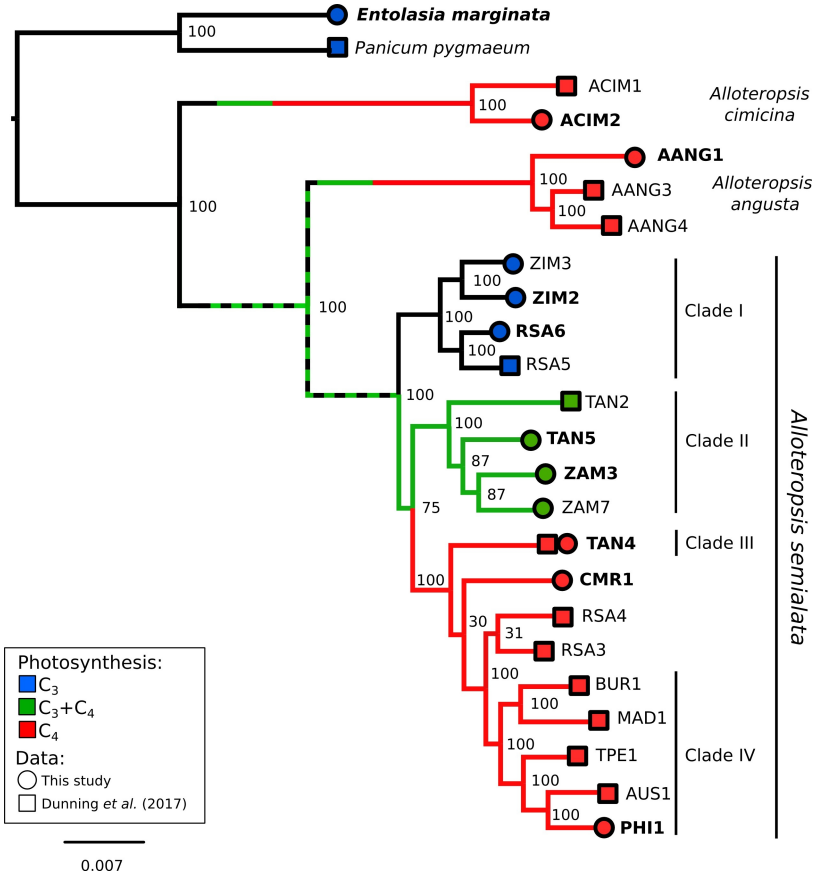


Figure II.1: Phylogenetic tree inferred from multiple nuclear markers.
This phylogeny was inferred under maximum likelihood using transcriptome-wide markers. Bootstrap values are indicated near branches. Population names indicate country of origin, with sample details available in Tables II.S1 and II.S2. Names of populations included in the differential expression analyses are in bold. Scale indicates number of nucleotide substitutions per site, and bootstrap support values are indicated near nodes. Nuclear clades from Olofsson *et al.* (2016) are indicated. Branch colours indicate the ancestral photosynthetic types, based on the transcriptomes and leaf anatomy detailed investigations of Dunning et al. (2017). The hashed green at the based of *A. semialata* indicates uncertainty between C_3 and C_3+C_4 states.

Phylogenetic identification of Panicoideae co-orthologs

Gene families were identified based on previously inferred homology among genes from eight publicly available genomes (Vilella *et al.*, 2009). The analyses were restricted to 6,407 gene families that contained sequences from *Arabidopsis* and at least one Panicoideae grass (*S. italica* or *S. bicolor*) in the Ensemble plant database (release version 30). These families covered a mean of 22,485 (SD=4,128) protein-coding genes for each model species in this dataset (78% of known *Arabidopsis* coding sequences). In

total, 5,454 (85.1%) of the gene families had a match with at least one transcriptome sequence from this study. After stringent alignment and trimming, 5,247 alignments for gene families were retained for downstream analyses, with a mean length of 814 bp (SD=545 bp), although monophyletic groups of Panicoideae were not necessarily detected in all of them. The phylogenetic annotation assigned 410,260 (54.9%) CDS from the transcriptomes to one of the 13,711 groups of Panicoideae co-orthologs (5,133 different gene families). In total, 23.0% of co-orthologs were represented by all 24 genotypes, and over 50% of co-orthologs were represented by at least 20 genotypes (Fig. II.S2).

On average, 24.9% of cleaned reads per individual mapped to at least one of the trimmed alignments of co-orthologs (Tables II.S1 and II.S3). This relatively low proportion of mapped reads results from the trimming of the reference alignments, which removed UTRs and poorly aligned regions, such as un-spliced introns. Indeed, a mapping performed on the alignments before trimming resulted in an average of 66.1% of cleaned reads per individual mapped to the same 410,260 sequences (Tables II.S1 and II.S3), which is similar to numbers achieved in other studies (e.g. Bräutigam et al. 2014). The remaining unmapped reads likely belong to genes that lack homologs in *Arabidopsis* (including contaminants from endosymbionts and transcriptome artefacts) or produced alignments that were too short after cleaning.

Transcriptome-wide patterns

Based on their expression profiles, samples group strongly by species, and the two C_3 outgroups cluster with the early diverging *A. cimicina* (Fig. II.2A). When focusing on *A. semialata*, the main phylogenetic groups are recovered, which match the photosynthetic types (Fig. II.2B). The Cameroonian samples (CMR1) are at a position intermediate between the C_4 and C_3+C_4 groups (Fig. II.2B). There is no apparent effect of the experiment, with previous and new transcriptomes of the same species grouping together (Fig. II.2). However, the Madagascan population (MAD1) appears as an outlier in Fig. II.2B, which might reflect an effect of sequencing depth as this sample had the smallest amount of data (<0.5 Gb).

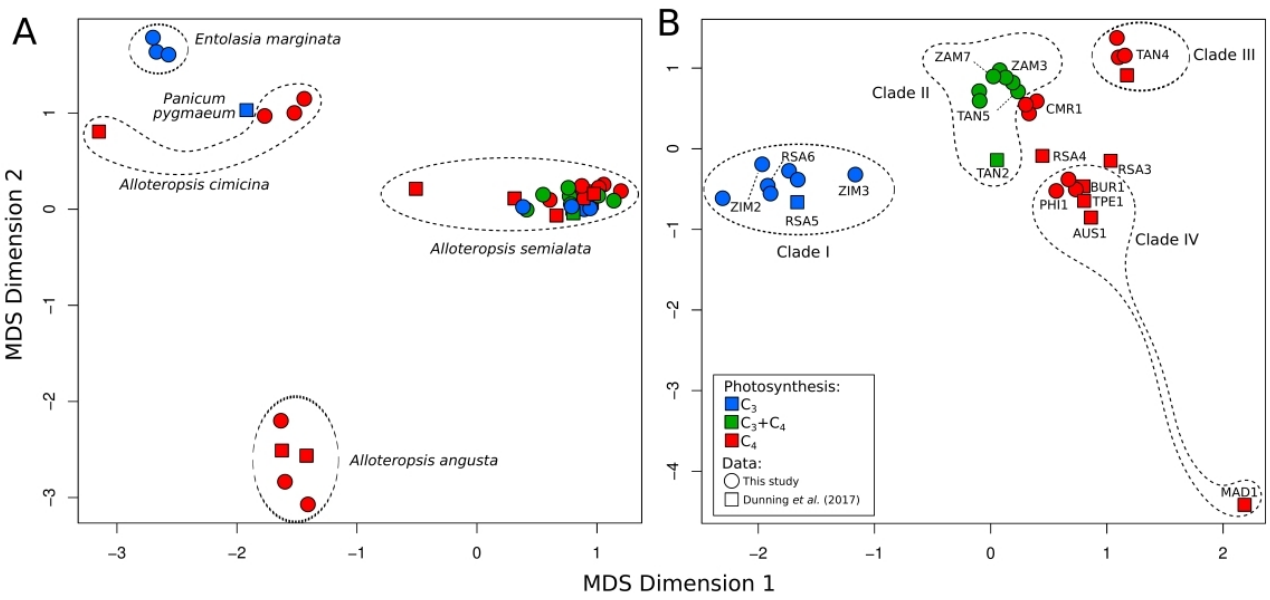


Figure II.2: Expression profile similarity across all samples.
Expression profiles are clustered in multidimensional scaling (MDS) plots using (A) all samples (B) only *A. semialata* samples. Species and nuclear clades from Olofsson *et al.* (2016) are delimited. Sample details can be found in Tables II.S1 and II.S2.

Differential expression analysis was performed for each pair of the ten populations that had three biological replicates. The 45 pairwise tests performed returned an average of 2,406 (SD=1,226) significantly (FDR<0.05) differentially expressed co-orthologs (Tables II.S5 and II.S6). The number of differentially expressed genes again matched the phylogenetic structure, being the highest between *E. marginata* and all others, and the smallest between groups of closely related *A. semialata* (Fig. II.3). Out of 13,475 genes, 14.3% (n=1,935) have conserved expression patterns across the phylogeny, not being significant in any of the pairwise comparisons (Tables II.S7 and II.S8). Another 11.0% (n=1,484) are shifted once in the phylogeny (Table II.S8; Fig. II.S3). The rest of the genes (74.5%, n=10056) are significant in at least one of the 45 pair-wise tests, but lack phylogenetic signal. The majority of these are significant in one or just a few pairwise comparisons (Table II.S8), and therefore likely represent false positives due to random variation among samples.

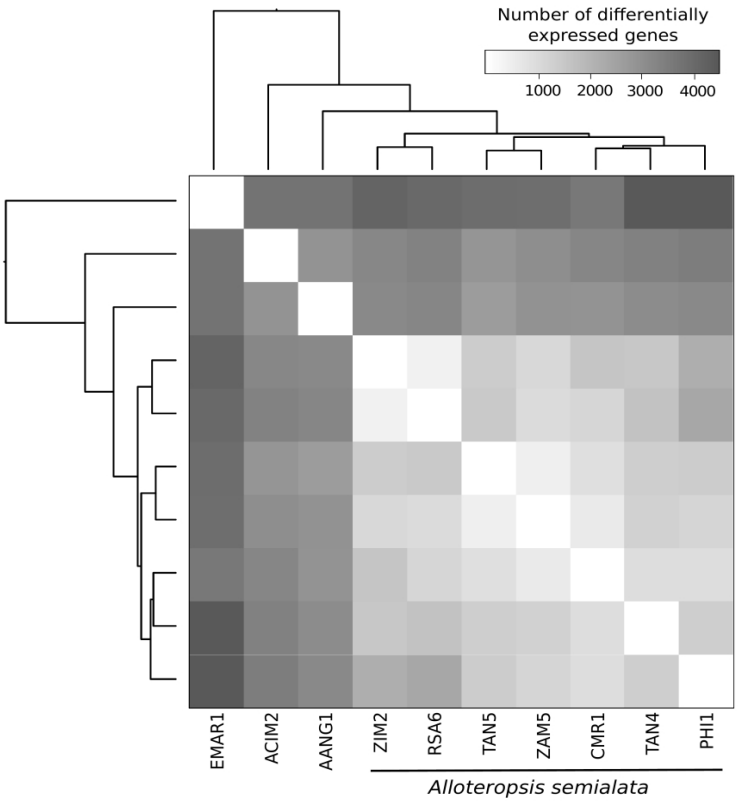


Figure II.3: Number of differentially expressed genes among pairs of populations.
The heatmap shows the number of significantly differentially expressed genes detected for each pair of populations. The phylogenetic relationships among populations are indicated on the side, using an ultrametric version of the tree presented in Fig. II.1. Sample details can be found in Tables II.S1 and II.S2.

Differences between the C_3 and C_3+C_4 states of A. semialata

As expected, the long divergence time between the C_3 outgroup (*Entolasia marginata*) and *A. semialata* results in a large number of significant expression changes (Fig. II.4). A total of 621 are downregulated along this branch, including two genes for phosphoenolpyruvate carboxylase (PEPC; *ppc-1P2* and *ppc-2P1*; Table II.S6), which drop to barely detectable levels in all *A. semialata* accessions, and are therefore unlikely to be linked to photosynthetic diversification. Two genes linked to the photorespiratory cycle are also downregulated along the same branch, which might represent a reduction of photorespiration in the ancestor of *A. semialata*. The genes uregulated in all *A. semialata* compared to the C_3 outgroup include one encoding PEPC (*ppc-1P3*), that reaches higher expression levels in C_3 accessions followed by further increases in C_3+C_4 and C_4 accessions (Fig. II.5), and might represent an early upregulation that was later

strengthened. One gene encoding malate dehydrogenase (NAD-MDH; *nadmdh-2P4*), one encoding adenosine monophosphate kinase (AK, *ak-3P3*), and one encoding glyceraldehyde 3-phosphate dehydrogenase (GAPDH, *gapdh-1P2*) were also upregulated along this branch, reaching moderate levels in all *A. semialata* independently of their photosynthetic type (Table II.S6). Finally, one gene for an enzyme linked to the photorespiratory pathway (*hpr-2P3*) was upregulated along the same branch, although levels remained very low within *A. semialata* (Table II.S6). The rest of the numerous genes varying in expression between the whole of *A. semialata* and the outgroup do not have known links to the C_4 pathway. A total of 30 genes are differentially expressed along the branch leading to the C_3 populations of *A. semialata* (Fig. II.4). None of these 30 genes encodes a protein known to function as part of the C_4 pathway (Table II.S6).

Within *A. semialata*, a weak C_4 cycle characterizes the monophyletic group of C_3+C_4 and C_4 populations, but not its C_3 sister group (Fig. II.1). Relatively few modifications to the transcriptome happen along the branch leading to C_3+C_4 and C_4 accessions, with only 16 significantly differentially expressed co-orthologs (Fig. II.4). Of those, only 12 are consistently upregulated in the C_3+C_4 and C_4 populations compared to the C_3 samples, including three genes that encode key C_4 enzymes: aspartate aminotransferase (ASP-AT), *aspat-3P4*; PEPC, *ppc-1P3*; and phosphoenolpyruvate carboxykinase (PCK), *pck-1P1* (Tables II.1 and II.S6). These three genes reach very high levels in the leaves of all C_3+C_4 and C_4 individuals, including the C_4 congener *A. angusta* (Fig. II.5; Tables II.1 and II.S6).

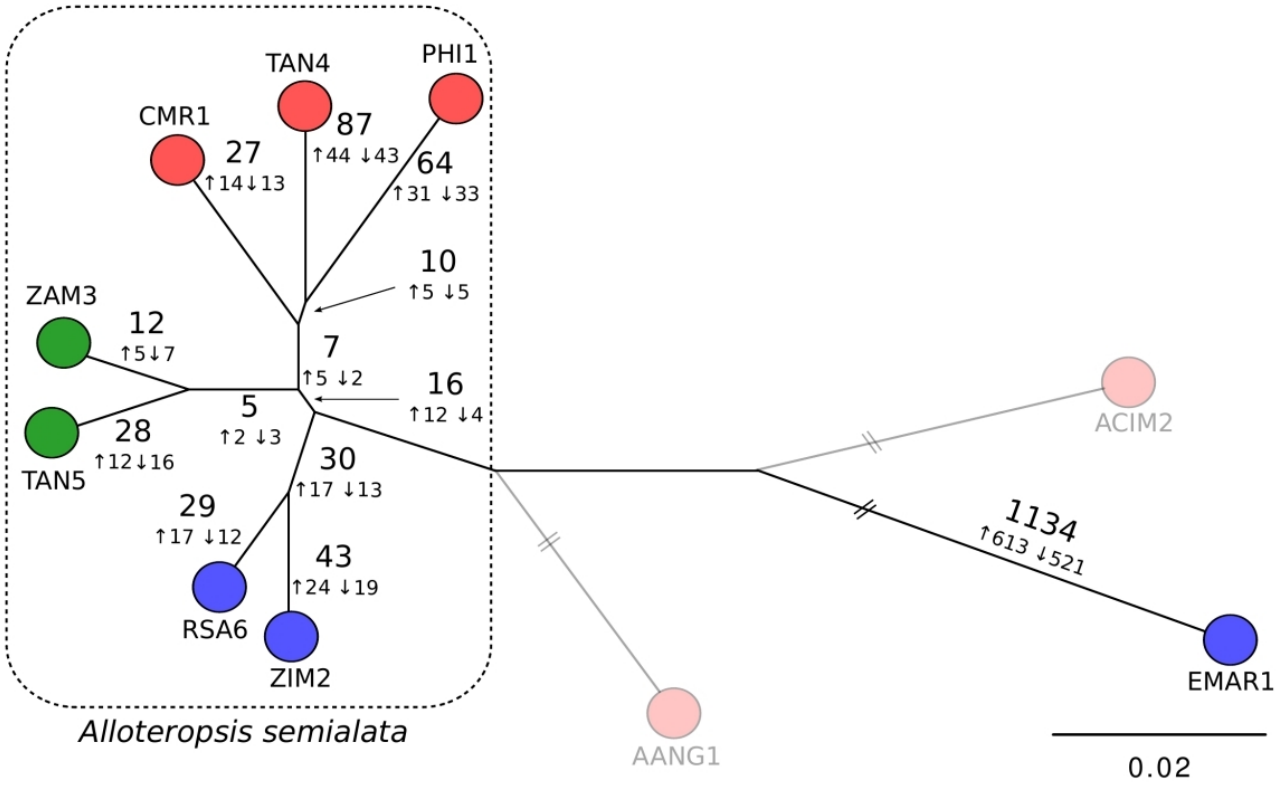


Figure II.4: Phylogenetic patterns of changes in gene expression.
For each branch of the unrooted phylogeny from Fig. II.1 showing only the populations used for expression analyses, the number of differentially expressed genes is indicated, with numbers next to arrows indicating those that are consistently up- or down-regulated. Each population has three biological replicates, and colours indicate the photosynthetic type (blue = C_3; green = C_3+C_4; red = C_4). Scale indicates number of nucleotide substitutions per site, with truncated branches highlighted by two bars. The two greyed out C_4 congeners were excluded from these analyses, and results that involve them can be found in Fig. II.S3.

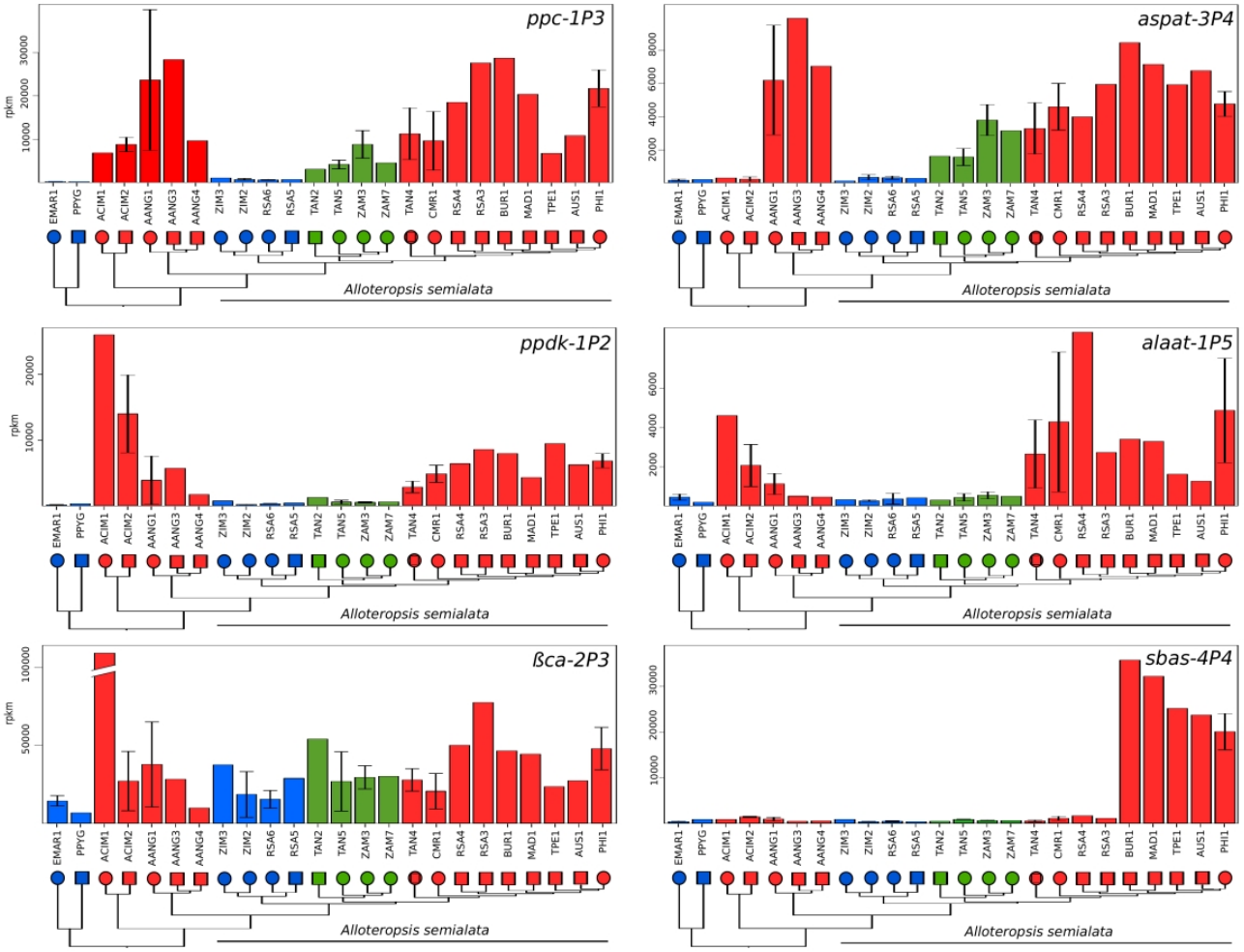


Figure II.5: Expression levels of selected genes across the phylogeny.
Expression levels in reads per kilobase of transcript per million mapped reads (RPKM) are shown for six example genes. Standard deviation for populations with biological replicates is indicated. Phylogeny is based on Fig. II.1; circles = sequenced in this study; squares = sequenced as part of Dunning et al. 2017; blue = C3; green = C3+C4; red = C4.

Changes during the transition from C_3+C_4 to C_4 in A. semialata

Within *A. semialata*, a strong C_4 cycle characterizes a monophyletic group of populations (Fig. II.1), but only seven co-orthologs were significantly differentially expressed along the branch separating this group from the other populations (Fig. II.4). Of these, only five were consistently upregulated in the C_4 populations, including two genes encoding core C_4 enzymes: pyruvate orthosphate dikinase (PPDK), *ppdk-1P2;*

and alanine aminotransferase (ALA-AT), *alaat-1P5* (Tables II.1 and II.S6). These genes reach very high levels in all C_4 populations, including the congeners *A. cimicina* and *A. angusta*, while they are present at moderate to low levels in all C_3 or C_3+C_4 populations (Fig. II.5; Table II.S6). A third gene upregulated in the C_4 *A. semialata* encodes AK (*ak-1P1*), an enzyme whose activity is linked in the C_4 cycle to PPDK. Two other genes were consistently downregulated in the C_4 populations, including one that encodes a peroxisomal photorespiratory enzyme (AGT; *agt-1P1*).

Adaptation of C_4 photosynthesis in independent lineages

The three C_4 populations included in the differential expression analyses come from geographically distant locations and diverged more than half a million years ago (Lundgren et al., 2015; Olofsson et al., 2016), explaining the high number of differentially expressed genes among them (Fig. II.3). Interestingly, several of these are potentially linked to the C_4 cycle, including genes encoding a NAD-MDH (*nadmdh-1P8*), a putative sodium bile acid symporter (SBAS; *sbas-4P4*), and phosphoenolpyruvate carboxylase kinase (PEPC-K; *pepck-1P3*), which are all upregulated in the C_4 from the Philippines (Table II.S6). A comparison of expression levels in the other transcriptomes indicates that the gene *sbas-4P4* is upregulated in all C_4 individuals from clade IV of *A. semialata*, but not in any other (Fig. II.5). This gene is orthologous to *Arabidopsis* BASS6 (At4g22840), which has been shown to have the ability to transport glycolate, and be involved in a process decreasing photorespiration (South et al., 2017). The *Arabidopsis* paralog previously related to C_4 photosynthesis transports pyruvate (BASS2; Furumoto *et al.*, 2011), but the precise function might differ between the *Alloteropsis* and *Arabidopsis* orthologs. In addition, a gene for a pyruvate kinase (PK; *pk-1P1.2*) is upregulated specifically in the Cameroonian C_4 samples (CMR1), while some photorespiratory genes are downregulated in only one of the three C_4 populations (Table II.S6). Finally, one of the genes encoding the NADP-malic enzyme (*nadpme-1P4*; NADP-ME) reaches high levels in some *A. semialata* populations (Table II.S5). It is significantly upregulated in *A. cimicina* and *A. angusta* and the differential expression analysis is significant for a number of pairwise comparisons involving C_3 and C_4 populations of *A. semialata* (Table II.S6). However, its level varies both among and within populations of *A. semialata* and *A. angusta*,

reaching in some C_4 individuals levels that are as low as in the C_3 plants (Table II.S5).

The number of genes significantly differentially expressed in the C_4 *A. cimicina* and *A. angusta* lineages is much higher due to only a single population representing these species (Fig. II.S3). As previously reported (Dunning et al., 2017), a high number of genes encoding core C_4 enzymes, regulatory proteins and transporters are upregulated in *A. cimicina*, and to a lesser extent in *A. angusta*, while some photorespiration genes are downregulated in these two species (Table II.S6). Besides the differentially expressed genes, a number of C_4-related genes are abundant in all samples independently of their photosynthetic type. This is especially the case of those for β-carbonic anhydrase (*βca-2P3*; Fig. II.5) and malate dehydrogenases (*nadpmdh-1P1*, *nadpmdh-3P4*, and *nadmdh-3P5*; Table II.S6). Transcripts for these genes were also abundant in the leaves of distantly related C_3 grasses, and their upregulation very likely predates the divergence of the group (Moreno-Villena et al. 2018).

Discussion

Sampling the natural diversity to limit false positives

RNA-Seq is routinely used to identify genes differentially expressed between individuals with distinct phenotypes, leading to candidate genes underpinning these differences (e.g. Shen et al., 2014; Dunning et al., 2016; Fracasso et al., 2016). When comparing distinct species, the risk of false positives is very high, as all changes in gene expression unrelated to the studied phenotypic transitions are detected. Here, 74.5% of genes are significantly differentially expressed in at least one pairwise comparison between our ten populations, which all belong to a relatively small group of closely related grasses. A powerful strategy to reduce false positives is to consider multiple independent origins of the trait of interest, and retain only those genes differentially expressed in all lineages (Rao et al., 2016). Such an approach would however miss non-convergent changes in gene expression, and would not allow deciphering of changes that occurred in specific stages of an evolutionary transition.

The alternative approach adopted here was to carry out multi-individual comparisons to infer changes along specific branches of the phylogenetic tree. The problem of false positives remains, as changes coinciding with the studied transitions would also be detected. However, working within a species complex decreases the

number of false positives, as lower divergence times result in fewer unrelated changes in gene expression. Because most changes cluster on terminal branches (Figs II.4 and II.S3), probably representing neutral changes that do not persist over evolutionary time, the inference of changes on short internal branches is more accurate. Indeed, a comparison of a C_3 *A. semialata* with the C_4 sister species *A. angusta* would identify over 3,000 differentially expressed genes (Fig. II.3; Table II.S5). This number drops by approximately 50% when comparing individual C_3 and C_4 populations within *A. semialata*, but this still includes all changes that occurred before and after the C_3 to C_4 transition.

After incorporating multiple populations of each type, only 16 genes differing in expression between the C_3 and C_3+C_4 phenotypes are identified, and seven between the C_3+C_4 and C_4 states. These genes represent the best candidates for a role in the emergence and subsequent strengthening of a C_4 cycle in the group, and the over-representation of core C_4 enzymes among them confirms the power of our approach consisting in sampling the diversity of realized phenotypes.

Emergence and reinforcement of the C_4 cycle in Alloteropsis semialata

The phylogenetic relationships and genus-wide comparisons of transcriptomes and leaf anatomical traits indicate that the last common ancestor of all *A. semialata* might have possessed a weak C_4 cycle based on the upregulation of some enzymes (Fig. II.1; Dunning et al. 2017). A high number of genes are differentially expressed between all *A. semialata* and the C_3 outgroup, which is unsurprising given the evolutionary distance. However, these include only a few genes encoding C_4 enzymes (Fig. II.5; Table II.S6). We conclude that the transcriptome of the C_3 *A. semialata* differs from that of other C_3 grasses by relatively few C_4-related genes. The C_3 group might still represent a reversal from a C_3+C_4 state, but this eventuality is covered in our phenotype comparisons, which would assign C_4-related changes in the last common ancestor of *A. semialata* to the C_3+C_4 and C_4 groups if the C_3 *A. semialata* achieved expression levels similar to the C_3 outgroup. Our transcriptome comparisons therefore provide an accurate quantification of the phenotypic gaps in gene expression between the C_3 state and those using a weak or strong C_4 cycle, which is not heavily influenced by potential evolutionary reversals or reticulate evolution.

Only 16 genes are differentially expressed in the group encompassing C_3+C_4 and C_4 phenotypes, and these include three genes encoding core C_4 enzymes that are upregulated in all C_3+C_4 and C_4 individuals (genes for ASP-AT, PCK and PEPC; Table II.1; Fig. II.5; Table II.S6). These three enzymes form the PCK shuttle, which theoretically cannot sustain a full C_4 pathway on its own without creating an energetic imbalance among cell types (Wang et al. 2014). However, it might create a weak CO_2-concentrating mechanism in C_3+C_4 plants that can function without dramatic energetic consequences due to its coexistence with a C_3 type of photosynthesis (Fig. II.6). Other small adjustments of the cellular metabolism might remain undetected, but none of the other major C_4 enzymes or transporters are significantly upregulated during the emergence of a weak C_4 cycle (Tables II.1 and II.S6). The apparently few genetic changes required to operate a weak C_4 cycle in the C_3+C_4 intermediates may be facilitated by C_4-like anatomical properties and an abundance of genes for some key enzymes in the ancestor, as observed specifically in the C_3 *A. semialata* (e.g. *ppc-1P3*; Fig. II.5), in the C_3 outgroups *E. marginata*, and *P. pygmaeum* (e.g. ßca-2P3*;* Fig. II.5), and in other C_3 grasses (Christin et al., 2013; Dunning et al., 2017; Moreno-Villena et al. 2018), and recent evidence suggests that some anatomical traits themselves might emerge via very few genetic changes (Wang et al. In press). While the C_4 cycle of the C_3+C_4 intermediates is weak, it may be responsible for the apparent reduction of photorespiration (Lundgren *et al.*, 2016). This confers a selective advantage similar to that of a complete C_4 cycle in tropical conditions (Sage *et al.*, 2012; Christin & Osborne, 2014; Lundgren & Christin, 2016), and allows the evolution of a stronger C_4 cycle under natural selection for faster biomass accumulation (Heckmann et al. 2013; Mallmann *et al.*, 2014; Bräutigam & Gowik 2016).

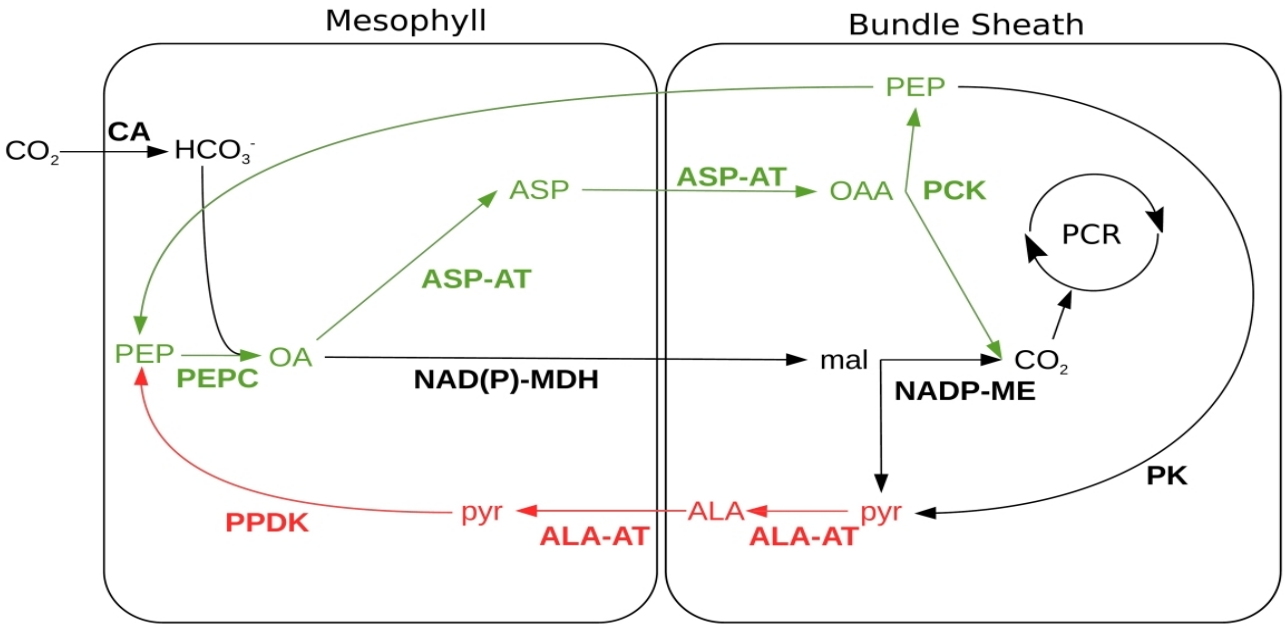


Figure II.6: Putative C_4 pathway in *Alloteropsis semialata*
A C_4 cycle is suggested for *A. semialata* based on the transcript abundance of C_4-related genes, and the literature (Frean et al., 1983; Ueno and Sentoku 2006). Pathway components are coloured per the differential expression analysis, with those in black being putatively sufficiently abundant in C_3 ancestors, parts of the pathway in green upregulated during the transition to C_3+C_4, and parts in red upregulated during the transition from C_3+C_4 to C_4. ALA-AT = alanine aminotransferase, ASP-AT = aspartate aminotransferase, CA = carbonic anhydrase, NADP-MDH = NADP malate dehydrogenase, NAD(P)-ME = NAD(P) malic enzyme, PCK = phosphoenolpyruvate carboxykinase, PEPC = phosphoenolpyruvate carboxylase, PEPP = phosphoenolpyruvate phosphatase, PK = pyruvate kinase, PPDK = pyruvate orthosphate dikinase, PCR = photosynthetic carbon reduction (Calvin-Benson cycle).

Table II.1: List of genes differentially expressed in key comparisons within *Alloteropsis semialata* from C_3 to C_3+C_4, and C_3+C_4 to C_4.

Regulation	Gene (Homolog_#)	Mean rpkm (± s.d.)[1] C_3	C_3+C_4	C_4	*Arabidopsis* ortholog	Brief description[2]
Up in C_3+C_4 and C_4	29490-G06	678 (±175)	5401 (±3079)	15673 (±8065)	AT2G42600 *	**Core C_4 enzyme:** Phosphoenolpyruvate carboxylase
	06567-G01	276 (±106)	3793 (±2236)	8351 (±4280)	AT5G65690, AT4G37870	**Core C_4 enzyme:** Phosphoenolpyruvate carboxykinase
	33440-G04	290 (±122)	2486 (±1230)	4976 (±1779)	AT1G62800, AT5G19550, AT5G11520	**Core C_4 enzyme:** Aspartate aminotransferase Asp4
	48666-G05	25 (±7)	172 (±71)	209 (±179)	AT3G12600	Nudix hydrolase homolog 16
	31859-G01	59 (±21)	189 (±72)	148 (±49)	AT2G18290	APC10 anaphase promoting complex 10. Plays an essential role in cell proliferation during leaf development
	01920-G08	1 (±1)	193 (±197)	156 (±147)	AT1G51190, AT5G17430, AT3G20840	Transcription factor belonging to the AINTEGUMENTA-like (AIL) subclass of the AP2 EREBP family
	16932-G04	23 (±8)	114 (±72)	100 (±59)	AT5G23360 *,AT5G23370*, AT5G23350 *,AT5G08350*	GRAM domain-containing protein ABA-responsive protein-like protein
	30221-G06	0 (±0)	66 (±49)	133 (±69)	AT4G18770	Transcription factor MYB98
	17622-G01	13 (±4)	91 (±61)	68 (±45)	AT3G02065	P-loop containing nucleoside triphosphate hydrolases superfamily protein
	08213-G09	0 (±0)	30 (±21)	47 (±42)	AT3G48350, AT3G48340, AT5G50260	Cysteine proteinases superfamily protein involved in tapetal programmed cell death and pollen development
	06020-G11	0 (±0)	33 (±42)	38 (±52)	AT2G03200 *	Eukaryotic aspartyl protease family protein
	11314-G11	0 (±0)	8 (±10)	37 (±33)	AT1G10370, AT1G69930, AT1G69920, AT1G59700, AT1G59670, AT1G27130, AT1G27140, AT1G10360	Glutathione transferase belonging to the tau class

Down in C_3+C_4 and C_4	36691-G02	182 (±105)	4 (±5)	2 (±4)	AT3G04770 *	40s ribosomal protein SA B
	31440-G07	73 (±19)	10 (±8)	9 (±8)	AT1G18590, AT2G03760, AT2G03770, AT1G74100, AT1G74090	Sulfotransferase with sulfating activity toward flavonoids, involved in the final step of glucosinolate core structure biosynthesis
	49585-G02	75 (±53)	8 (±8)	7 (±5)	AT3G02960 , AT5G63530, AT5G50740	Heavy metal transport detoxification superfamily protein
	09708-G05	14 (±11)	0 (±0)	0 (±0)	AT4G10310	Sodium transporter HKT1 expressed in xylem parenchyma cells
Up in C_4	29424-G01	315 (±185)	702 (±373)	5582 (±2262)	AT4G15530	**Core C_4 enzyme:** Pyruvate orthophosphate dikinase
	19748-G07	320 (±162)	447 (±152)	3704 (±2543)	AT1G72330, AT1G17290	**Core C_4 enzyme:** Alanine aminotransferase ALAAT2
	02503-G07	267 (±153)	391 (±271)	1262 (±378)	AT5G47840	**Accessory C_4 enzyme:** Adenosine monophosphate kinase
	20625-G12	17 (±17)	70 (±52)	242 (±128)	AT1G17020, AT4G25310, AT4G25300, AT1G17010, AT1G78550	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
	27406-G04	0 (±0)	0 (±0)	76 (±20)	AT2G47490 *	Chloroplast-localized NAD+ transporter that transports NAD+ in a counter exchange mode with ADP and AMP in vitro.
Down in C_4	24452-G01	1077 (±499)	918 (±228)	292 (±156)	AT2G13360	**Photorespiratory enzyme:** Peroxisomal photorespiratory enzyme catalysing transamination reactions with multiple substrates
	02733-G05	15 (±13)	21 (±8)	0 (±0)	AT5G06750	Phosphatase 2C family protein

[1] Raw values can be found in Table II.S6. [2] The brief description is based on the *Arabidopsis* ortholog (*=paralog if no direct ortholog) annotation.

The transition from a weak to a strong C_4 cycle in *A. semialata* switches carbon isotope signatures (the method most often used to identify photosynthetic types) from non-C_4 values to values diagnostic of C_4 plants (von Caemmerer, 1992; Lundgren *et al.*, 2015). This key transition occurred through the upregulation of relatively few genes encoding C_4-related enzymes (AK, ALA-AT, and PPDK). Based on the literature and our transcriptome data, the C_4 cycle of *A. semialata* relies on a minimum of eight enzymes, and we propose a possible involvement of PK (Fig. II.6; Frean et al., 1983;

Ueno and Sentoku, 2006). Genes for some of these enzymes (PEPC, NAD-MDH, and AK) increased in the common ancestor of the whole group, potentially as part of an ancestral weak C_4 cycle (Fig. II.1; Dunning et al., 2017). Within *A. semialata*, further increases in transcript abundance are observed in the C_3+C_4/C_3 or C_4/C_3+C_4 comparisons (Table II.1) for genes encoding PEPC and four other enzymes (i.e. ALA-AT, ASP-AT, PCK, PPDK), while others are specific to some C_4 populations (i.e. PK). The expression of genes encoding two others (CA and NAD(P)-MDH) in the C_3 ancestor of the group might have been sufficient to sustain a functioning C_4 cycle (Tables 1 and S6; Moreno-Villena et al. 2018). Genes for the last of these enzymes (NADP-ME) are abundant in some C_4 individuals (Table II.S5), and might be expressed only in specific conditions, as suggested previously (Frean et al., 1983). C_4 populations of *A. semialata* are also characterized by a set of specific anatomical modifications that support the C_4 pump (Lundgren *et al.* 2016; Dunning et al. 2017). Genetic changes responsible for these modifications would not be captured by transcriptome analyses if they act during leaf development, and the evolution of the C_4 phenotype almost certainly involves more genetic changes than those altering the biochemical cycle. Our comparative transcriptomics therefore show that, once the required enablers are present, the transition between C_3 to C_3+C_4 with some C_4 activity, and C_3+C_4 to a rudimentary C_4 metabolism might require fewer changes in gene expression than previously suggested (Bräutigam *et al.*, 2011, 2014; Gowik *et al.*, 2011; Külahoglu *et al.*, 2014; Li *et al.*, 2015). These changes were spread between the C_3/C_3+C_4 and C_3+C_4/C_4 transitions, supporting a stepwise model of evolution (Mallmann *et al.*, 2014), where evolutionarily stable adaptive peaks can be reached with few mutations.

Adaptation continued after the emergence of a rudimentary C_4 pathway

The CO_2-pump generated by the C_4 cycle of *A. semialata* is less efficient than that of other C_4 species, as illustrated by the incomplete segregation of enzymes between different cell types (Ueno & Sentoku, 2006) and slightly elevated CO_2-compensation points lying at the upper limit of those observed in C_4 species (Lundgren *et al.*, 2016). Therefore, *A. semialata* may be considered to exhibit an incipient C_4 cycle, which has not been optimised through protracted evolutionary periods, as suggested in the most recent models (Bräutigam & Gowik 2016). The analyses conducted here, which compared all C_4 individuals to the C_3+C_4 or C_3 conspecifics, can detect the changes that

happened in the early C_4 members of the group, before the diversification of the C_4 genotypes. However, transcriptome comparisons across C_4 individuals of *A. semialata* show evidence of additional alterations of the leaf biochemistry subsequent to the initial emergence of a C_4 cycle, with the abundance of some C_4-related enzymes varying in abundance across C_4 populations (Fig. II.5), and photorespiratory proteins downregulated in only some of the C_4 populations (Table II.S6). These changes likely represent the adaptation of the C_4 cycle after its initial emergence, previously illustrated by variation in the identity of genes responsible for an abundance of the key C_4 enzyme PEPC across C_4 genotypes (Fig. II.S4; Dunning *et al.*, 2017).

The C_4 pathway proposed for *A. semialata*, based on the upregulation of few enzymes in addition to those present in C_3 ancestors (Fig. II.6), might serve as an intermediate stage toward more complex and more efficient C_4 cycles. The congeneric C_4 *A. cimicina* and *A. angusta* have transcriptomes more typical of C_4 species, with very high levels of numerous C_4-related enzymes, including a number of regulatory proteins and metabolite transporters, as would be predicted from other study systems, and an abundance of amino acid transitions adapting the proteins for the new catalytic context (Table II.S6; Bräutigam *et al.*, 2011, 2014; Gowik *et al.*, 2011; Mallmann *et al.*, 2014; Christin et al. 2015; Dunning *et al.*, 2017). These two species might have undergone more adaptive changes, due to an earlier C_4 origin or faster evolutionary rate.

Conclusions

In this study, we analyse transcriptomes in a phylogenetic context to show that the changes in gene expression required for a physiological innovation can be spread over long evolutionary time scales, with the relatively few changes required for the initial emergence of a new metabolism contrasting with the numerous modifications involved in the adaptation of the new pathway. Indeed, a weak C_4 cycle emerged in our study system through the upregulation of a handful of enzymes, and allowed the evolution of a stronger C_4 cycle via the upregulation of a few other genes. However, adaptation of C_4 photosynthesis, illustrated here by population-specific expression of C_4-specific enzymes, continues when the plants are already in a C_4 state, and is spread over a longer evolutionary period. The evolutionary time interval required to generate a rudimentary C_4 pathway can therefore be relatively short in species possessing C_4 enablers, but even

a suboptimal C_4 pathway is important as it changes the environmental responses of the species. This creates an opportunity for natural selection to act on the standing variation, new mutations and, in some cases laterally acquired genes, to assemble a trait of increasing complexity, allowing the colonization and gradual dominance in a larger spectrum of ecological conditions.

Acknowledgements

This paper is dedicated to the memory of Mary Ann Cajano, from the University of the Philippines at Los Banos, who helped with the identification of plant specimens. This work was funded by the Royal Society University Research Fellowship (grant URF120119) and the Royal Society Research Grant (grant number RG130448) to PAC. LTD is funded by a NERC grant (grant NE/M00208X/1), and JKO and MRL are funded by an ERC grant (grant number ERC-2014-STG-638333).

Chapter II: Supporting information

Figure II.S1: Geographic origins of the *A. semialata* individuals analyzed in this study.

The approximate worldwide distribution of the different photosynthetic types of A. semialata is indicated based on Lundgren et al. 2016, in blue for the C_3, green for the C_3+C_4, and red for the C_4. The origins of the samples analysed here is indicated with squares for those without replicates, and with circles for those sampled in triplicates, with colours indicating the photosynthetic types.

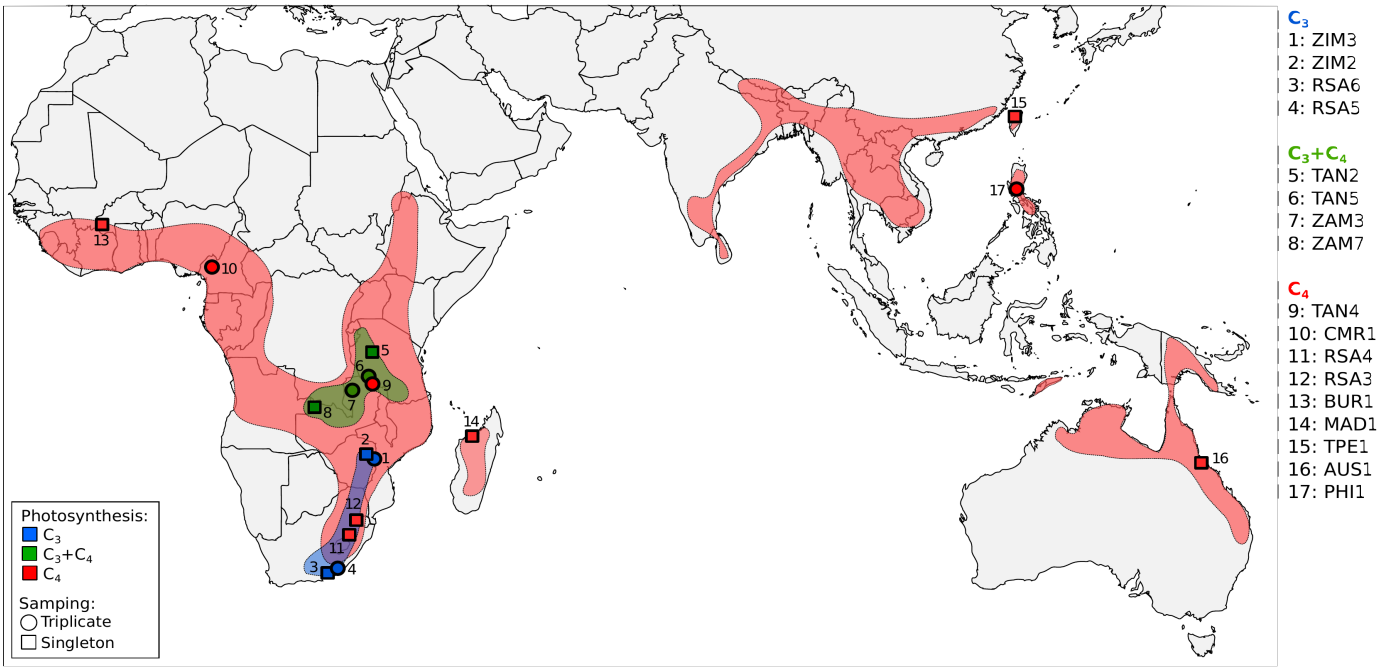


Figure II.S2: Distribution of the number of genotypes represented in each group of (co-)orthologs.

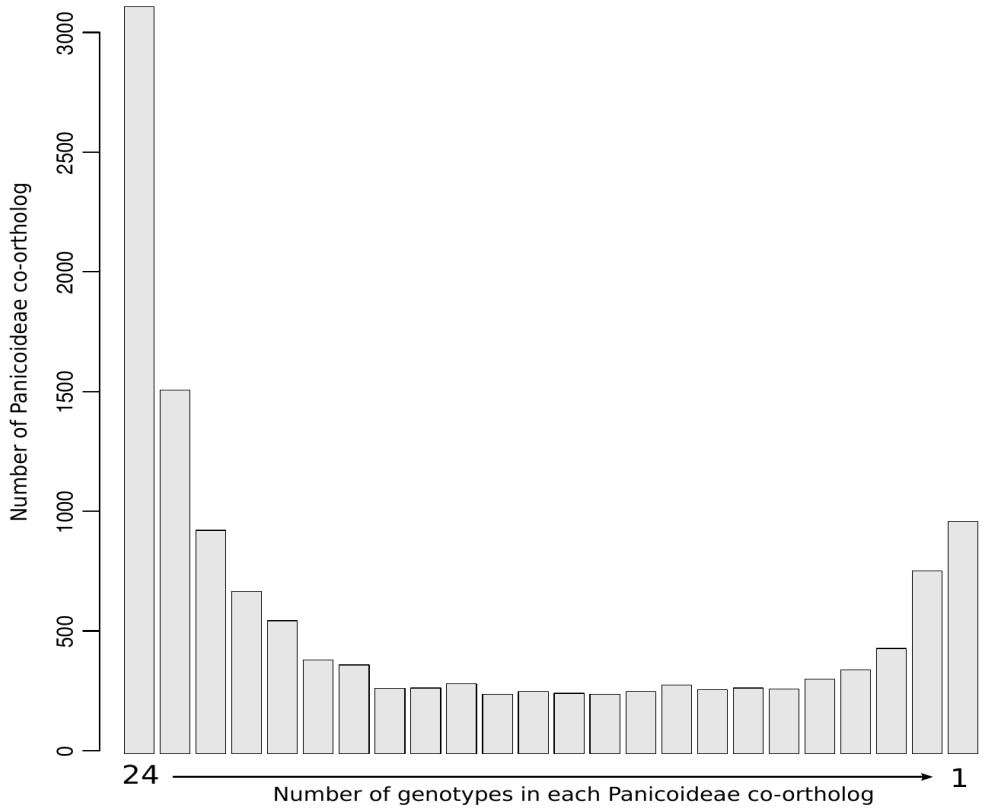


Figure II.S3: Phylogenetic patterns of changes in gene expression in (A) all populations, (B) *Alloteropsis angusta*, and (C) *Alloteropsis cimicina*. For each branch of the unrooted phylogeny from Fig. II.1, the number of differentially expressed genes is indicated, with numbers next to arrows indicating those that are consistently up- or down-regulated. Each population has three biological replicates, and colors indicate the photosynthetic type (blue = C_3; green = C_3+C_4; red = C_4). Scale indicates number of nucleotide substitutions per site, truncated branches are highlighted by two bars. For each tree, taxa excluded from the analyses are greyed.

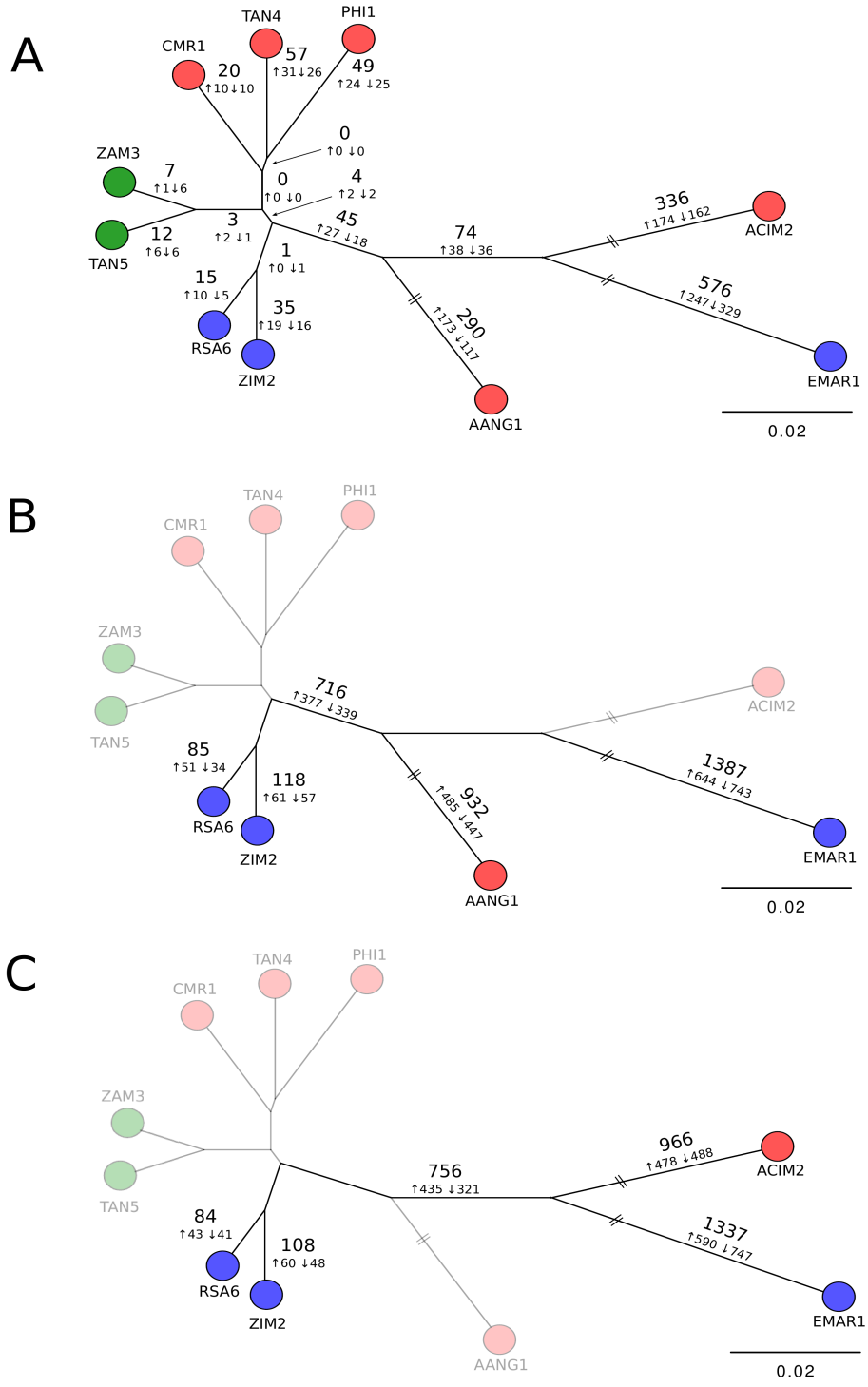


Figure II.S4: Expression levels of ppc isoforms in reads per kilobase of transcript per million mapped reads (RPKM). Colors indicate the photosynthetic type (blue = C_3; green = C_3+C_4; red = C_4).

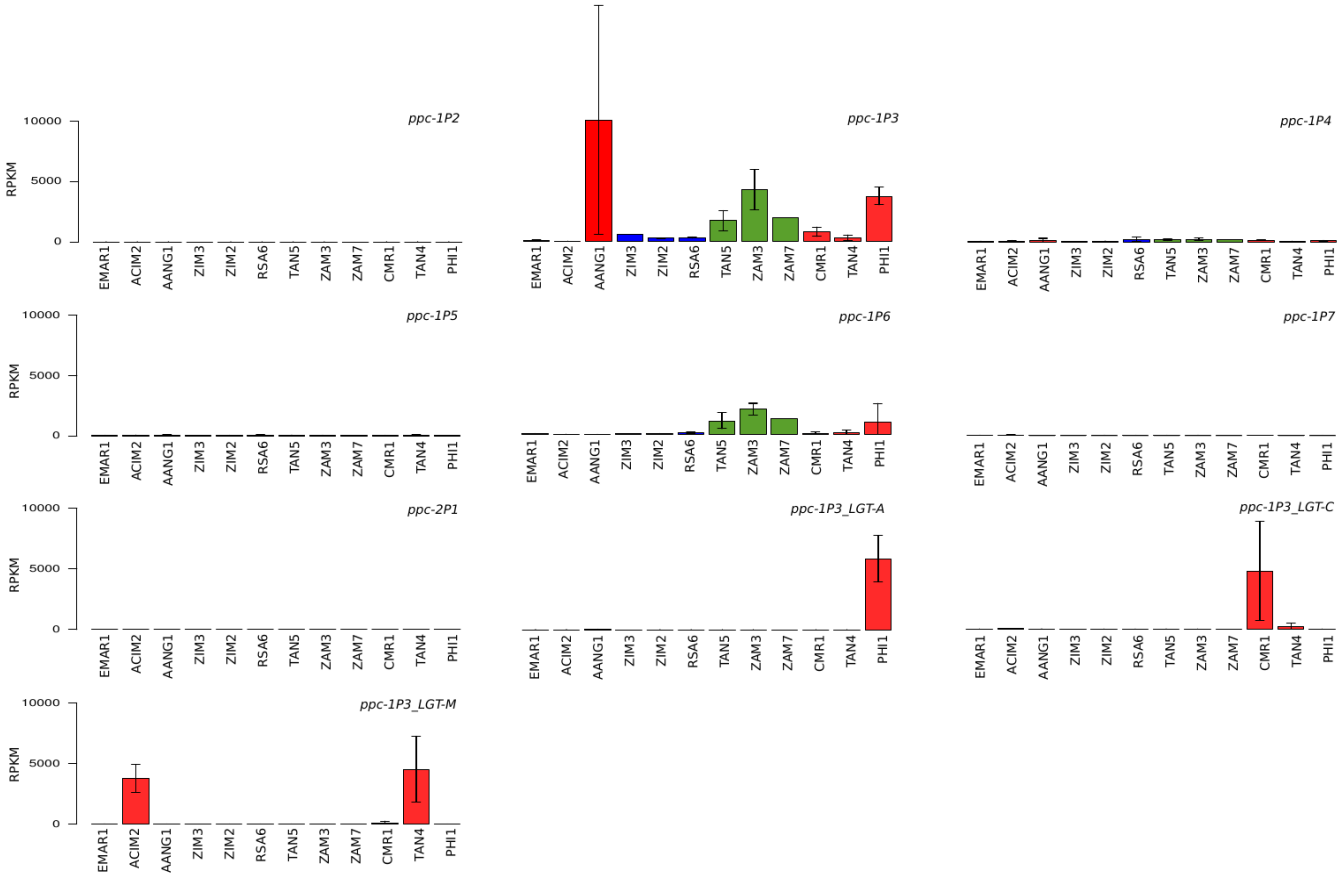


Table II.S1: Transcriptomes without replicates used in this study.

Species	Photo.	Pop.	Country	SRA accession[2]	TSA accession[2]	Percentage of clean reads mapped	
						Raw co-orthologs	Trimmed co-orthologs
Panicum pygmaeum	C_3	PPYG	Seed bank	SRR3323220[3]	GFYP00000000	65.43	28.53
Alloteropsis cimicina	C_4	ACIM1	Australia	SRR3994072[3]	GFYN00000000	69.66	30.38
Alloteropsis angusta	C_4	AANG33	Uganda	SRR3994075[3]	pending	72.67	30.95
		AANG48	Uganda	SRR3994077[3]	GFYLO0000000	65.45	26.70
Alloteropsis semialata	C_3	RSA5	South Africa	SRR3323049[3]	GFYL00000000	68.43	26.59
		ZIM3	Zimbabwe	pending	pending	68.36	28.72
	C_3+C_4	TAN2	Tanzania	SRR3323088[3]	GFYF00000000	68.71	27.85
		ZAM7	Zambia	pending	pending	69.37	22.56
	C_4	AUS1	Australia	SRR3322358[3]	GFYB00000000	67.97	27.53
		BUR1	Burkina Faso	SRR3322973[3]	GFYE00000000	65.00	27.37
		MAD1	Madagascar	SRR3323132[3]	GFYI00000000	67.06	20.47
		RSA3	South Africa	SRR3323137[3]	GFYJ00000000	67.13	26.11
		RSA4	South Africa	SRR3323220[3]	GFYK00000000	63.44	27.33
		TAN4[1]	Tanzania	SRR3323124[3]	GFYG00000000	67.36	29.13
		TPE1	Taiwan	SRR3323243[3]	GFYL00000000	71.23	30.19

[1] This population was resampled in triplicates (Table S2); [2] SRA = NCBI Sequence Read Archive; TSA = NCBI Transcriptome Shotgun Assemblies; [3] samples published in Dunning et al. 2017.

Table II.S2: Information for populations sampled in triplicates.

Species	Photo.	Population	Country	Lat/long
Entolasia marginata	C_3	EMAR1	Australia	-26.57,150.55
Alloteropsis cimicina	C_4	ACIM2	Madagascar	-
Alloteropsis angusta	C_4	AANG1	Uganda	0.34,31.89
Alloteropsis semialata	C_3	ZIM2	Zimbabwe	-18.42,32.77
		RSA6	South Africa	-33.32,26.53
	C_3+C_4	TAN5	Tanzania	-8.35,31.28
		ZAM3	Zambia	-10.23,29.83
	C_4	CMR1	Cameroon	5.93,10.62
		TAN4	Tanzania	-9.04,32.48
		PHI1	Philippines	15.94,121.01

Table II.S3: RNA-Seq data and mapping statistics for ten populations with triplicates.

					Percentage of clean reads mapped	
Pop.	Sample	raw PE reads	Gbp	Clean PE reads	Raw co-orthologs	Trimmed co-orthologs
ACIM2	ACIM2-1	5967276	1.29	5810460	66.36	27.89
	ACIM2-2	3501357	0.76	3419398	62.65	26.88
	ACIM2-3	3102005	0.67	3033994	66.99	20.82
EMAR1	EMAR1-2	3928822	0.85	3849825	62.82	27.33
	EMAR1-4A	3898055	0.84	3810471	66.24	25.14
	EMAR1-4B	2500722	0.54	2439927	68.21	24.68
AANG1	AANG1-6A	4618386	1.00	4332270	68.46	16.69
	AANG1-6B	5997583	1.30	5836626	62.52	19.25
	AANG1-6C	6009876	1.30	5866831	65.47	26.01
ZIM2	ZIM2-1	2815050	0.61	2622817	68.74	23.51
	ZIM2-5	3357287	0.73	3236198	69.95	17.72
	ZIM2-10	5308033	1.15	4989635	60.80	19.63
RSA6	RSA6-1	5284768	1.14	5193728	66.85	20.12
	RSA6-7	4397850	0.95	4299225	65.61	30.31
	RSA6-9	8129836	1.76	7964933	63.85	29.00
TAN5	TAN5-1	10558384	2.28	10304872	65.60	21.57
	TAN5-2	8101223	1.75	7890990	66.88	25.98
	TAN5-3	2623196	0.57	2494198	67.26	17.25
ZAM3	ZAM3-2	11178860	2.41	10897999	65.26	26.27
	ZAM3-4	5146745	1.11	4992771	70.22	23.10
	ZAM3-8	6118516	1.32	5853276	67.88	25.90
CMR1	CMR1-2	12525719	2.71	12273903	68.67	27.91
	CMR1-7	4385675	0.95	4282371	53.39	23.43
	CMR1-10	2724267	0.59	2652363	56.09	27.98
TAN4	TAN4-1	6581391	1.42	6393229	63.57	20.81
	TAN4-3	7547915	1.63	7335642	65.06	23.66
	TAN4-8	6089921	1.32	5932271	63.55	21.08
PHI1	PHI1-1	4911514	1.06	4705864	67.80	24.25
	PHI1-13	4643128	1.00	4441009	64.58	22.01
	PHI1-17	5167456	1.12	5003163	64.32	24.58

Table II.S4: Transcriptome assembly statistics for ten populations with triplicates.

Population	Trinity 'unigenes'	Trinity Contigs	N50	CDS
EMAR1	94080	138268	1486	39538
ACIM2	88770	148193	1131	38285
AANG1	100594	140272	1349	36465
ZIM2	62149	82565	1202	23087
RSA6	66024	91789	1577	30527
TAN5	78718	114745	1486	34838
ZAM3	86692	127283	1381	37637
CMR1	126654	181572	958	37979
TAN4	87506	126211	1455	36750
PHI1	63727	86299	1334	25155

Table II.S5: Number of significantly differentially expressed genes between populations.

	EMAR1	ACIM1	AANG1	ZIM2	RSA6	TAN5	ZAM3	CMR1	TAN4	PHI1
EMAR1	-									
ACIM1	3730	-								
AANG1	3746	2885	-							
ZIM2	4139	3212	3180	-						
RSA6	4026	3345	3265	374	-					
TAN5	3929	2857	2677	1381	1472	-				
ZAM3	3867	3027	2958	1075	979	426	-			
CMR1	3594	3257	2901	1531	1109	883	554	-		
TAN4	4480	3478	3145	2165	2348	1349	1144	920	-	
PHI1	4491	3371	3076	1520	1619	1324	1224	921	1321	-

Table II.S6: Leaf abundance, annotation, and summary of significance for all Panicoideae co-orthologs.

https://www.dropbox.com/s/689mmnzu6k3d76v/Table_S6.xlsx?dl=0

Table II.S7: Pairwise differential expression tests for all Panicoideae co-orthologs.

https://www.dropbox.com/s/27fpwzpn7w63w48/Table_S7.xlsx?dl=0

Table II.S8: Summary of classes of differentially expressed genes based on their expression pattern across the phylogeny.

Class		#genes (all samples)	#genes (No *A. cimicina* & *A. angusta*)
Conserved		1935	3632
Shifted in the phylogeny		1484	1465
Remaining	Sig. 1 test	742	1308
	Sig. 2 tests	755	1215
	Sig. 3 tests	739	1013
	Sig. 4 tests	748	910
	Sig. 5 tests	687	865
	Sig. 6 tests	716	918
	Sig. 7 tests	693	558
	Sig. 8 tests	691	375
	Sig. 9 tests	556	301
	Sig. 10 tests	461	244
	Sig. 11 tests	446	194
	Sig. 12 tests	392	181
	Sig. 13 tests	385	94
	Sig. 14 tests	339	60
	Sig. 15 tests	295	70
	Sig. 16 tests	301	34
	Sig. 17 tests	238	22
	Sig. 18 tests	175	9
	Sig. 19 tests	148	4
	Sig. 20 tests	118	1
	Sig. 21 tests	107	2
	Sig. 22 tests	77	0
	Sig. 23 tests	68	0
	Sig. 24 tests	51	0
	Sig. 25 tests	38	0
	Sig. 26 tests	25	0
	Sig. 27 tests	18	0
	Sig. 28 tests	14	-
	Sig. 29 tests	14	-
	Sig. 30 tests	7	-
	Sig. 31 tests	4	-
	Sig. 32 tests	1	-
	Sig. 33 tests	6	-
	Sig. 34 tests	0	-
	Sig. 35 tests	1	-
	Sig. 36 tests	0	-
	Sig. 37 tests	0	-
	Sig. 38 tests	0	-
	Sig. 39 tests	0	-
	Sig. 40 tests	0	-
	Sig. 41 tests	0	-
	Sig. 42 tests	0	-
	Sig. 43 tests	0	-
	Sig. 44 tests	0	-
	Sig. 45 tests	0	-

Chapter III: Introgression and repeated co-option facilitated the recurrent emergence of C_4 photosynthesis among close relatives

Luke T. Dunning[*,1], Marjorie R. Lundgren[*,1], Jose J Moreno-Villena[*,1], Mary Namaganda[2], Erika J. Edwards[3], Patrik Nosil[2], Colin P. Osborne[2], Pascal-Antoine Christin[1,4]

[*] These authors contributed equally to the work.

[1] Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom.

[2] Makerere University, Kampala, Uganda.[3] Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912, USA.

[4] Corresponding author: p.christin@sheffield.ac.uk, +44-114-222-0027

This work was published in 2017 in **Evolution, Volume 71, Issue 6, Pages 1541–1555.**

Personal contributions: I generated the transcriptome data, which I analyzed jointly with Dr Luke Dunning. The anatomical data was generated and analyzed by Dr Marjorie Lundgren. I helped design the study and co-wrote the paper with Dr Dunning and Dr Lundgren, with the help of my co-authors.

Abstract

The origins of novel traits are often studied using species trees and modeling phenotypes as different states of the same character, an approach that cannot always distinguish multiple origins from fewer origins followed by reversals. We address this issue by studying the origin of C_4 photosynthesis, an adaptation to warm and dry conditions, in the grass *Alloteropsis*. We dissect the C_4 trait into its components, and show two independent origins of the C_4 phenotype via different anatomical modifications, and the use of distinct sets of genes. Further, inference of enzyme adaptation suggests that one of the two groups encompasses two transitions to a full C_4 state from a common ancestor with an intermediate phenotype that had some C_4 anatomical and biochemical components. Molecular dating of C_4 genes confirms the introgression of two key C_4 components between species, while the inheritance of all others matches the species tree. The number of origins consequently varies among C_4 components, a scenario that could not have been inferred from analyses of the species tree alone. Our results highlight the power of studying individual components of complex traits to reconstruct trajectories toward novel adaptations.

Keywords: ancestral state, complex trait, co-option, species tree, reticulate evolution

Introduction

Inferences of transitions among character states along species phylogenies provide powerful tools to test specific hypotheses about the timing and rate of functional diversification, correlations among functional and ecological traits (e.g., Pagel 1999; Edwards et al. 2010; Danforth et al. 2013; Moreau and Bell 2013; McGuire et al. 2014; Halliday et al. 2016), and speciation rates (Rabosky et al. 2013; Cantalapiedra et al. 2017; Cooney et al. 2017). However, distinguishing between a single origin of a trait with subsequent losses versus multiple independent origins can be problematic (Whiting et al. 2002; Pagel 2004; Wiens et al. 2006; Gamble et. al. 2012), particularly when some character states affect the rates of speciation and/or extinction, when rates of transitions are high and asymmetrical, or variable among clades and through time (Maddison 2006; Goldberg and Igic 2008; Beaulieu et al. 2013; Igic and Busch 2013; King and Lee 2015). Indeed, transition rates might be higher in taxonomic groups possessing evolutionary precursors that increase the likelihood of evolving a specific trait (Blount et al. 2008, 2012; Marazzi et al. 2012; Christin et al. 2013a, 2015; Werner et al. 2014). This can lead to an unbalanced distribution of character states across the tree, with clusters forming in certain clades. However, a low rate of origins would lead to similar patterns if the rate of reversals is high (Wiens 1999; Danforth et al. 2003; Trueman et al. 2004; Pyron and Burbink 2014). Difficulties worsen if hybridization and introgression disconnect the history of underlying traits from the species tree (Pardo-Diaz et al. 2012; Meier et al. 2017).

An alternative approach to analysing the phenotypes as different character states is to decompose them into their constituent parts, then carefully analyse the evolution of each element independently to understand how the trait has been assembled or lost (Christin et al. 2010; Oliver et al. 2012; Niemiller et al. 2013; Kadereit et al. 2014). The use of distinct components can be interpreted as evidence for multiple origins, while reversals could leave a signature of the lost trait that can be detected when components are compared with those from species that never evolved it (Protas et al. 2006; Christin et al. 2010; Oliver et al. 2012; Niemiller et al. 2013). Identifying the mutations that underlie a trait further helps to distinguish shared origins and reversals (Igic et al. 2006; Shimizu et al. 2008; Niemiller et al. 2013; Meier et al. 2017). Evaluating the number of origins of each component of a complex trait would reconstruct the order of

modifications that led to the trait of interest. This approach is applied here to the photosynthetic diversity exhibited within a five-species taxonomic group.

C_4 photosynthesis is a complex phenotype that improves the efficiency of carbon fixation in warm and dry conditions when compared to the ancestral C_3 photosynthetic pathway (Sage et al. 2012; Atkinson et al. 2016). The C_4 advantages are achieved by increasing the concentration of CO_2 around Rubisco, the enzyme responsible for inorganic carbon fixation in the Calvin cycle of all photosynthetic organisms (von Caemmerer and Furbank 2003; Sage et al. 2012). To function, C_4 photosynthesis requires the co-ordinated action of numerous anatomical and biochemical components that lead to the emergence of a novel biochemical pathway, usually across two types of cells; the mesophyll and bundle sheath cells (Hatch 1987; Prendergast et al. 1987; Gowik et al. 2011; GPWGII 2012; Bräutigam et al. 2014). Besides the increased expression of genes co-opted for a C_4 function, several other changes are known to occur during the evolution of C_4 photosynthesis, including: an expansion of bundle sheath tissue, a concentration of chloroplasts within it, and the adaptation of the enzymes to the new catalytic context (Fig. III.1; Bläsing et al. 2000; von Caemmerer and Furbank 2003; McKown and Dengler 2007; Sage et al. 2012).

Despite its apparent complexity, C_4 photosynthesis evolved multiple times independently, and is present in distantly related groups of plants (Sinha and Kellogg 1996; Kellogg 1999; Sage et al. 2011). As with any complex trait, C_4 photosynthesis likely evolved in incremental steps, via stages that are functionally intermediate and gradually increase carbon assimilation in warm and dry conditions (Fig. III.1; Sage et al. 2012; Heckmann et al. 2013; Williams et al. 2013; Mallmann et al. 2014; Christin and Osborne 2014). An increase in bundle sheath size and the relocation of the chloroplasts/Rubisco to these cells can sustain a photorespiratory bypass (Hylton et al. 1988; Bräutigam and Gowik 2016). Subsequent increases in C_4 enzyme abundances can generate a weak C_4 cycle, which assimilates some of the atmospheric CO_2, complementing the C_3 cycle in C_3+C_4 plants (referred to as 'type II C_3-C_4 intermediates' in the specialized literature; Fig. III.1; Heckmann et al. 2013; Mallmann et al. 2014). The transition to a full C_4 state involves further increases of the bundle sheath tissue and gene expression, while selective pressures adapt the C_4 enzymes for the new biochemical context (Fig. III.1; Bläsing et al. 2000; McKown and Dengler 2007).

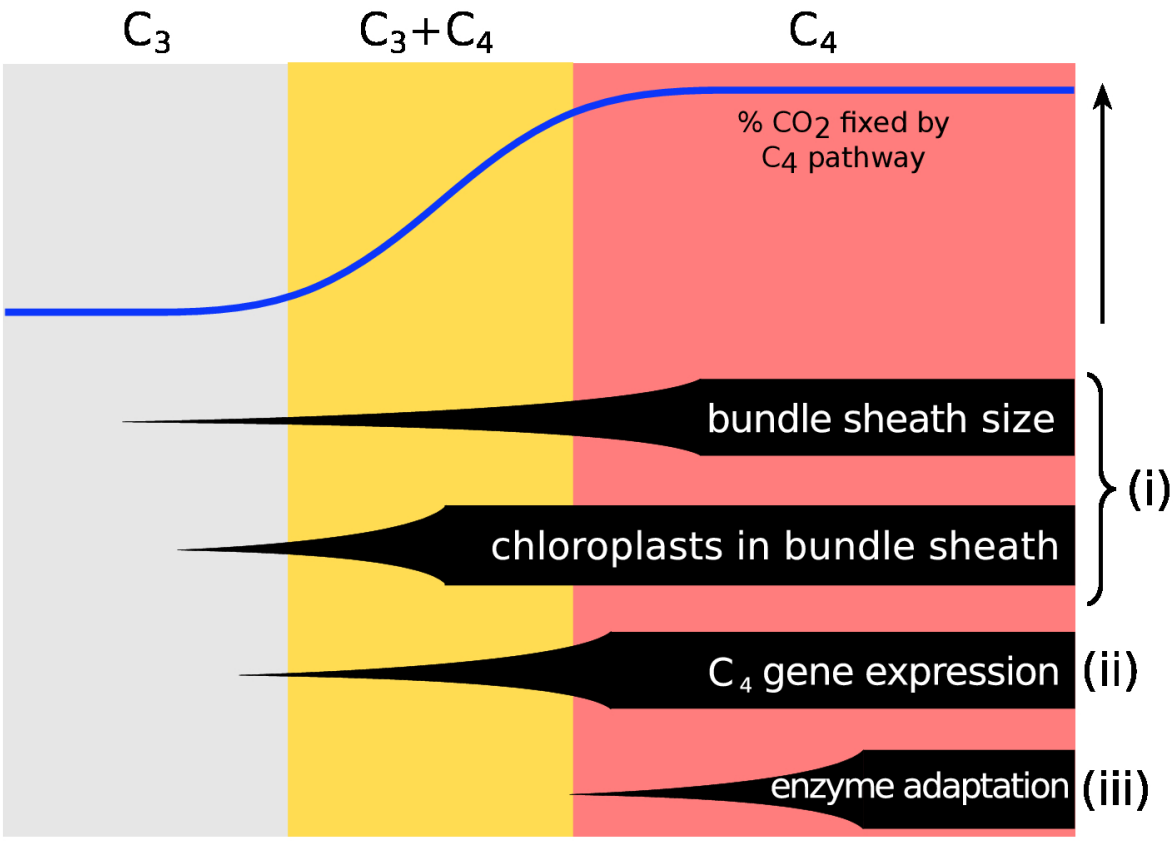


Figure III.1. Schematic of expected changes during the transition from C_3 to C_4.
The continuous variation in anatomical and biochemical components can be simplified using three phenotypic categories; C_3 plants, C_3+C_4 intermediates, and C_4 plants. A schematic indicates the gain in efficiency of carbon assimilation, and the expected order of modifications is shown at the bottom for four categories of changes, with the number on the right indicating the section of our analyses where they are investigated.

In the angiosperm phylogeny, C_4 taxa form clusters, many of which have multiple C_4 clades that are separated by non-C_4 branches (Sage et al. 2011; GPWGII 2012). Thus, establishing past photosynthetic transitions is difficult when photosynthetic type is modeled as a simple binary character (Ibrahim et al. 2009; Christin et al. 2010; Hancock and Edwards 2014; Bohley et al. 2015; Fisher et al. 2015; Washburn et al. 2015). Overall, non-homology of key C_4 components among some closely related C_4 groups, including the cells, enzymes, and genes modified to generate the C_4 pathway (Prendergast et al. 1987; Soros and Dengler 2001; Bräutigam et al. 2014; Lundgren et al. 2014; Wang et al. 2014), points to a predominance of C_4 origins (Sinha and Kellogg 1996; Christin and Besnard 2009; Christin et al. 2010). However, the possibility of evolutionary reversals to a non-C_4 state is still debated (e.g. Kadereit et al. 2014; Bohley

et al. 2015; Washburn et al. 2015). Furthermore, some components of the C_4 phenotype (e.g. expansion of bundle sheaths and migration of chloroplasts; Fig. III.1) may have evolved relatively few times, and have then been recurrently utilized for independent transitions to C_3+C_4 or C_4 photosynthesis (Christin et al. 2011, 2013a).

One of the proposed candidates for an evolutionary reversal from C_4 to C_3 is in the grass genus *Alloteropsis* (Ibrahim et al. 2009). Within this genus, the species *Alloteropsis semialata* contains C_3, C_3+C_4, and C_4 genotypes (Ellis 1974; Brown 1975; Lundgren et al. 2016). In molecular phylogenies based on either plastid or nuclear markers, this species is sister to the C_4 *A. angusta*, and the two species form a monophyletic clade sister to the three remaining closely-related C_4 species: *A. cimicina*, *A. paniculata*, and *A. papillosa* (Ibrahim et al. 2009; Christin et al. 2012; Olofsson et al. 2016). The C_4 *A. semialata* and *A. cimicina* use different cell types for the segregation of C_4 reactions (Renvoize 1987), which suggests independent realizations of C_4 photosynthesis (Christin et al. 2010). However, the evolutionary origins of C_4 biochemistry and the situation within the *A. angusta* / *A. semialata* group remain largely unexplored.

In this study, we focus on the genus *Alloteropsis* and its C_3 outgroup, to test the competing hypotheses of multiple origins versus fewer origins followed by reversals, independently for each C_4 component. A C_4 phenotype generated via distinct cells, genes, or amino acid mutations would indicate independent origins. In contrast, a reversal may lead to a derived state that retains traces of its past C_4 state when compared to the ancestral one (i.e. approximated by the C_3 outgroup here). We combined different approaches to investigate different components of the complex C_4 trait. (i) Focusing on anatomical characters, we evaluate the most likely number of episodes of movement of chloroplasts to the bundle sheath, and expansion of this tissue. (ii) Using transcriptome analyses to estimate gene expression, we then determine the most likely number of origins of a C_4 cycle via the upregulation of known C_4 photosynthetic genes. (iii) The number of episodes of enzyme adaptation for the C_4 cycle is estimated using positive selection analyses, with scenarios corresponding to episodes of adaptation along different sets of branches. (iv) Finally, we compare divergence times across genes, to detect potential introgression of C_4 components, as suggested within this genus for two C_4 genes (Olofsson et al. 2016). Our multifaceted effort highlights the power of

comparative analyses that directly consider genes and other components involved in the trait of interest, rather than modeling complex phenotypes as states of a single character. Using this approach, we show that recurrent origins of C_4 photosynthesis in *Alloteropsi*s arose via a complex mixture of co-option of traits increasing C_4 accessibility, hybridization, and independent adaptation of the phenotype.

Methods

Taxon Sampling

The different datasets were obtained from plants grown under controlled conditions (See Supplementary Methods 1.1 for detailed description of growth conditions), including one *Alloteropsis cimicina* (C_4) accession, one *Alloteropsis paniculata* (C_4) accession, two *Alloteropsis angusta* (C_4) accessions, and up to ten different *Alloteropsis semialata* accessions collected from separate populations that encompass the global genetic and photosynthetic diversity of this species (one C_3, two C_3+C_4 intermediates with a weak C_4 cycle, and seven C_4 accessions; Table III.S1; Lundgren et al. 2016). The over representation of C_4 *A. semialata* accessions mirrors their natural abundance, with C_4 accessions spread throughout Africa, Asia and Australia, C_3 accessions only reported in Southern Africa, and C_3+C_4 individuals restricted to central East Africa (Lundgren et al. 2015). We also make use of species representing the C_3 sister group to *Alloteropsis* (*Panicum pygmaeum* and *Entolasia marginata*), previously identified using plastid markers (GPWGII 2012). Using the above taxa, we conduct four complementary sets of analyses, each providing insight into the origins or spread of distinct components of C_4 in *Alloteropsis*.

i) Comparing leaf anatomies among photosynthetic types

Leaf cross sections were analyzed to identify the leaf compartment being used for the segregation of Rubisco and the modifications that increased the proportion of bundle sheath tissue in C_3+C_4 and C_4 accessions. Co-option of different tissues and distinct modifications among accessions would support independent origins, while a reversal should result in the leaves of C_3 individuals having reverted to a state that retain traces of their past C_4 state when compared to the ancestral condition (e.g. enlarged bundle sheath cells and/or chloroplasts in the bundle sheath).

We generated new anatomical data for nine *A. semialata* accessions and *A. angusta* (Table III.S2), which supplemented previously published anatomical data for *E. marginata*, *P. pygmaeum*, *A. cimicina*, and *A. paniculata* (Christin et al. 2013a). Images of *A. semialata* and *A. angusta* leaves in cross-section were obtained by fixing the centre portion of a mature leaf blade in 4:1 ethanol:acetic acid, embedding them in methacrylate embedding resin (Technovit 7100, Heraeus Kulzer GmbH, Wehrhein, Germany), sectioning on a manual rotary microtome (Leica Biosystems, Newcastle, UK), staining with Toluidine Blue O (Sigma-Aldrich, St. Louis, MO, USA), then photographing them with a camera mounted atop a microscope (Olympus DP71 and BX51, respectively. Olympus, Hamburg, Germany), as described in Lundgren et al. (2016).

All species used in this study have two bundle sheath layers, differentiated as inner and outer bundle sheaths, which create concentric circles around each vein (Fig. III.S1). The sheath co-opted for the segregation of Rubisco was identified by a concentration of chloroplasts producing starch. We also recorded the presence of minor veins, and measured the following traits on one cross-sectional image per accession, as described in Christin et al. (2013a), using ImageJ software (Schneider et al. 2012): the interveinal distance (IVD; the average distance between centers of consecutive veins), the number of mediolateral mesophyll cells between veins, the average width of all outer and inner bundle sheath cells within a leaf segment, and the ratio of outer to inner bundle sheath cell widths (OS:IS). One leaf cross section was used per accession, with previous work showing the traits we are measuring exhibit little variation within species or populations (Lundgren et al. 2016).

ii) Comparing gene expression profiles among photosynthetic types

We use RNA-Seq to identify the genes co-opted by the different accessions performing a C_4 cycle, as those encoding C_4-related enzymes that reach high abundance in C_4 leaves. Variation in the co-opted loci would support multiple origins of a weak C_4 cycle, while a reversal might lead to high expression of C_4-related genes in individuals without a C_4 cycle or loss of functions of genes previously used for the C_4 cycle.

For RNA-Seq, we sampled the highly photosynthetically active distal halves of fully expanded new leaves and fresh roots midway into the photoperiod, which were

subsequently flash frozen. Two different photoperiods (i.e., 10 and 14 hours) were used to ensure that the identification of the most highly expressed genes did not differ among light regimes. Data from root libraries were only used in this study for transcriptome assembly, while all leaf samples were used for both assembly and quantification of transcript abundances. For a full list of individuals, conditions, and tissues sampled please see Supplementary Table III.S3.

Total RNA was extracted, Illumina TruSeq libraries generated, and sequencing performed using standard laboratory procedures, and transcriptomes were assembled using available pipelines (see Supplementary Methods 1.2 for a detailed description of RNA-Seq protocol and assembly statistics). For each assembled contig, the transcript abundance was calculated as reads per million of mapped reads (rpm). Using a previously developed phylogenetic annotation pipeline (Christin et al. 2013b, 2015), the transcript abundance was then calculated for each gene lineage encoding C_4-related enzymes. For each gene family, all sequences descending from a single gene in the common ancestor of grasses via speciation and/or duplication were considered as the same gene lineage (i.e., these are grass co-orthologs). These groups include potential lineage specific paralogs (i.e., also known as inparalogs). When different *Alloteropsis* genes were identified within the same group of co-orthologs through detailed phylogenetic analyses, the abundance of each group was estimated independently. In *Alloteropsis*, this is the case only for genes previously shown to have been acquired laterally from distantly related C_4 lineages (Christin et al. 2012; see Results). In short, the reference datasets, composed of *Arabidopsis thaliana* coding sequences annotated as encoding C_4-related enzymes, and homolog sequences from other completely sequenced plants including five grasses, were retrieved from Christin et al. (2013b; 2015), or generated following the same approach for additional C_4-related enzymes identified in more recent studies (Mallmann et al. 2014; Li et al. 2015; Fig. III.S2). Contigs with similar sequences from the transcriptomes generated here were identified using BLASTn, with a minimal e-value of 0.01, and a minimal matching length of 50bp. Only the portion of the contig matching the references was considered to remove UTRs, potential introns, and other very variable segments. Each sequence retrieved this way was then aligned independently to the reference dataset using Muscle (Edgar 2004), and a phylogenetic tree was inferred using Phyml (Guindon and Gascuel 2003) with a

GTR+G+I model, a model that fits the vast majority of genes (e.g., Fisher et al. 2016) and is appropriate to infer a large number of trees. Phylogenetic trees were automatically screened, and each contig was assigned to the previously identified gene lineages in which it was nested. The sum of rpm values of all transcriptome contigs assigned to the same gene lineage produced transcript abundance per group of grass co-orthologs or distinct genes within these groups, which were subsequently transformed into rpm per kilobase (rpkm) values. Rpkm values were then compared among accessions to identify similarities and differences in the expression of C_4 photosynthetic genes.

iii) Gene trees and detection of enzyme adaptation for C_4 photosynthesis

Phylogenetic trees were inferred for C_4-related genes that were highly abundant in the leaf transcriptomes of at least two C_4 *Alloteropsis* samples (identified from transcriptome data; see Results) and their co-orthologs in other C_3 and C_4 grasses (see Supplementary Methods 1.3 for a detailed description of phylogeny construction). The inferred gene trees were used to verify that C_4-related genes were placed as expected based on the species tree, as opposed to a position suggesting an acquisition from distant C_4 relatives. In addition, the gene tree topologies were used for positive selection analyses to detect traces of past episodes of enzyme adaptation for the new catalytic context after the initial emergence of a C_4 cycle (Fig. III.1; Blasing et al. 2000; Christin et al. 2007; Besnard et al. 2009; Wang et al. 2009; Heckmann et al. 2013; Mallmann et al. 2014; Huang et al. 2017). Positive selection on branches leading to each C_4 group would support independent transitions to a full C_4 cycle via enzyme adaptation, while an early origin followed by a reversal should result in positive selection in the common ancestor of all C_4 accessions and possibly in the lineages that reversed back to the previous state.

For each set of genes encoding core C_4 enzymes in at least two *Alloteropsis* accessions, identified via transcriptome analyses, we optimized several codon models (site and branch-site models) to test for adaptive evolution using codeml as implemented in PAML (Yang 2007). The best-fit model was identified among those that assume (0) no positive selection (M1a null model), and the branch-site models that assume shifts in selection pressure, either to relaxed selection (model BSA) or to

positive section (model BSA1), at the base of: (1) *Alloteropsis* (one round of enzyme adaptation), (2) both *A. cimicina* and *A. angusta* + *A. semialata* (two rounds of enzyme adaptation), and (3) *A. cimicina*, *A. angusta*, and *A. semialata* (three independent episodes of enzyme adaptation). Foreground branches for all models were specified as the branch leading to the identified node plus all descending branches (i.e., using a '$' sign as opposed to a '#'). Models involving positive selection in only one of the C_4 lineages were also considered (see Supplementary Methods 1.3 for additional details of positive selection analysis). For each gene lineage, the best-fit model was identified based on the corrected Akaike information criterion (AICc), selecting the model with the lowest AICc after checking that its ΔAICc score was at least 5.22 units below that of the M1a null model. An ΔAICc score = 5.22 corresponds to a p-value threshold of 0.01 for a likelihood ratio test comparing these two models using two degrees of freedom. C_4 species other than *Alloteropsis* were removed prior to analysis to avoid an influence of positive selection in these taxa affecting our conclusion. Analyses were repeated using only codons with fixed nucleotides within each lineage (i.e., *A. angusta*, C_3 *A. semialata*, C_3+C_4 *A. semialata*, and C_4 *A. semialat*a), to verify that short terminal branches with unfixed mutations did not significantly inflate the dN/dS ratio, and therefore alter our conclusion. Finally, to assess the effect of gene tree topology on our conclusions, we repeated the positive selection analyses using 100 bootstrap pseudoreplicate topologies.

iv) Dating the divergence of adaptive loci to identify introgression

To determine whether introgression has spread C_4 adaptations among species, we performed molecular dating of markers from across the transcriptomes, including those used for C_4 by at least two *Alloteropsis* accession and their paralogs. The divergence times between species estimated from introgressed genes are expected to be younger than those estimated from other genes (e.g. Smith & Kronforst 2013; Li et al. 2014; Marcussen et al. 2014; Li et al. 2016), resulting either in outliers (if few genes are introgressed) or a multimodal distribution of ages (if many genes are introgressed).

Groups of genes descending from a single gene in the common ancestor of Panicoideae (Panicoideae co-orthologs), the grass subfamily that includes *Alloteropsis*, were identified through phylogenetic analyses of our transcriptomes and completely

sequenced genomes that were publicly available. Our automated pipeline started with gene families previously inferred for eight plant genomes (homologs: i.e. all the paralogs and orthologs; Vilella et al. 2009), including two Panicoideae grasses (*Setaria italica* and *Sorghum bicolor*), two non-Panicoideae grasses (*Brachypodium distachyon* and *Oryza sativa*), and four non-grass species (*Amborella trichopoda*, *Arabidopsis thaliana*, *Populus trichocarpa*, and *Selaginella moellendorffii*). To ensure accurate annotation, we restricted the analysis to gene families that included at least one *Arabidopsis thaliana* sequence. The coding sequences (CDS) from the above genomes were then used to identify similar sequences in our transcriptomes using BLASTn with a minimum alignment length of 500 bp.

Stringent alignment and filtering methods were used to ensure reliable alignments of the above sequences for each gene family for phylogenetic inference (see Supplementary Methods 1.4 for full details). In total, 2,797 1:1 Panicoideae co-ortholog datasets were used for subsequent molecular dating, as implemented in Beast v. 1.5.4 (Drummond and Rambaut 2007). For each dataset, divergence times were estimated based on 3rd codon positions, to decrease the risk of selective pressures biasing the outputs. A log-normal relaxed clock was used, with a GTR+G+I substitution model, and a constant coalescent prior. The *Sorghum* sequence was selected as the outgroup and the root of the tree was fixed to 31 Ma (using a normal distribution with a standard deviation of 0.0001), based on estimates from Christin et al. (2014). There is uncertainty around this date, and the low species sampling used here probably leads to overestimation of both divergence times and confidence intervals, but the use of consistent sampling and calibration points among markers allows for the comparison of relative (rather than absolute) ages, which is the point of these analyses. Each Beast analysis was run for 2,000,000 generations, sampling a tree every 1,000 generations after a burn-in period of 1,000,000. For nodes of interest, divergence times were extracted from the posterior distribution as medians.

Divergence times were also estimated for key genes used for C_4 photosynthesis in *Alloteropsis* (identified based on transcriptomes; see Results), using the same parameters. To guarantee a consistent species sampling, the taxa included in the transcriptome-wide analyses were retrieved from manually curated alignments for C_4-specific genes as well as other groups of orthologs from the same gene families,

obtained as described above for C_4-specific forms. In addition, plastid genomes for the same species were retrieved from Lundgren et al. (2015), and reanalysed with the same parameters. For each of these datasets, the median, 95% CI, and 0.25 and 0.75 quantiles were extracted from the posterior distribution, using the R package APE (Paradis et al. 2004).

Results

i) Different realizations of C_4 leaf anatomy in A. cimicina *and* A. semialata/A. angusta

Grasses ancestrally possess two concentric rings of bundle sheath cells and either can be co-opted for C_4 photosynthesis (Brown 1975; Lundgren et al. 2014). The closely related C_4 *A. cimicina* and *A. paniculata* co-opted the outer bundle sheath for Rubisco segregation, as evidenced by the proliferation of chloroplasts in this tissue (Fig. III.S1; Table III.S2). In these species, the overall proportion of outer bundle sheath tissue within the leaf is increased via enlarged outer bundle sheath cells. Indeed, the outer sheath is 7.8-fold larger than the inner sheath in C_4 *A. cimicina* and *A. paniculata*, compared to a 1.2-0.6 fold differences in C_4 *A. semialata* and *A. angusta* (Table III.S2). This contrasts strongly with the anatomy of the C_4 *A. semialata* and *A. angusta* (Fig. III.S1). Both of these species use the inner bundle sheath for Rubisco segregation and increase the overall proportion of this tissue via the proliferation of minor veins, and enlargement of the inner sheath cell size (Fig. III.S1; Table III.S2).

Staining by Toluidine Blue O indicates some starch production occurs in the inner bundle sheaths of both the C_3 and C_3+C_4 *A. semialata* (Fig. III.S1), which implies some Rubisco activity in these cells, confirming previous reports (Ueno and Sentoku 2006; Lundgren et al. 2016). The absence of minor veins in the C_3 and C_3+C_4 *A. semialata* results in a larger proportion of mesophyll compared to C_4 *A. semialata* (Table III.S2; Fig. III.S1). In the C_3 and C_3+C_4 *A. semialata*, the outer bundle sheath is slightly larger than the inner one (1.2-1.8 fold; Table III.S2), while the C_3 outgroup species *P. pygmaeum* and *E. marginata* have outer bundle sheaths that are considerably larger than their small inner sheaths (4.5 and 5.3 fold; Fig. III.S1; Table III.S2).

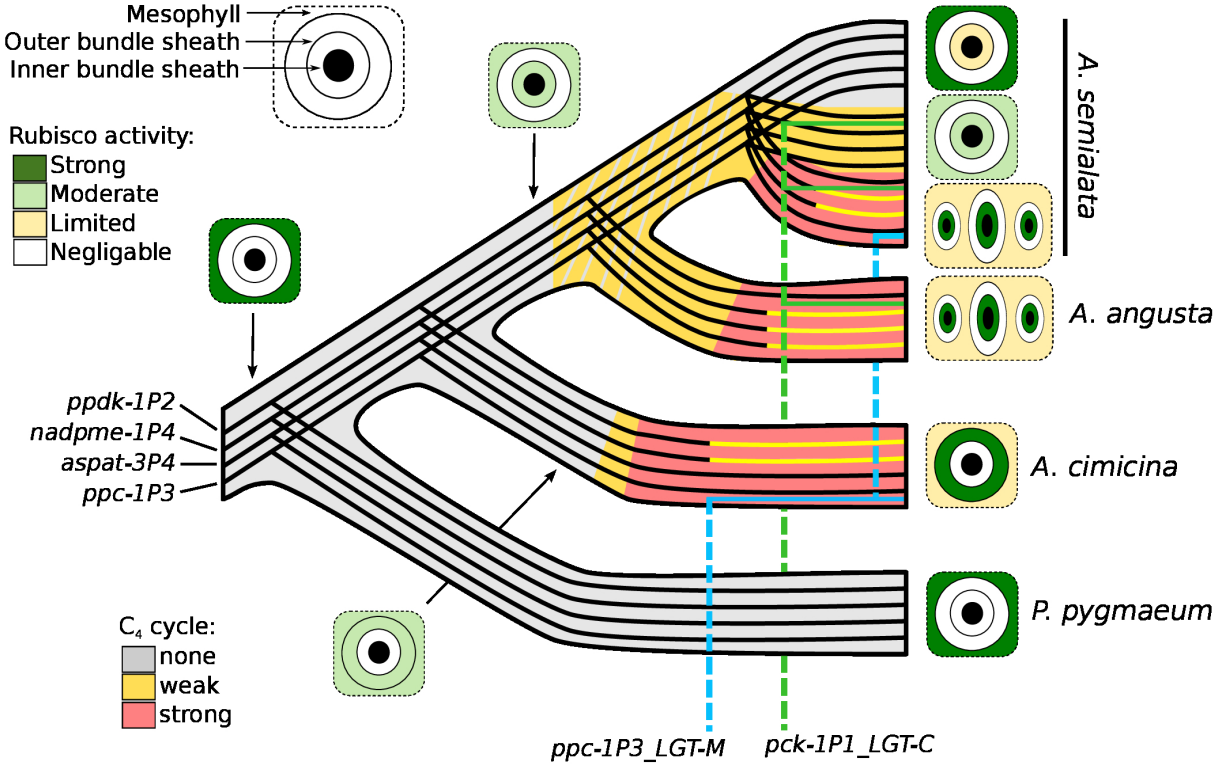


Figure III.2: Inferred transitions among C_4 components.
A schematic phylogenetic tree is presented, based on previous genome-wide analyses (Lundgren et al. 2015; Olofsson et al. 2016). Individual lines represent the transmission of individual genes within the species complex. For each of the four genes subject to C_4-related selection, episodes of positive selection are indicated by changes to yellow. Other lines track the spread of two genes that were originally laterally-acquired from distant relatives, and have subsequent been introgressed among *Alloteropsis* species. The inferred phenotype is represented by the background colour, in grey for C_3, in yellow for C_3+C_4, and in red for C_4. The grey hatching indicates uncertainty about the ancestral state. A simplified version of leaf anatomy is represented, for extant taxa and some hypothetical ancestors (see Fig. III.S1 for details of leaf anatomy of extant accessions).

In summary, our comparative studies of leaf anatomy indicate that the C_4 *A. cimicina* and *A. semialata*/*A. angusta* use different tissues for Rubisco segregation and achieve high bundle sheath proportions via distinct modifications, supporting independent origins of C_4 anatomical components in these two groups. Some Rubisco activity is suggested in the inner sheath of the C_3 *A. semialata*, which supports an early origin migration of chloroplasts to this tissue (Fig. III.2). In addition, a slight enlargement of the inner sheath, absent in the C_3 outgroup, is common to all non-C_4 *A. semialata*.

ii) A. cimicina *uses different enzymes and genes for C_4 biochemistry than* A. semialata/A. angusta

All *Alloteropsis* C_3+C_4 and C_4 accessions have high expression abundance in their leaves of co-orthologs encoding phosphoenolpyruvate carboxylase (PEPC), the enzyme used for the initial fixation of atmospheric carbon into organic compounds in C_4 plants. However, the gene lineage most highly expressed varies among accessions (Fig. III.3 and III.4). The close relationships between some of the genes for PEPC and one for phosphoenolpyruvate carboxykinase (PCK) isolated from *Alloteropsis* and those of distantly-related C_4 species was confirmed by our phylogenetic analyses (Fig. III.S3 and III.S4), supporting the previous conclusion that these genes were acquired by *Alloteropsis* via lateral gene transfer (LGT; Christin et al. 2012). Based on the read abundance, *A. cimicina* uses *ppc-1P3_LGT-M*, while *A. angusta* uses *ppc-1P3* (Fig. III.4). There is variation within *A. semialata*, with C_3+C_4 and C_4 accessions using either a combination of one or several gene lineages (Fig. III.4).

From the expression profiles (Fig. III.3), the carbon shuttle of *A. cimicina* relies on enzymes and transporters associated with the most common form of C_4 photosynthesis (NADP-malic enzyme type; Gowik et al. 2011; Bräutigam et al. 2014; Mallman et al. 2014). This expression profile differs markedly from that observed in the C_4 *A. semialata* and *A. angusta* accessions. These two species mainly use the PCK decarboxylating enzyme, through the high expression of the same gene (*pck-1P1_LGT-C*; Fig. III.4). There is little evidence in these species for an involvement of the auxiliary transporters observed in *A. cimicina* (Fig. III.3; Table III.S4), and some of the core enzymes are not shared by *A. cimicina* and *A. semialata*/*A. angusta* (Fig. III.3). Furthermore, even when the same enzyme family is used, it is not necessarily encoded by the same locus (e.g., *A. cimicina* expresses *aspat-2P3* and *A. semialata*/*A. angusta* express *aspat-3P4*; Fig. III.3).

The transcriptomes of the C_3+C_4 *A. semialata* show elevated levels of some of the genes used by the C_4 *A. semialata*, with a slightly higher abundance of those encoding the NADP-malic enzyme (*nadpme-1P4*; Fig. III.3; Table III.S4). In terms of the expression levels of genes encoding C_4-related enzymes, the transcriptome of the C_3 *A. semialata* is not markedly different from that of the C_3 outgroup *P. pygmaeum* (Fig. III.3; Table III.S4).

Our comparative transcriptomics therefore indicate that *A. cimicina* uses different genes and different enzymes for the C_4 pathway than *A. semialata* / *A. angusta*, suggesting multiple origins of the C_4 cycle (Fig. III.2). The only C_4-related genes used by some C_4 *Alloteropsis* that are abundant in the C_3 *A. semialata* (*bca-2P3* and *tpt-1P1*) are also highly expressed in the C_3 outgroup and in other distantly related C_3 taxa (Fig. III.3; Külahoglu et al. 2014; Ding et al. 2015), indicating that high levels in leaves is not specific to our group of species. For the C_4-related used by the C_4 *Alloteropsis*, but not abundant in the outgroup, there is no evidence for high expression or pseudogeneization in the C_3 *A. semialata*. Evidence is thus lacking that the C_3 *A. semialata* represent a reversal from an ancestor with a C_4 cycle.

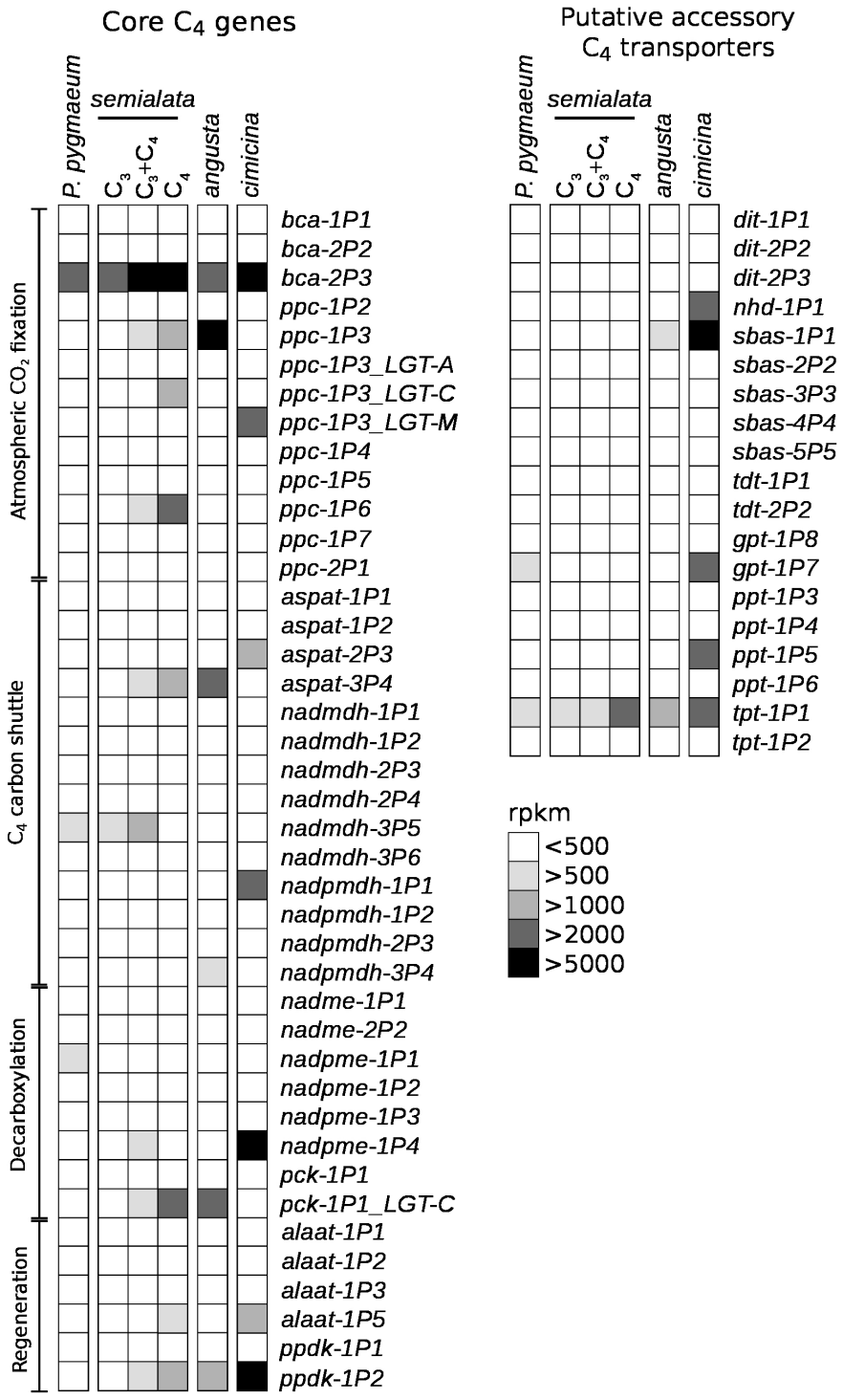


Figure III.3: Expression of C_4-related enzymes in *Alloteropsis.*
For each gene encoding a C_4-related enzyme, the shade indicates the category of transcript abundance, using averages per group. For raw values, see Table III.S4. Note that *ppc* abundance varies among C_4 accessions of *A. semialata* (Fig. III.4). The enzymes involved in core C_4 reactions (left column) are grouped by functional property, gene names are written in italics on the right of the expression values.

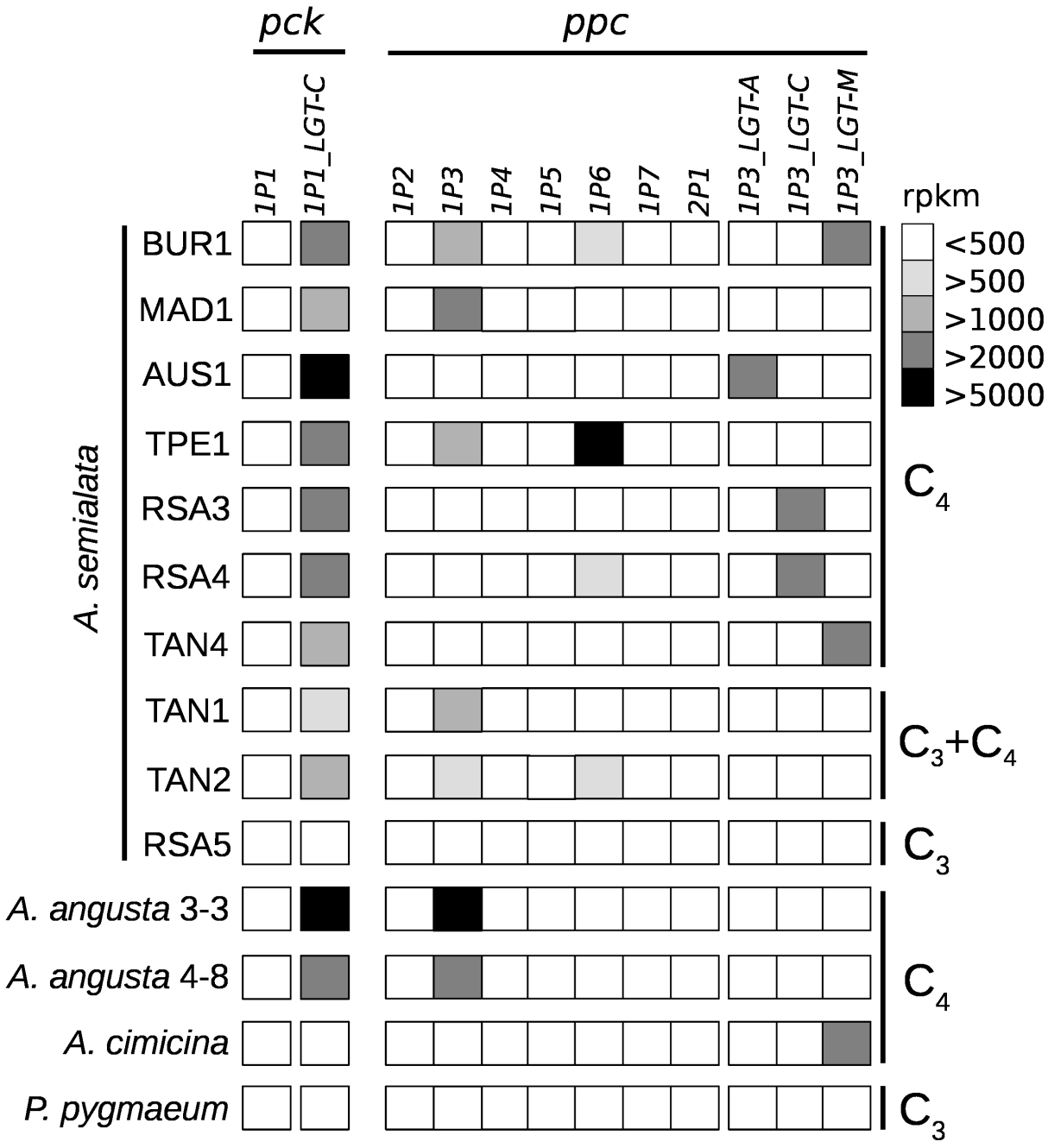


Figure III.4: Leaf abundance of *pck* and *ppc* genes in the different accessions.
The shade indicates the relative expression (in rpkm) in the different accessions. For each accession, the averages are used. For raw values, see Table III.S4.

iii) Independent episodes of C_4-related positive selection in each C_4 species

The codon models do not support positive selection on any genes involved in C_4 photosynthesis at the base of *Alloteropsis* or along the branch leading to the *A. angusta*/*A. semialata* group (Table III.1). In two cases (*nadpme-1P4* and *ppdk-1P2*), analyses including all *Alloteropsis* accessions clearly point to changes in selective pressures specifically in the branch leading to *A. cimicina* (Table III.1; Fig. III.S5). No evidence of positive selection was found for the two other genes analyzed on the three *Alloteropsis* species (*aspat-2P3* and *alaat-1P5*; Table III.1). When testing for selection only in the *A. angusta*/*A. semialata* clade, no positive selection was found on *ppdk-1P2*, while positive selection on *ppc-1P3* was identified only on the branch leading to *A. angusta* (Table III.2). For the two other genes (*nadpme-1P4* and *aspat-3P4*), the model that assumes positive selection after the split of the two species was favored (Table III.2). A majority of the amino acid sites identified as under positive selection by the Bayes Empirical Bayes analysis overlapped with those previously identified in other C_4 taxa (e.g. site 241 in *nadpme-1P4;* Fig. III.5; Christin et al. 2009), or were shared with other C_4 species in our phylogenies (e.g. Fig. III.5), supporting their link to C_4 photosynthesis. For *aspat-3P4*, more amino acid substitutions were fixed in *A. angusta* than in *A. semialata*. This variation among *A. semialata* C_4 accessions indicates repeated bouts of positive selection during the diversification of this species (Fig. III.S6). Conclusions based on the selection tests were also supported using only codons with fixed nucleotides within a lineage (i.e., photosynthetic type in *A. semialata*, and *A. angusta*), with the exception of *nadpme-1P4* for which no positive selection was inferred after removing the unfixed codons (Tables S5, S6). Furthermore, gene tree topology had no effect on our conclusions, since all bootstrap replicates supported the same model, with the exception of 2% of *nadpme-1P4* bootstrap replicates (Tables S7, S8).

Overall, our positive selection tests point to independent episodes of enzyme adaptation for the C_4 context in each of the C_4 groups (Fig. III.2). None of the models that included adaptive evolution on branches leading to C_3 and/or C_3+C_4 *A. semialata* were favoured, suggesting a lack of evolutionary loss of a full C_4 cycle.

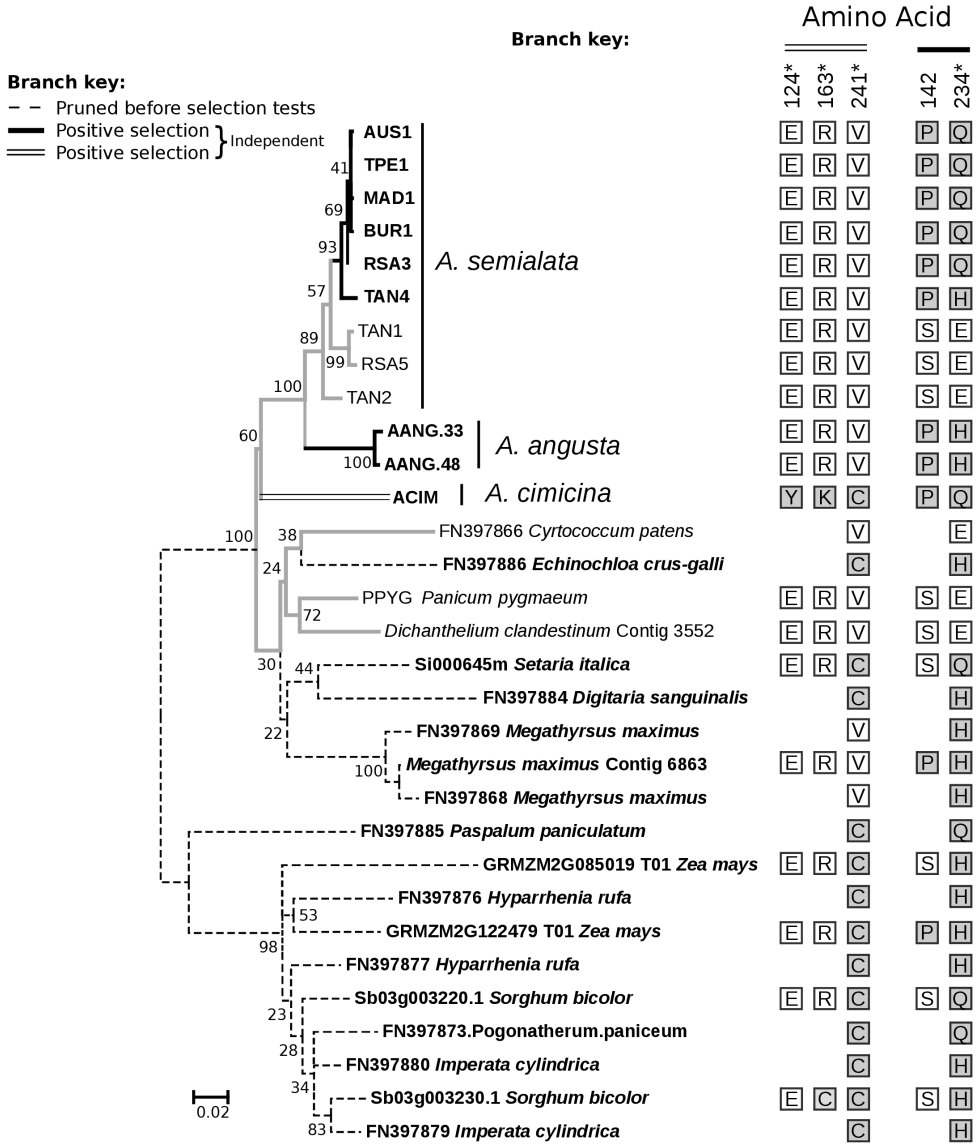


Figure III.5: Evolution of *nadpme-1P4* genes in *Alloteropsis* and other Panicoideae.
This phylogenetic tree was inferred on 3rd positions of codons of *nadpme-1P4* genes of Panicoideae. Bootstrap values are indicated near branches. Names of C_4 accessions are in bold. Amino acid at positions under positive selection are indicated on the right, with those associated with C_4 accessions in gray. Positions are indicated on the top, based on Sorghum gene Sb03g003220.1. Amino acid positions with a posterior probability >0.90 of being under positive selection are indicated on the right, asterisks indicate positions with a posterior probability >0.95.

Table III.1: Results of positive selection analyses inferring the episodes of enzymatic adaptation in *Alloteropsis*[1].

Gene	Number of sequences	Site model M1a	One origin		Two origins		Three origins		Only *A. cimicina*	
			BSA	BSA1	BSA	BSA1	BSA	BSA1	BSA	BSA1
aspat-2P3	14	**0.00***	4.02	4.02	4.02	4.02	3.94	4.02	4.00	4.00
nadpme-1P4	15	35.07	30.44	27.26	26.34	24.28	19.52	13.31	3.34	**0.00***
ppdk-1P2	15	29.45	32.00	26.35	32.31	27.17	26.37	23.37	3.55	**0.00***
alaat-1P5	14	**0.00***	1.74	1.74	2.03	2.03	0.69	0.69	4.02	4.02

[1] The ΔAICc values compared to the best fit model for that gene are shown. The most appropriate model is indicated with an asterisk, with the null model (M1a) only rejected if the ΔAICc was at least 5.22 (equivalent to a p-value of 0.01 with a likelihood ratio test with df = 2). Two branch-site models were used to test for a relaxation of purifying selection (BSA), and potential positive selection (BSA1).

Table III.2: Results of positive selection analyses inferring the episodes of enzymatic adaption in the *A. angusta/A. semialata* clade[1].

Gene	Number of sequences	Site model M1a	One origin		Two origins		Only *A. angusta*	
			BSA	BSA1	BSA	BSA1	BSA	BSA1
aspat-3P4	13	12.33	10.20	6.70	6.37	**0.00***	5.45	5.29
nadpme-1P4	14	10.19	14.19	9.66	13.52	**0.00***	14.18	14.18
ppc-1P3	9	72.43	66.62	66.58	11.70	9.85	5.66	**0.00***
ppdk-1P2	14	**0.00***	4.01	4.01	4.01	4.01	3.91	3.91

[1] The ΔAICc values compared to the best fit model for that gene are shown. The most appropriate model is indicated with an asterisk, with the null model (M1a) only rejected if the ΔAICc was at least 5.22 (equivalent to a p-value of 0.01, with a likelihood ratio test with df = 2). Two branch-site models were used to test for a relaxation of purifying selection (BSA), and potential positive selection (BSA1).

iv) Genes for PEPC and PCK were spread across species boundaries

The 2,797 groups of orthologs extracted from genomes and transcriptomes led to a wide range of estimated divergence times, with 95% of the medians falling between 6.51 and 17.92 Ma for the crown of *Alloteropsis*, and between 4.17 and 11.27 Ma for the split of *A. semialata* and *A. angusta* (Fig. III.6). The peak of values (i.e., 50% of the points) ranged between 9.38 and 13.07 Ma for the crown of *Alloteropsis* and 5.93 and 8.18 Ma for the split of *A. semialata* and *A. angusta* (Fig. III.6). Finally, 95% of the markers estimated the crown of *A. semialata* between 1.88 and 7.77 Ma, with a peak between 3.12 and 5.07 Ma (Fig. III.6). Note that monophyly of the groups was not enforced, and various combinations of *A. semialata* accessions were included across markers, contributing to the observed variation.

Most of the C_4-related genes, as well as the plastomes, provided age estimates ranging from 5.54 to 10.32 Ma for the split of *A. semialata* and *A. angusta*, which matches the distribution of estimates from the transcriptome-wide data (Fig. III.6A), and indicates their transmission followed the species tree. The only exception is the gene *pck-1P1_LGT-C*, for which the last common ancestor of *A. semialata* and *A. angusta* was estimated at 2.77 Ma (Fig. III.6A), which is smaller than all but four of the 2,797 estimates from the transcriptome-wide markers. While the confidence intervals of the estimate for this gene do overlap with those of almost all other markers, this estimate matches more closely the diversification of *A. semialat*a accessions (Fig. III.6A).

The different markers selected for detailed analyses similarly yielded estimates for the crown of *Alloteropsis* matching those obtained from transcriptome-wide data, between 9.38 and 16.46 Ma (Fig. III.6B). The only exception is the gene *ppc-1P3_LGT-M*, for which the last common ancestor of *A. cimicina* and *A. semialata* is estimated at 3.25 Ma (Fig. III.6B), which is smaller than all estimates based on markers extracted from the transcriptomes. The 95% CI of the divergence estimate based on this gene does not overlap with many of those based on other markers, and again matches closely with the diversification of *A. semialat*a accessions (Fig. III.6B).

Overall, our dating analyses support an introgression of these two genes among *Alloteropsi*s species after their divergence, whilst the other genes were transmitted following the species tree (Fig. III.2).

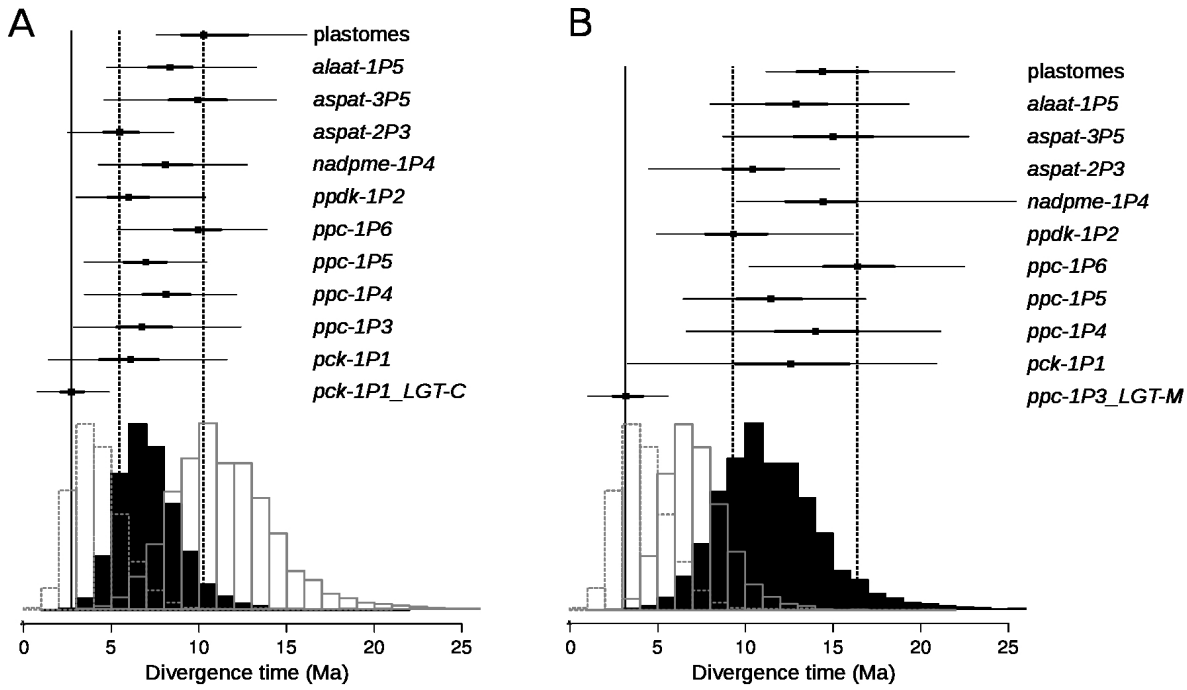


Figure III.6: Estimates of divergence times.
On the top, divergence times are shown for selected nuclear genes and plastomes for **A)** the split of *A. angusta* and *A. semialata* and **B)** the crown of Alloteropsis. For each marker, the median of the estimates is indicated by a square, with thick bars connecting the 25 and 75 percentiles and thin bars connect the 2.5 and 97.5 percentiles. The distribution of medians for the crown of A. semialata (left), the split of *A. angusta* and *A. semialata* (middle), and the crown of Alloteropsis (right) over 2,797 markers extracted from the transcriptomes is shown at the bottom. The scale is given in million years ago (Ma).

Discussion

Two independent transitions from C_3 to C_4

The earliest split in *Alloteropsis* separates the lineage containing *A. cimicina* from *A. angusta* and *A. semialata* (Fig. III.2). These two lineages co-opted different tissues for the segregation of Rubisco activity and achieved a large proportion of bundle sheath tissue via different modifications (Fig. III.S1). The evidence therefore strongly supports two independent origins of C_4 anatomical properties, which is generally accepted as the first step during the C_3 to C_4 transition (Fig. III.1; Sage et al. 2012; Heckmann et al. 2013). Gene expression analyses show that the two clades use different enzymes for parts of the C_4 cycle, express different genes encoding the same enzyme family when there is an overlap (Fig. III.3), and positive selection analyses show that the enzymes were independently adapted for their C_4 function (Table III.1). We therefore conclude that the different transitions to C_4 biochemistry occurred independently after the split of

these two lineages (Fig. III.2). The only exception to the distinctiveness of *A. cimicina* and the two other C_4 species is the gene *ppc-1P3_LGT-M*, used by both *A. cimicina* and some C_4 *A. semialata* accessions (Fig. III.4). This gene is absent from other accessions (Olofsson et al. 2016) and, as such, we previously concluded that it was acquired early during the diversification of the group and then recurrently lost (Christin et al. 2012). This hypothesis is falsified by our dating analyses here, which show that this gene was only recently transferred among species boundaries, likely as a result of a rare hybridization event (Fig. III.6).

One independent C_3 to C_4 transition includes two separate C_3+C_4 to C_4 shifts

The C_4 phenotype is realized in *A. angusta* and *A. semialata* via identical anatomical modifications, using the same enzymes, and the same genes encode these enzymes. Chloroplasts are present in the inner sheaths of all *A. semialata* and *A. angusta* accessions, independent of their photosynthetic type, which suggests that this characteristic represents the ancestral condition for the clade (Fig. III.2). The C_4 cycle is realized using the same set of genes in *A. angusta* and *A. semialata*, which can be explained by convergent evolution (e.g., as indicated for other C_4 grasses; Christin et al. 2013b) or a single origin of a weak C_4 cycle (C_3+C_4), followed by a reversal to expression levels that resemble the ancestral condition in the C_3 accessions (Fig. III.2). Differentiating these two scenarios would require retracing the origin of the mutations responsible for the increased expression of C_4 enzymes to identify where they occurred on the phylogeny. Unfortunately, the molecular mechanisms controlling C_4 gene expression are poorly known, and can involve both cis- and trans-acting elements (Gowik et al. 2004; Brown et al. 2011; Williams et al. 2016).

The positive selection analyses indicate that enzyme adaptation happened independently in *A. angusta* and *A. semialata* (Table III.2). Together with the variation observed within the C_4 *A. semialata* (Fig. III.4, III.S6), this evidence strongly suggests that the biochemical adaptation allowing the transition to a full C_4 cycle happened recently, and independently in the two species (Fig. III.2). The dramatic increase in the proportion of the inner bundle sheath tissue via the proliferation of minor veins is limited to the C_4 *A. semialata* and *A. angusta* (Fig. III.S1). The genetic control of these features is unknown, preventing a comparison of the causal mutations. However, the

distribution of anatomical characters among grasses indicates that the vast majority of C_4 lineages that co-opted the inner bundle sheath increased its proportion via the addition of minor veins (Renvoize 1987; Christin et al. 2013a).

With the current state of knowledge, we hypothesize that the common ancestor of *A. semialata* and *A. angusta* had chloroplasts in the inner bundle sheath, and that this facilitated the emergence of a weak C_4 cycle via the upregulation of some enzymes. Following their split, *A. angusta* strengthened its C_4 anatomy via the proliferation of minor veins, and enzyme adaptations led to a strong C_4 cycle (Fig. III.2). In the *A. semialata* lineage, some isolated populations acquired mutations that added minor veins and adapted the enzymes, leading to a C_4 cycle. Other populations, potentially under pressures linked to the colonization of colder environments (Lundgren et al. 2015), might have lost the weak C_4 cycle by down-regulating the genes (Fig. III.2). However, the details of the changes leading to C_3 photosynthesis in some *A. semialata* will need to be confirmed by comparative genomics, when mutations regulating expression of C_4 enzymes and anatomy are identified.

Introgression of C_4 components among species

Our dating analyses suggest that the gene *pck-1P1_LGT-C* that encodes the decarboxylating enzyme PCK was introgressed among some members of *A. semialata* and *A. angusta* (Fig. III.2 and III.6). The C_4 cycle carried out before this event was likely based on NADP-malic enzyme, an enzyme still abundant in the C_3+C_4 *A. semialata* and some C_4 accessions (Fig. III.4; Frean et al. 1983). The acquisition of *pck-1P1_LGT-C*, a gene already adapted for the C_4 context, probably added a PCK shuttle, which alters the stoichiometry of the pathway and the spatial distribution of its energy requirements, increasing its efficiency under some conditions (Bellasio and Griffiths 2014; Wang et al. 2014). This important component of the C_4 cycles of extant *A. semialata* and *A. angusta* populations first evolved its C_4-specific properties in the distantly-related *Cenchrus* (Fig. III.S3; Christin et al. 2012), and therefore never evolved within *Alloteropsis*. Instead, it represents the spread of a component of a complex physiology across multiple species boundaries. Therefore, in addition to the possibility that the sequential steps generating a complex physiology can happen on

different branches of a species phylogeny (Fig. III.2), introgression among close relatives can disconnect the origins of key components from the species tree.

On the inference of transitions among character states

Inferences of transitions among character states are a key component of numerous macro-evolutionary studies (e.g. Cantalapiedra et al. 2017; Cooney et al. 2017). However, species trees *per se* are not always able to disentangle the complex scenarios underlying the appearance or losses of multi-component adaptations, especially when complex phenotypes are modeled as different states of a single character (e.g. Goldberg and Igic 2008; Pardo-Diaz et al. 2012; Niemiller et al. 2013; Igic and Busch 2013; King and Lee 2015). In the case of photosynthetic transitions within *Alloteropsis* depicted here, considering the photosynthetic type as a binary character would lead to a single C_4 origin as the most plausible scenario (Ibrahim et al. 2009), and modeling photosynthetic types based on their category of C_4 cycle does not improve the inference (Washburn et al. 2015). For traits assumed to evolve via sequential stages, the accepted sequence of changes can be incorporated in the model (e.g. Marazzi et al. 2012). However, the power of character modeling remains inherently limited by the small number of informative characters. Decomposing the phenotype into its components can solve this problem, especially when the underlying genetic determinism is considered (Oliver et al. 2012; Niemiller et al. 2013; Glover et al. 2015; Meier et al. 2017), and good mechanistic models exist for the evolution of DNA sequences (Liberles et al. 2013). Violation of model assumptions can still mislead the conclusions, but the multiplication of sources of information, coupled with the possibility to track the history of specific genes independently of the species tree, limits the risks of systematic errors. We therefore suggest that efforts to reconstruct the transitions leading to important traits should integrate as many underlying components as possible. As progresses in genome biology increase data availability and improve our understanding of causal mutations, modeling phenotypes as the results of cumulative changes in genomes will be able to solve the problems raised by the paucity of informative characters.

Conclusions

In this study, we dissect the genetic and anatomical components of C_4 photosynthesis in *Alloteropsis*, a genus of grasses with multiple photosynthetic types. Our comparative efforts strongly support at least two independent origins of C_4 photosynthesis within this genus. The C_4 phenotype within these separate origins is realized via divergent anatomical modifications, the upregulation of distinct sets of genes, and independent enzyme adaptations. One of these lineages includes a range of photosynthetic types, and based on our analyses, we suggest that some C_4 components in this group evolved in the shared common ancestor, while others were acquired independently after the lineages diverged. The history of photosynthetic transitions within *Alloteropsis* is furthermore complicated by the introgression of C_4 genes across species boundaries. This disconnects the spread of C_4 components from the species tree, and means that the number of origins varies among the different components of the complex C_4 trait. This scenario is unlikely to have been inferred from traditional macroevolutionary approaches based on species trees alone. We suggest that integrating genomic data and phenotypic details in future studies of character transitions might resolve similarly complicated scenarios in other groups, enabling a better understanding of the trajectories followed during the evolution of novel adaptations.

Acknowledgments

This work was funded by a Royal Society Research Grant (grant number RG130448) to PAC. PAC and PN are supported Royal Society University Research Fellowships (grant numbers URF120119 and URF130423, respectively), LTD is supported by a NERC grant (grant number NE/M00208X/1), and MRL is supported by an ERC grant (grant number ERC-2014-STG-638333). All raw RNA-Seq data have been deposited in the NCBI Sequence Read Archive (project identifier SRP072730), and transcriptome assemblies are deposited in the NCBI Transcriptome Shotgun Assembly repository (Bioproject PRJNA310121).

Chapter III: Supplementary Information

Supplementary Methods

1.1 Plant growth conditions

Alloteropsis semialata, A. angusta, and *Panicum pygmaeum* plants were grown from seeds or propagated vegetatively from cuttings collected in the field. All individuals were maintained in controlled environment chambers (Conviron BDR16; Manitoba, Canada) set to 60% relative humidity, 500 µmol m^{-2} s^{-1} photosynthetic photon flux density (PPFD), and 25/20°C day/night temperatures with 14h of light at the University of Sheffield. Plants were grown in John Innes No. 2 potting compost (John Innes Manufacturers Association, Reading, England), maintained under well-watered conditions, and fertilised every two weeks (Scotts Evergreen Lawn Food; The Scotts Company, Surrey, England). After a minimum of 30 days in the above conditions, samples were taken for RNA-Seq and leaf anatomy. For RNA-Seq, certain individuals were then were resampled after 30 days under a 10-hour photoperiod (Table II.S3). Leaf samples from *A. cimicina* were collected from two individual plants grown under ambient glasshouse conditions at Brown University.

1.2 RNA-Seq protocol

Total RNA was extracted from *A. semialata*, *A. angusta*, and *P. pygmaeum* samples using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany), following the manufacturer's protocol. An on-column DNA digestion step was performed using the RNase-Free Dnase Set (Qiagen, Hilden, Germany). Total RNA was eluted in RNAse-free water with 20 U/µL of SUPERase-IN RNase Inhibitor (Life Technologies, Carlsbad, CA). RNA quality and concentration were determined using the RNA 6000 Nano kit with an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, California). Extractions used for library preparation contained at least 0.5 µg of total RNA, with an RNA integrity number (RIN) greater than 6.5. Each sample was prepared individually using the TruSeq RNA Library Preparation Kit v2 (Illumina, San Diego, CA), following the manufacturer's protocol with an eight-minute fragmentation step. Indexed libraries were paired-end sequenced by the Sheffield Diagnostic Genetics Service on an Illumina HiSeq 2500 platform for 100 cycles in rapid mode, with 24

libraries pooled per lane of the flow cell. The two *A. cimicina* leaf samples were sequenced as described in Christin et al. (2015).

The RNA-seq data were filtered and assembled using the Agalma pipeline v.0.5.0 with default parameters (Dunn et al. 2013). In brief, this pipeline removes the reads that are low quality (Q <30), adaptor contaminated, or correspond to rRNA, prior to constructing *de novo* assemblies using Trinity (version trinityrnaseq_r20140413p1; Grabherr et al. 2011). One assembly was generated per genotype, using all reads available for each accession (Table II.S3). All raw RNA-Seq data have been deposited in the NCBI Sequence Read Archive (project identifier SRP072730, Table II.S3), and transcriptome assemblies are deposited in the NCBI Transcriptome Shotgun Assembly repository (Bioproject PRJNA310121). To generate transcript abundances, the paired-end reads from each library were mapped back onto their respective reference transcriptome assembly using bowtie2 v.2.0.5 (Langmead and Salzberg 2012).

In total, 38 individually sequenced RNA-Seq libraries from 14 different accessions/species generated over 300 million 100 bp paired-end reads. This represents 66.44 Gb of data, with a mean of 1.75 Gb per library (SD = 1.56 Gb; Table II.S3). Over 80% of the data were kept after removing low-quality reads and ribosomal RNA sequences (Table II.S3). One transcriptome was assembled per genotype, pooling all the reads obtained for each genotype (mean per genotype = 3.87 Gb, SD = 1.76 Gb). The 14 assembled transcriptomes were all of comparable quality, with a mean of 44,578 trinity 'unigenes' (i.e. putative loci in the transcriptome assembly; SD = 6,905), 65,725 contigs (SD = 12,282), and a 1,543 bp N50 (SD = 167 bp).

1.3 Positive selection analysis

For each gene lineage, additional sequences were retrieved from complete published genomes for Panicoideae, the NCBI non-redundant nucleotide database, and other published transcriptomes (Bräutigam et al. 2014). The *pck-1P1* gene is expressed at extremely low levels in the C_4 *A. semialata* and *A. angusta* (see Results), and was therefore not assembled as part of the transcriptomes. As an alternative, we used coding sequences previously generated by Sanger sequencing when available (Christin et al. 2012), or manually assembled PCK coding regions from low coverage genome sequencing data (Olofsson et al. 2016). Each set of genes was aligned as codons using

ClustalW (Thompson et al. 2002), and the resulting alignments were manually refined, including truncating the 5' or 3' ends to remove poorly aligned segments. A phylogenetic tree was inferred on 3rd positions of codons, using PhyML, with the GTR+G+I model and 100 bootstrap pseudoreplicates. The gene tree topologies were used for subsequent selection analyses, after removing sequences belonging to C_4 species other than *Alloteropsis* to avoid an influence of positive selection in these taxa affecting our conclusion. C_3 species outside *Alloteropsis* were however kept for positive selection analyses.

For genes not involved in the C_4 cycle of *A. cimicina*, we repeated the positive selection analyses to distinguish between a single (common ancestor of *A. angusta* and *A. semialata*) and two episodes of adaptive evolution (*A. angusta* and C_4 *A. semialata* separately) within the *A. angusta*/*A.semialata* clade. This was also preformed with the hypothesis of positive selection acting only in *A. angusta*. In addition, we also performed these tests on the genes for which selection was detected on the branch leading to *A. cimicina*, after excluding *A. cimicina* sequences, to evaluate the possibility that selection operated on different sites in the different lineages.

1.4 Alignment and filtering

Stringent alignment and filtering methods were used to ensure reliable alignments of each gene family for phylogenetic inference. First, sequences within each gene family were translated and aligned as proteins with four different assemblers (maaft, muscle, kalign, t-coffee) using m-coffee (Wallace et al. 2006) as part of the t-coffee package v.11.0 (Notredame et al. 2000). Consensus alignments from the four different methods were then trimmed so that only amino acids aligned in the same position by all of the assemblers were retained. Alignments were further parsed using the tcs residue filter (Chang et al. 2014), only retaining the highest confidence residues. The trimmed protein alignments were reverse-translated into nucleotide alignments using the original sequences, and further filtered using gblocks v.0.91 (parameters: -t=c -b2=b1 -b5=h; Castresana 2000). Finally, sequences shorter than 100 bp were removed, and maximum likelihood trees were inferred with PhyML. Putative groups of Panicoideae co-orthologs were identified as monophyletic groups that contained only sequences from Panicoideae species. The alignment process was repeated for each of these groups of putative co-

orthologs, starting again from the initial untrimmed sequences, producing high quality alignments for each individual group. Subsequent analyses were restricted to groups containing at least one sequence of each *Alloteropsis* species and *Sorghum* (used as the outgroup), and phylogenetic trees were again inferred with PhyML. Datasets where at least one of the six Panicoideae species (*Sorghum*, *Setaria*, *P. pygmaeum*, *A. cimicina*, *A. angusta*, and *A. semialata*), the *Alloteropsis* genus, or the *semialata*/*angusta* clade was not monophyletic in the maximum likelihood tree were discarded to remove genes that were duplicated after the divergence from the outgroup or poorly informative datasets. Of the 4,969 datasets originally screened, 1,042 were discarded because at least one of *Sorghum*, *Setaria*, *P. pygmaeum*, or the *Alloteropsis* genus was not monophyletic. These include potential cases of paralogy problems, sequencing or assembly errors, and poor phylogenetic resolution in the deep nodes of the trees. A further 1,130 datasets were discarded because one of the *Alloteropsis* species or the *A. semialata*/*A. angusta* clade was not monophyletic. This category includes potential *Alloteropsis*-specific duplicates and datasets lacking resolution among these closely-related taxa. While it cannot be excluded that some of these incongruences reflect true biological phenomena, the remaining 2,797 datasets (56% of the original ones) represent reliable markers for dating analyses. Finally, we removed species-specific duplicates, or transcript variants, by only retaining the longest sequence for each accession when several sequences from that accession formed a monophyletic clade.

Figure III.S1: Comparisons of leaf anatomy in *Alloteropsis* and relatives.

Leaf cross-sections are shown for each group. The red bar at the bottom represents 0.5 mm. Black arrows indicate mesophyll cells (M), red arrows inner sheath (IS) cells and orange arrows outer sheath (OS) cells. The species and photosynthetic type are indicated on the right.

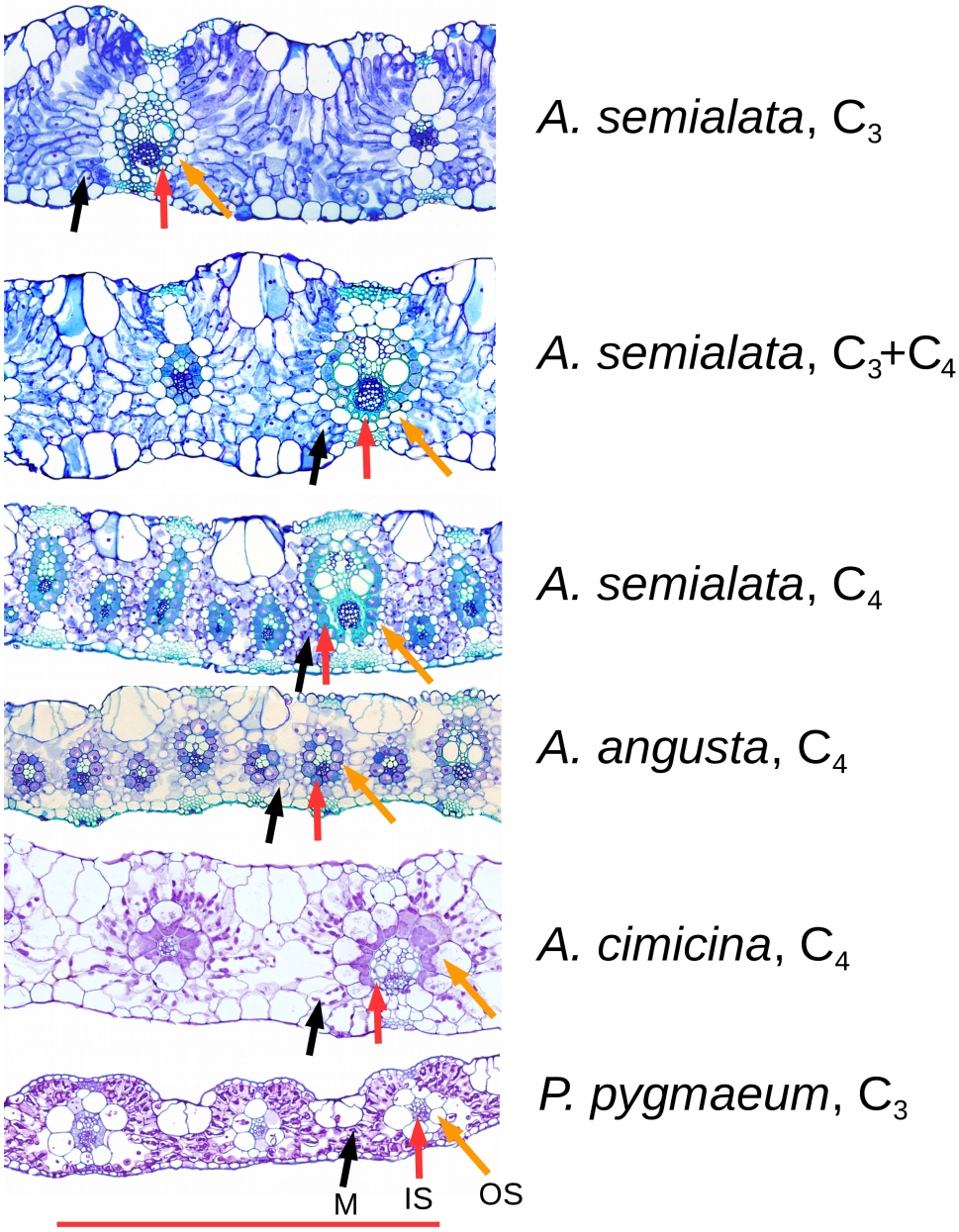
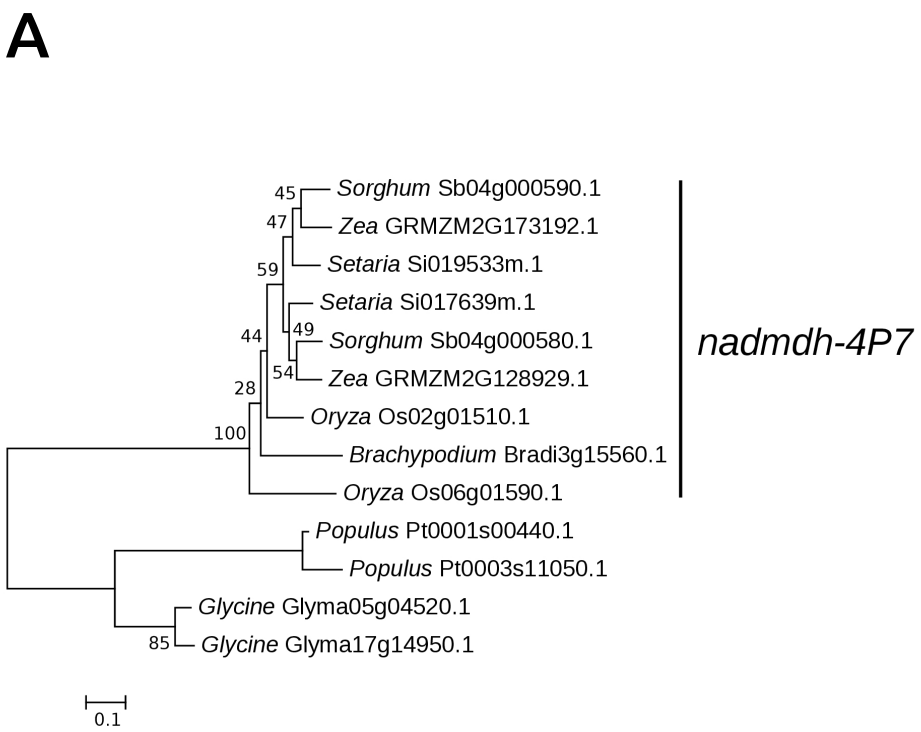
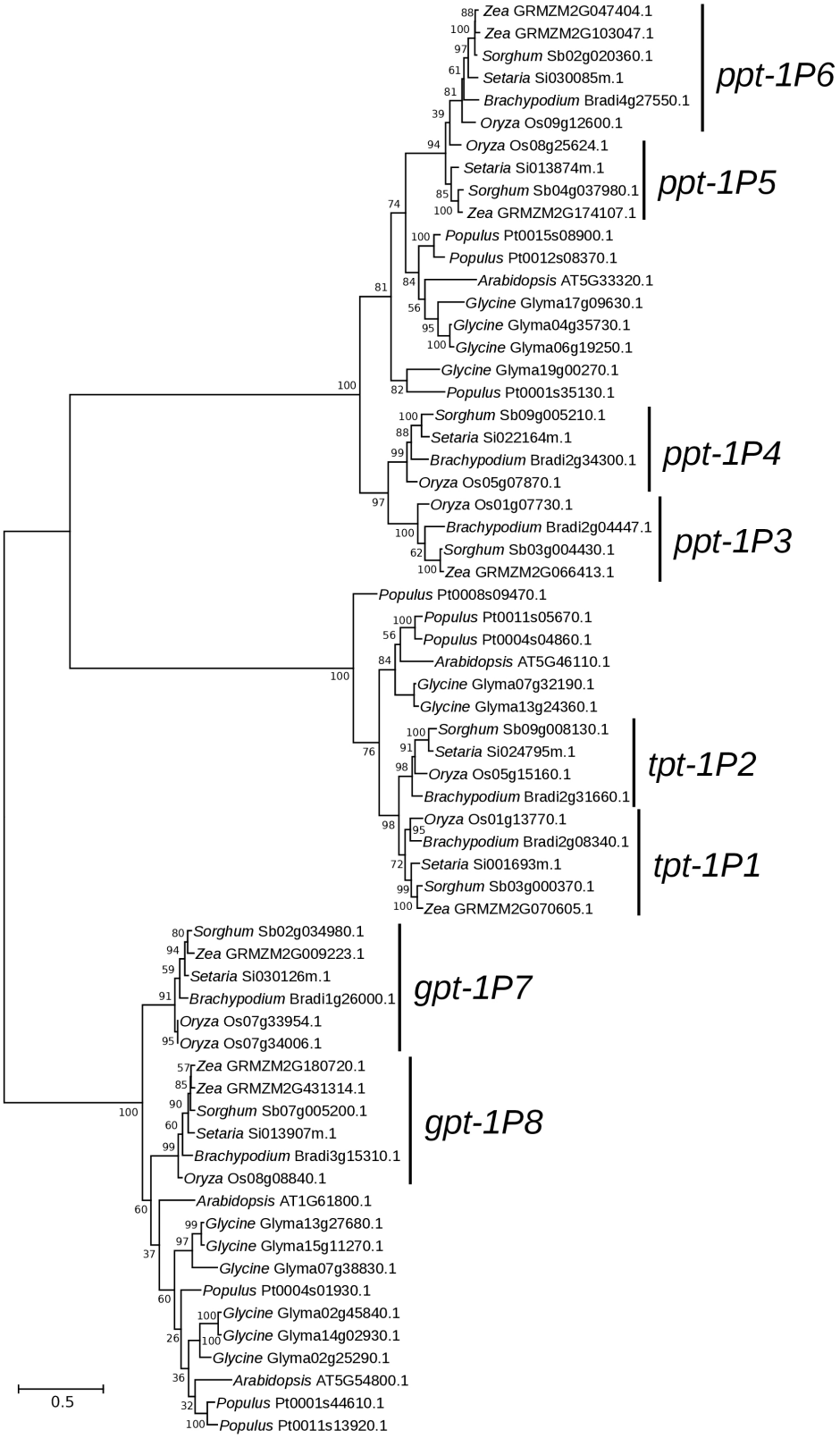


Figure III.S2: Phylogenetic trees for genes encoding three C_4-related enzymes.

These maximum likelihood trees show the relationships among genes used to circumscribe grass co-orthologs for the phylogenetic annotation of contigs. The trees are shown for three families changed compared to Christin et al. (2013, 2015); A) NAD-malate dehydrogenase (*nadmdh-4*), B) phosphoenolpyruvate-phosphate translocator (*ppt*)/triosephosphate-phosphate translocator (*tpt*)/glucose-6-phosphate/phosphate translocator (*gpt*), C) Sodium bile acid symporter (*sbas*). For each tree, grass co-orthologs are delimited on the right, with names following the approach of Christin et al. (2015). Bootstrap values are indicated near branches.



B


Zea GRMZM2G047404.1
Zea GRMZM2G103047.1
Sorghum Sb02g020360.1
Setaria Si030085m.1
Brachypodium Bradi4g27550.1
Oryza Os09g12600.1
ppt-1P6
Oryza Os08g25624.1
Setaria Si013874m.1
Sorghum Sb04g037980.1
Zea GRMZM2G174107.1
ppt-1P5
Populus Pt0015s08900.1
Populus Pt0012s08370.1
Arabidopsis AT5G33320.1
Glycine Glyma17g09630.1
Glycine Glyma04g35730.1
Glycine Glyma06g19250.1
Glycine Glyma19g00270.1
Populus Pt0001s35130.1
Sorghum Sb09g005210.1
Setaria Si022164m.1
Brachypodium Bradi2g34300.1
Oryza Os05g07870.1
ppt-1P4
Oryza Os01g07730.1
Brachypodium Bradi2g04447.1
Sorghum Sb03g004430.1
Zea GRMZM2G066413.1
ppt-1P3
Populus Pt0008s09470.1
Populus Pt0011s05670.1
Populus Pt0004s04860.1
Arabidopsis AT5G46110.1
Glycine Glyma07g32190.1
Glycine Glyma13g24360.1
Sorghum Sb09g008130.1
Setaria Si024795m.1
Oryza Os05g15160.1
Brachypodium Bradi2g31660.1
tpt-1P2
Oryza Os01g13770.1
Brachypodium Bradi2g08340.1
Setaria Si001693m.1
Sorghum Sb03g000370.1
Zea GRMZM2G070605.1
tpt-1P1
Sorghum Sb02g034980.1
Zea GRMZM2G009223.1
Setaria Si030126m.1
Brachypodium Bradi1g26000.1
Oryza Os07g33954.1
Oryza Os07g34006.1
gpt-1P7
Zea GRMZM2G180720.1
Zea GRMZM2G431314.1
Sorghum Sb07g005200.1
Setaria Si013907m.1
Brachypodium Bradi3g15310.1
Oryza Os08g08840.1
gpt-1P8
Arabidopsis AT1G61800.1
Glycine Glyma13g27680.1
Glycine Glyma15g11270.1
Glycine Glyma07g38830.1
Populus Pt0004s01930.1
Glycine Glyma02g45840.1
Glycine Glyma14g02930.1
Glycine Glyma02g25290.1
Arabidopsis AT5G54800.1
Populus Pt0001s44610.1
Populus Pt0011s13920.1
0.5

C Fig. S2 C

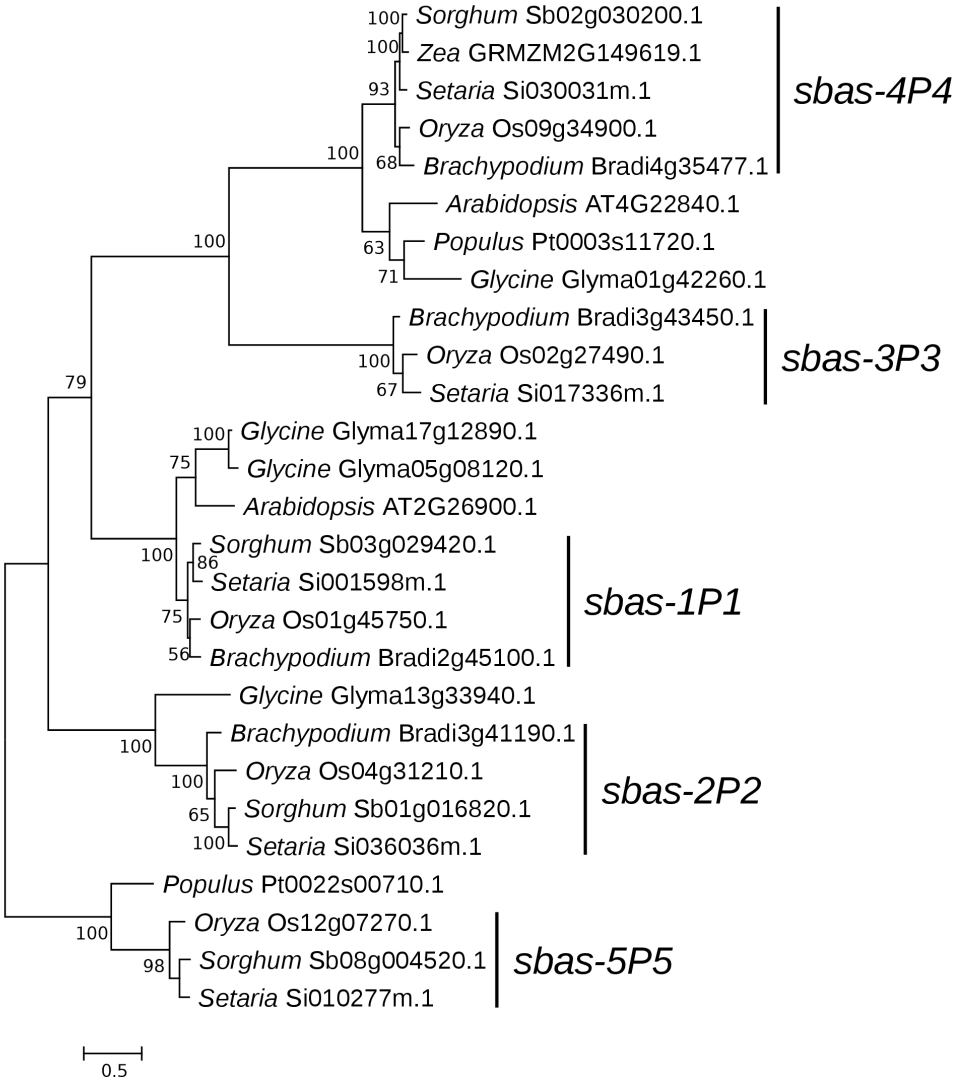

Sorghum Sb02g030200.1
Zea GRMZM2G149619.1
Setaria Si030031m.1
Oryza Os09g34900.1
Brachypodium Bradi4g35477.1
Arabidopsis AT4G22840.1
Populus Pt0003s11720.1
Glycine Glyma01g42260.1
Brachypodium Bradi3g43450.1
Oryza Os02g27490.1
Setaria Si017336m.1
Glycine Glyma17g12890.1
Glycine Glyma05g08120.1
Arabidopsis AT2G26900.1
Sorghum Sb03g029420.1
Setaria Si001598m.1
Oryza Os01g45750.1
Brachypodium Bradi2g45100.1
Glycine Glyma13g33940.1
Brachypodium Bradi3g41190.1
Oryza Os04g31210.1
Sorghum Sb01g016820.1
Setaria Si036036m.1
Populus Pt0022s00710.1
Oryza Os12g07270.1
Sorghum Sb08g004520.1
Setaria Si010277m.1
sbas-4P4
sbas-3P3
sbas-1P1
sbas-2P2
sbas-5P5
0.5

Figure III.S3: Phylogeny of *pck-1P1* genes in Panicoideae.

This phylogenetic tree was inferred on 3rd positions of codons. Bootstrap values are indicated near branches. Branches leading to genes that have been co-opted for C_4 photosynthesis are in green, following Christin et al. (2012). Tribes are delimited on the right. The laterally acquired *pck-1P1-C* gene is indicated.

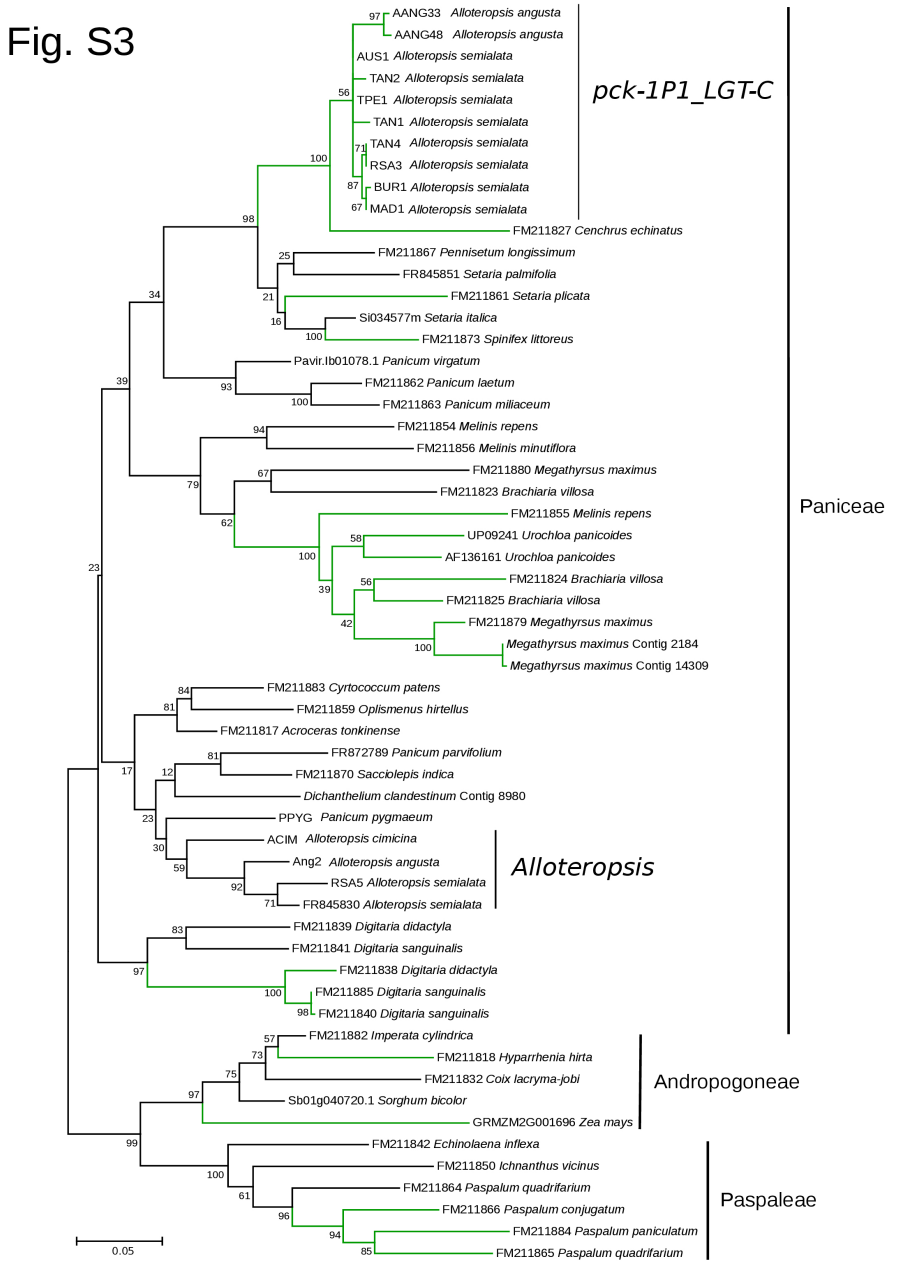


Figure III.S4: Evolution of *ppc-1P3* genes in *Alloteropsis* and other Panicoideae.

This phylogenetic tree was inferred on 3rd positions of codons of *ppc-1P3* genes of Panicoideae. Bootstrap values are indicated near branches. Names of C_4 accessions are in bold. Gray branches were pruned before selection tests. Positive selection was detected on the thick branch. Groups of *Alloteropsis* genes are delimited on the right.

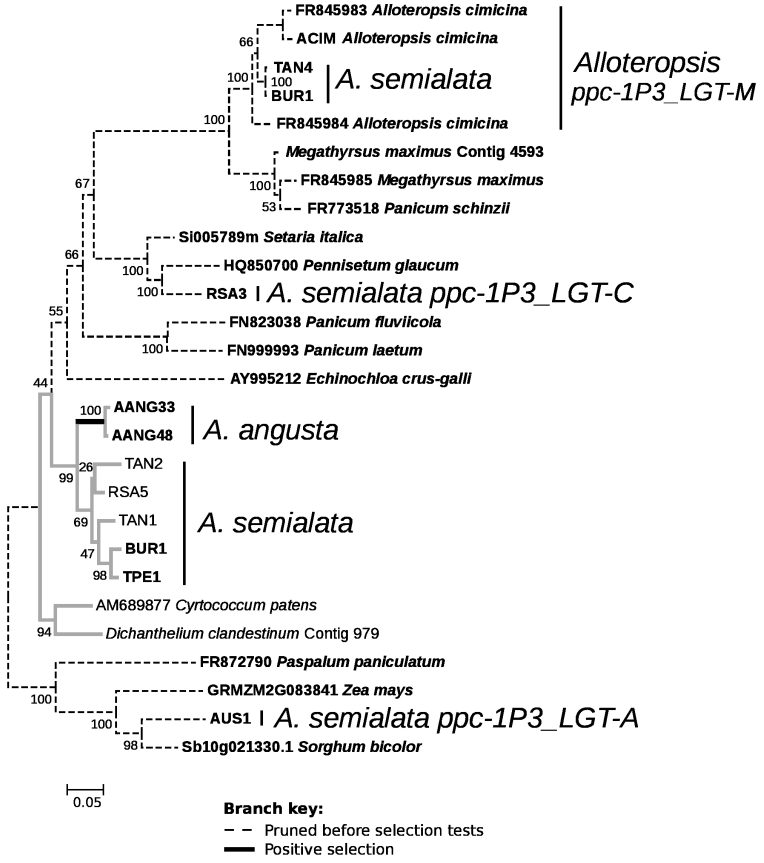


Figure III.S5: Evolution of *ppdk-1P2* genes in *Alloteropsis* and other Panicoideae.

This phylogenetic tree was inferred on 3rd positions of codons of *ppdk-1P2* genes of Panicoideae. Bootstrap values are indicated near branches. Names of C_4 accessions are in bold. Gray branches were pruned before selection tests. Positive selection was detected on the thick branch. Amino acid positions with a posterior probability >0.90 of being under positive selection are indicated on the right, asterisks indicate positions with a posterior probability >0.95, with those associated with C_4 accessions in gray. Positions are indicated on the top, based on *Sorghum* gene Sb09g019930.1.

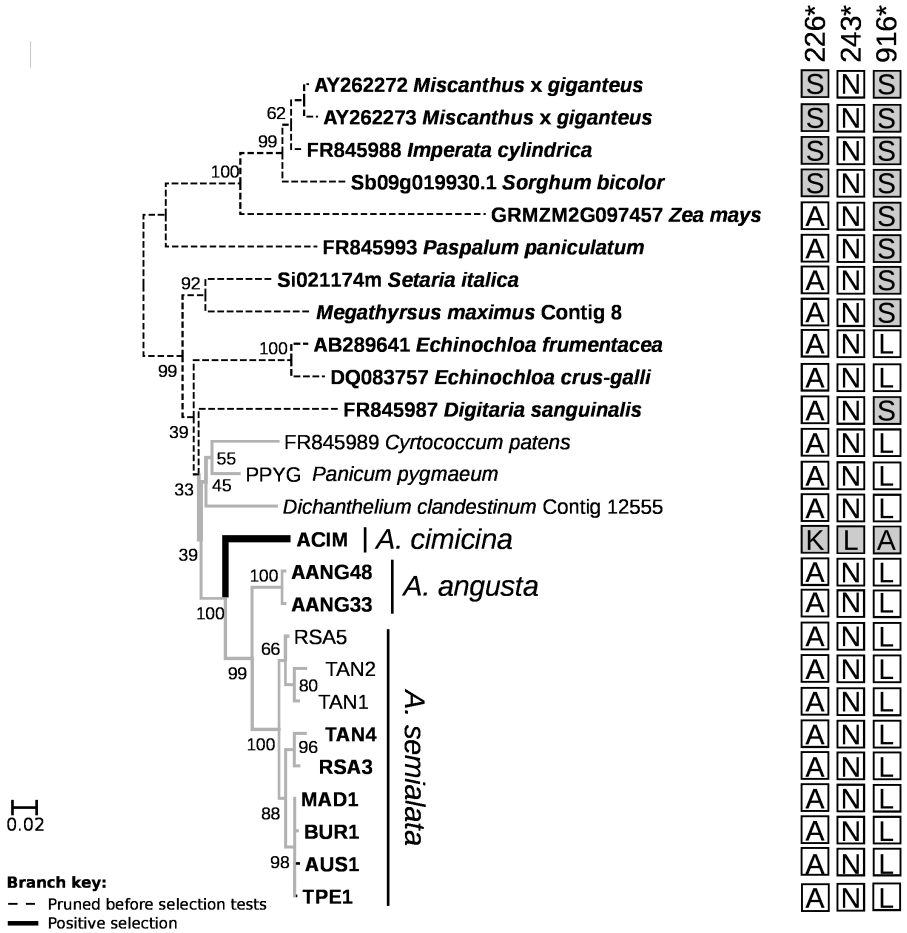


Figure III.S6: Evolution of *aspat-3P4* genes in *Alloteropsis* and other Panicoideae.

This phylogenetic tree was inferred on 3rd positions of codons of *aspat-3P4* genes of Panicoideae. Bootstrap values are indicated near branches. Names of C_4 accessions are in bold. Gray branches were pruned before selection tests. Positive selection was detected on thick branches. Amino acid positions with a posterior probability >0.90 of being under positive selection are indicated on the right, asterisks indicate positions with a posterior probability >0.95, with those associated with C_4 accessions in gray. Positions are indicated on the top, based on *Sorghum* gene Sb03g035220.1. Asterisks indicate positions with a posterior probability >0.9.

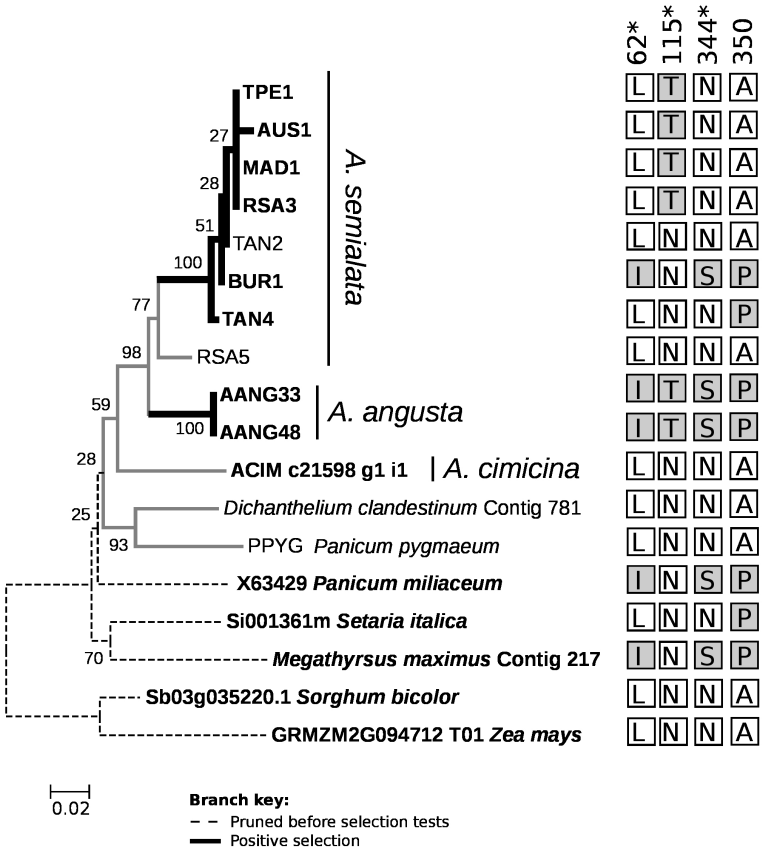


Table III.S1: *Alloteropsis semialata* accessions used in this study[1].

ID	Sample name	Country	Latitude	Longitude	Type	$\delta^{13}C$
RSA5	KWT3	South Africa	-32.70	27.53	C_3	-26.3
TAN2	L01	Tanzania	-5.63	32.69	C_3+C_4	-26.3
TAN1	L04	Tanzania	-8.51	35.17	C_3+C_4	-23.1
TAN4	L02	Tanzania	-9.04	32.48	C_4	-11.4
BUR1	BF3	Burkina Faso	10.85	-4.83	C_4	-11.3
MAD1	Maj	Madagascar	-15.67	46.37	C_4	-11.8
RSA3	MDB8	South Africa	-25.76	29.47	C_4	-12.7
RSA4	SFD3	South Africa	-28.39	29.04	C_4	-12.7
AUS1	Aus2	Australia	-19.62	146.96	C_4	-12.1
TPE1	TW10	Taiwan	24.47	120.72	C_4 [2]	-11.8

[1] Collection localities and photosynthetic type with the diagnostic physiology data come from Lundgren et al. (2016). Stable isotope data from Lundgren et al. (2015). [2]Inferred from stable isotope data, adjusted for anthropogenic CO_2 sources per Lundgren et al. (2016).

Table III.S4. Transcript abundance (in rpkm) for each C_4-related gene and sample.

http://onlinelibrary.wiley.com/store/10.1111/evo.13250/asset/supinfo/evo13250-sup-0002-tableS4.xls?v=1&s=9a2546d2bcceec96af525c83e4c11c081a1375d3

Table III.S2: Leaf anatomical data for the study species and accessions[1].

Species/Accession	Pathway	C_4 sheath	minor veins[2]	IVD (µm)	nb.M	OS.width (µm)	IS.width (µm)	OS:IS
Entolasia marginata[3]	C_3	na	absent	255.5	7.5	24.9	4.7	5.3
Panicum pygmaeum[3]	C_3	na	absent	219.9	7.0	27.9	6.2	4.5
Alloteropsis semialata, RSA5	C_3	na	absent	186.2	7.4	12.2	6.8	1.8
Alloteropsis semialata, TAN2	C_3+C_4	inner	absent	167.4	4.4	12.4	10.1	1.2
Alloteropsis semialata, TAN1	C_3+C_4	inner	absent	159.8	5.4	11.8	9.4	1.3
Alloteropsis semialata, TAN4	C_4	inner	present	127.8	2.7	9.7	14.0	0.7
Alloteropsis semialata, AUS1	C_4	inner	present	79.1	1.8	10.4	12.0	0.9
Alloteropsis semialata, BUR1	C_4	inner	present	77.9	1.3	11.3	10.6	1.1
Alloteropsis semialata, MAD1	C_4	inner	present	92.9	1.8	9.7	13.3	0.7
Alloteropsis semialata, RSA3	C_4	inner	present	86.0	1.8	9.6	11.7	0.8
Alloteropsis semialata, RSA4	C_4	inner	present	97.2	1.8	8.4	13.3	0.6
Alloteropsis semialata, TPE1	C_4	inner	present	84.5	1.2	8.7	7.2	1.2
Alloteropsis angusta	C_4	inner	present	83.4	1.0	9.5	9.8	1
Alloteropsis cimicina[3]	C_4	outer	absent	292.3	3.4	46.7	6.0	7.8
Alloteropsis paniculata[3]	C_4	outer	absent	198.0	2.7	43.9	5.6	7.8

[1] Column headings and abbreviations: C_4 sheath, bundle sheath used for CO_2 reduction; IVD, interveinal distance; nb.M, number of mediolateral mesophyll cells separating vein units; OS.width, the width of the outer bundle sheath cells; IS.width, the width of the inner bundle sheath cells; OS:IS is the ratio of outer to inner bundle sheath cell size. [2] Minor veins are considered 4th and 5th order veins here, while the midrib, secondary and tertiary vein orders are excluded from this category. [3] Data taken from Christin et al. 2013.

Table III.S3: RNA-Seq data, NCBI SRA accession numbers, and growth conditions.

Genotype	Species	SRA	Tissue	Photoperiod	Raw PE	Clean PE reads	No. Trinity contigs
ACIM	*A. cimicina*	SRR3994072	Leaf	Glasshouse	36087907	27351333	51195
		SRR3994073	Leaf	Glasshouse	28714973	4854902	
ANG33	*A. angusta*	SRR3994075	Leaf	14hr	7334658	6612013	72468
ANG48	*A. angusta*	SRR3994077	Leaf	14hr	7498597	6826154	71835
AUS1	*A. semialata*	SRR3321311	Leaf	10hr	5308967	4675722	54197
		SRR3322358	Leaf	14hr	11184717	9624467	
		SRR3322714	Root	14hr	3950267	3613034	
BUR1	*A. semialata*	SRR3322990	Leaf	10hr	11153698	9575675	75444
		SRR3322973	Leaf	14hr	16859344	14948845	
		SRR3323003	Root	14hr	2223260	3042111	
RSA5	*A. semialata*	SRR3323066	Leaf	10hr	5500526	2402701	63273
		SRR3323049	Leaf	14hr	13458893	12135442	
		SRR3323067	Root	14hr	3107385	2915544	
TAN2	*A. semialata*	SRR3323068	Leaf	10hr	17997033	16131010	74639
		SRR3323088	Leaf	14hr	5423680	4804035	
		SRR3323114	Root	14hr	4282037	4018356	
TAN4	*A. semialata*	SRR3323124	Leaf	14hr	3218981	3218981	58125
		SRR3323125	Root	14hr	4689624	4689624	
TAN1	*A. semialata*	SRR3323127	Leaf	10hr	25015574	22153077	74400
		SRR3323128	Root	10hr	11154350	9983220	
		SRR3323129	Root	14hr	3137368	2928130	
MAD1	*A. semialata*	SRR3323131	Leaf	10hr	1338699	1146407	75444
		SRR3323132	Leaf	14hr	2546427	2190484	
		SRR3323133	Root	14hr	10980770	9826198	
		SRR3323134	Root	10hr	3460229	3201082	
RSA3	*A. semialata*	SRR3323186	Leaf	10hr	5001282	2021239	74023
		SRR3323137	Leaf	14hr	11922494	10671710	
		SRR3323187	Root	14hr	3950750	4185063	
PPYG	*P. pygmaeum*	SRR3330791	Leaf	14hr	4093890	3793221	72117
		SRR3323220	Leaf	14hr	8925603	8087404	
		SRR3330803	Root	14hr	4106624	3482683	
		SRR3330803	Root	14hr	2026469	1752009	
RSA4	*A. semialata*	SRR3323240	Leaf	10hr	4467544	3470414	87362
		SRR3323220	Leaf	14hr	16357704	14849292	
		SRR3323241	Root	14hr	3248828	4614268	
TPE1	*A. semialata*	SRR3323242	Leaf	10hr	12422286	6546514	57350
		SRR3323243	Leaf	14hr	7457117	10995742	
		SRR3323244	Root	14hr	2604885	3699333	

Table III.S5: Results of positive selection analyses inferring the episodes of enzymatic adaptation in *Alloteropsis*[1] using only codons with fixed nucleotides for each photosynthetic type within *A. semialata* and *A. angusta.*

Gene	Number of sequences	Number codons removed	Site model M1a	One origin		Two origins		Three origins		Only *A. cimicina*	
				BSA	BSA1	BSA	BSA1	BSA	BSA1	BSA	BSA1
aspat-2P3	7	23	**0.00***	4.05	4.05	4.05	4.05	3.54	3.54	4.05	4.05
nadpme-1P4	8	29	25.14	6.13	4.43	6.13	4.43	11.68	9.83	3.77	**0.00***
ppdk-1P2	8	48	26.34	11.49	8.91	11.49	8.91	9.67	4.73	4.27	**0.00***
alaat-1P5	7	33	**0.00***	4.03	4.03	4.04	4.04	3.61	3.61	4.04	4.04

[1] The ΔAICc values compared to the best fit model for that gene are shown. The most appropriate model is indicated with an asterisk, with the null model (M1a) only rejected if the ΔAICc was at least 5.22 (equivalent to a p-value of 0.01 with a likelihood ratio test with df = 2). Two branch-site models were used to test for a relaxation of purifying selection (BSA), and potential positive selection (BSA1).

Table III.S6: Results of positive selection analyses inferring the episodes of enzymatic adaptation in the *A. angusta/A. semialata* clade[1] using only codons with fixed nucleotides for each photosynthetic type within *A. semialata* and *A. angusta.*

Gene	Number of sequences	Number codons removed	Site model M1a	One origin		Two origins		Only *A. angusta*	
				BSA	BSA1	BSA	BSA1	BSA	BSA1
aspat-3P4	7	13	8.62	12.67	12.67	**0.00***	**0.00***	**0.00***	**0.00***
nadpme-1P4	7	29	**0.00***	4.04	4.04	3.47	1.19	3.71	3.71
ppc-1P3	6	70	65.71	23.63	22.33	9.50	5.83	6.17	**0.00***
ppdk-1P2	7	48	**0.00***	4.02	4.02	3.38	3.38	3.20	3.20

[1] The ΔAICc values compared to the best fit model for that gene are shown. The most appropriate model is indicated with an asterisk, with the null model (M1a) only rejected if the ΔAICc 5.22 (equivalent to a p-value of 0.01 with a likelihood ratio test with df = 2). Two branch-site models were used to test for a relaxation of purifying selection (BSA), and potential positive selection (BSA1).

Table III.S7: Effect of gene tree topology on the conclusions of the positive selection analyses in *Alloteropsis*[1].

Gene	Site model M1a	One origin		Two origins		Three origins		Only *A. cimicina*	
		BSA	BSA1	BSA	BSA1	BSA	BSA1	BSA	BSA1
aspat-2P3	**100**	0	0	0	0	0	0	0	0
nadpme-1P4	0	0	0	0	0	0	0	0	**100**
ppdk-1P2	0	0	0	0	0	0	0	0	**100**
alaat-1P5	**100**	0	0	0	0	0	0	0	0

[1] The number of topologies favouring each modelled, out of 100 bootstrap pseudoreplicates, is indicated.

Table III.S8: Assessing the effect of gene tree topology on the conclusions of the positive selection analyses within *A. semialata* and *A. angusta*[1].

	Site model	One origin		Two origins		Only *A. cimicina*	
Gene	**M1a**	**BSA**	**BSA1**	**BSA**	**BSA1**	**BSA**	**BSA1**
aspat-3P4	0	0	0	0	**100**	0	0
nadpme-1P4	**2**	0	0	0	**98**	0	0
ppc-1P3	0	0	0	0	0	0	**100**
ppdk-1P2	**100**	0	0	0	0	0	0

[1] The number of topologies favouring each modelled, out of 100 bootstrap pseudoreplicates, is indicated.

Chapter IV: Evidence for ancient and recurrent reticulate evolution in Panicoideae

Jose J. Moreno-Villena[1], Claudia Solis-Lemus[2], Colin P. Osborne[1], Pascal-Antoine Christin[1,3]

[1] Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom.

[2] Department of Human Genetics, Emory University, Atlanta GA 30322, United States of America.

[3] Corresponding author: Pascal-Antoine Christin, p.christin@sheffield.ac.uk, +44-114-222-0027

Personal contribution: I designed the research, analysed the data and wrote the paper with the help of all co-authors.

Abstract

Grasses are among the most ecologically and economically groups of plants, and their phylogenetic relationships have been extensively investigated to understand their evolutionary history and its impact on ecological, genomic and functional diversification. Historically, these efforts relied heavily on markers from the plastid genomes, which were recently expanded to whole plastomes and some genome-wide analyses of nuclear genes. These revealed numerous discrepancies among markers, but their causes were never formally investigated. In this study, we infer the phylogenetic relationships panicoid grasses from thousands of nuclear genes. Different approaches are used to infer a species tree from individual genes, and to explicitly test for incomplete lineage sorting and hybridization events. Our analyses reveal discrepancies among genes regarding relationships among members of the Paniceae subtribe, which are partially explained by incomplete lineage sorting due to rapid diversification. However, hybridization events are also supported, and these involve in several cases species representing presumably independent origins of C_4 photosynthesis. We conclude that hybridization has punctuated the history of Paniceae, on top of important lineage sorting. Because the inferred events involved distant species, we hypothesize that they might have contributed to the lateral spread of this trait boosting productivity in tropical conditions.

Keywords: Hybridization, phylogenetic networks, grasses, species tree, transcriptomics.

Introduction

The grass family (Poaceae) includes more than 11,000 species, which are spread around the world (Soreng et al. 2017). They occur in most habitats, ranging from the tropical rainforests, to the warm deserts and subarctic tundras. This ecological diversity has been made possible by the acquisition of various physiological attributes, including cold or desiccation tolerance, and the capacity to regrow after grazing or fire disturbance (Fernandez & Reynolds et al. 2000; Sandve et al. 2011; Scheiter et al. 2012). One key adaptation underlying the success of many grasses of tropical and subtropical regions is C_4 photosynthesis. This combination of anatomical and biochemical novelties boosts productivity in high-light, warm conditions (Atkinson et al. 2016). It evolved more than 60 times independently in flowering plants, and at least 22 times just in grasses (Sage et al. 2011; GPWGII 2012). C_4 grasses are extremely successful in tropical regions, where they dominate all open habitats, and together account for one quarter of terrestrial primary production (Still et al. 2003).

Over the years, many efforts have been put in establishing the phylogenetic relationships among grasses (GPWG 2001; Bouchenak-Khelladi et al. 2008; GPWG 2012; Bouchenak-Khelladi et al. 2014; Kellogg 2015; Soreng et al. 2015, 2017). The grass family has been subdivided into 13 subfamilies, but most species belongs to one of the two sister clades referred to as BEP and PACMAD clades (GPWG 2001; GPWGII 2012; Soreng et al. 2017). While they are of similar size, with ca. 5000 species each, the BEP clade contains only C3 species, and all 22-24 C_4 lineages are found in the PACMAD clade, alongside many C3 taxa (GPWGII 2012). Of the six PACMAD subfamilies, the Panicoideae contains the highest diversity of photosynthetic types, with 18 C_4 lineages separated by C3 species in the phylogeny and several C3-C_4 intermediates (GPWGII 2012; Christin et al. 2013). This group has consequently been the focus of repeated phylogenetic investigations, which resulted in important taxonomic changes (Aliscioni et al. 2001, 2003; Duvall et al. 2003; Giussani et al. 2003; Sanchez-Ken et al. 2010; Morroe et al. 2012; Burke et al. 2016). Historically, phylogenetic relationships were primarily investigated using few selected markers from the plastid genomes, approaches that were recently expanded to complete chloroplast genomes (Besnard et al. 2014). With some exceptions (e.g. Vicentini et al. 2008), the species tree was not evaluated with nuclear markers, which were only used to infer the

evolution of specific genes (e.g. Mathews et al. 1999; Christin et al. 2007, 2009). Discrepancies between individual nuclear markers and plastid genomes appeared (Christin et al. 2009), but these were difficult to interpret with only few nuclear markers. More recently, high-throughput sequencing enabled inferences based on numerous markers, which provided nuclear trees that recovered the main families and tribes, but presented differences with the plastid analyses (Moreno-Villena et al. 2018; Washburn et al. 2017). The causes of these discrepancies have however never been statistically evaluated.

Differences among markers might be due to a lack of information, phylogenetic errors due to systematic biases (e.g. Christin et al. 2012), or other processes such as incomplete lineage sorting (Maddison et al. 2006). These are particularly frequent in the case of rapid diversification (Koblmüller et al. 2010; Rannala & Yang 2008) . Alternatively, discrepancies among markers can arise from reticulate evolution, including hybridization, allopolyploidization, and lateral gene transfers (Walker et al. 2017; Maddison et al. 2006). All of these have been reported for specific groups of grasses (e.g. Manson-Gamer 2004, 2010; Marcussen et al. 2014), including panicoid grasses (Christin et al. 2012; Estep et al. 2014; Dunning et al. 2017), but their contribution to large-scale phylogenetic problems remains unaddressed. The advent of high-throughput sequencing enabled inferring phylogenetic trees based on genome-wide datasets, although such studies generally still looked for a single dichotomous species trees (Uddin et al. 2003; Leache et al. 2014). Other approaches however assume reticulate evolution, producing trees where multiple ancestral groups can contribute to the same descendant, and provide explicit tests for historical reticulate evolution (Huson et al. 2010; Genner & Turner 2012). To our knowledge, these methods have never been applied to panicoid grasses.

In this study, we combine a number of approaches to infer the phylogenetic relationships among Panicoideae from thousands of nuclear markers. Using different approaches, we (i) test for incongruence between nuclear and plastid genomes and among individual nuclear markers. By expanding the traditional species tree into a set of alternative topologies, we then explicitly test for (ii) incomplete lineage sorting in the whole or subgroups of Panicoideae, and (iii) reticulate evolution across the subfamily. Our investigations shed new light on the phylogenetic history of Panicoideae, which is

used to discuss the influence of non-dichotomous trees on the spread of adaptive traits, such as C_4 photosynthesis.

Methods

Species sampling and identification of one-to-one orthologs

The species sampling focused on the Panicoideae subfamily of grasses. Coding Sequences (CDS) were retrieved from three publicly available grass genomes (*Zea mays*, *Sorghum bicolor*, and *Setaria italica*), while CDS for another 13 grass species were retrieved from published transcriptomes from Moreno-Villena et al. (2018). This resulted in one representative of the outlaying Panicoideae (*Chasmanthium*, which served as outgroup), two Andropogoneae, three Paspaleae, and ten Paniceae (Table IV.1).

Table IV.1. List of Panicoideae species used in this study.

Species	Abbreviation	Subtribe	Tribe
*Setaria italica**	Sitalica	*Cenchrinae*	*Paniceae*
Setaria barbata	SBAR	*Cenchrinae*	*Paniceae*
Panicum queenslandicum	PQUE	*Panicinae*	*Paniceae*
Digitaria ciliaris	DCIL	*Anthephorinae*	*Paniceae*
Sacciolepis striata	SSTR	*Sacciolepis clade*	*Paniceae*
Homopholis proluta	HPRO	*Homopholis clade*	*Paniceae*
Alloteropsis semialata	ASEM	*Boivinellinae*	*Paniceae*
Panicum pygmaeum	PPYG	*Boivinellinae*	*Paniceae*
Echinochloa stagnina	ESTA	*Boivinellinae*	*Paniceae*
Lasiacis sorghoidea	LSOR	*Boivinellinae*	*Paniceae*
*Sorghum bicolor**	Sbicolor	*Sorghinae*	*Andropogoneae*
*Zea mays**	Zmays	*Tripsacinae*	*Andropogoneae*
Paspalum fimbriatum	PFIM	*Paspalinae*	*Paspaleae*
Hymenachne amplixecaulis	HAMP	*Otachyriinae*	*Paspaleae*
Steinchisma ssp	OSSP	*Otachyriinae*	*Paspaleae*
Chasmanthium latifolium	CLAT	-	*Chasmanthieae*

* coding sequences extracted from complete genome.

Prior to any phylogenetic analysis, orthology was established among the CDS from the 16 species. The coding sequences were first translated into proteins and clustered into groups of homologs via reciprocal blast searches as implemented in OrthoFinder v1.0.7 (Emms & Kelly 2015). The protein sequences of each group of homologs were then aligned using MAFFT v7.130b (Katoh & Standley 2016), and the

alignments were then translated back into nucleotides. Gene trees were then generated using FastTree v2.1.8 (Price et al. 2010), with parameters suggested by the OrthoFinder authors (Emms & Kelly 2015). One sequence was kept per monophyletic group of sequences belonging to the same species, to remove alleles, splice variants, and duplicates specific to that species. One-to-one orthologs were then identified from the trees as monophyletic groups including each species once only, keeping groups of orthologs including sequences from all 16 taxa for downstream analyses. For each group of one-to-one orthologs, a new nucleotide alignment was obtained using the more stringent method described by Dunning et al. (2017). In short, nucleotide sequences were translated into proteins and aligned with different programs, retaining only residues that were consistently aligned with high confidence. The sequences were then translated back to nucleotides. The resulting alignments were finally trimmed with Gblocks v.0.91 (parameters: -t=c -b2=b1 -b5=h; Castresana, 2000), and alignments longer than 100bp after trimming were used in further analyses. Our approach therefore produced high-quality alignments for a large number of one-to-one orthologs for Panicoideae.

Different approaches to generate the species trees

Four different methods were used to produce a species tree based on all groups of one-to-one orthologs assuming that the individual gene trees follow a common topology. These methods present different ways of summarizing multiple genetic markers, either by combining the alignments before the phylogenetic inference (first apporach) or by summarizing trees inferred individually for each marker (other three methods).

First, groups of one-to-one orthologs were concatenated into a single alignment, and invariant sites and gaps were removed. The concatenated alignment was then used to infer a maximum likelihood phylogenetic using PhyML (Guindon and Gascuel 2003) under the GTR+G +I substitution model, which was the best-fit substitution model, as determined through hierarchical likelihood ratio tests.

Second, individual gene trees were obtained using Bayesian inference, and the consensus of the individual topologies was subsequently obtained, summarizing the gene tree topologies rather than concatenating the markers. For each orthologs,

phylogenetic analyses were performed using MrBayes 3.2.2 (Ronquist and Huelsenbeck 2003) with a GTR+G+I model. Three independent analyses, each composed of three chains, were run per gene for 1,000,000 generations, sampling a tree every 200 generations after a burn-in period of 200,000. Convergence of runs was monitored and analyses were deemed successful when the standard deviation of split frequencies dropped below 0.05 before the end of the burn-in period. Genes were discarded if the analyses failed to converge during the burn-in period. A consensus tree was then obtained for each marker using all sampled trees. These trees were rooted and summarized into a single consensus multigene topology using the R package phangorn v2.2.0 (Schliep, 2011) implemented in R v3.4.1 (R Core Team, 2017).

Third, the individual consensus trees per gene were combined in a DensiTree, as implemented in the R package phangorn v2.2.0 (Schliep, 2011). This was first performed with a randomly selected subset of 510 gene trees, which is the maximum number of trees that can be plotted at once, and then repeated using only those topologies for genes with alignments scores above 92% pairwise identity and length above 1,500bp

Fourth, a topological species tree was inferred using the concordance factors (CF) among gene trees. The CFs were estimated using Bayesian concordance analyses (Ane et al, 2007), which consider the frequency of each of the three possible clades (quartets) of four-taxa sets among all trees sampled post burn-in during the Bayesian tree searches. For each set of four taxa, the CF was calculated using BUCKy 1.4.4 (Larget et al, 2010), with one million post burn-in generations and alpha=1 assuming that gene trees are independent (default parameters in the TICR *Tree Incongruence Checking in R pipleine*.; Stenz et al, 2015). The quartets with the highest CF were then extracted and used by Quartet MaxCut (Snir & Rao, 2012) to generate an amalgamated species tree. Tree sampled during the Bayesian analyses of the different orthologs were all used to infer support of the topology.

Test for reticulate evolution

Because incomplete lineage sorting (ILS) can generate discordance among gene trees, we used the TICR method (Stenz et al, 2015) to test for significant deviation from the equal frequency of minor quartets expected under a pure ILS model, independently for

each four-taxa set. An excess of four-taxa sets with an over-representation of one of the minor quartets will reject ILS as the sole explanation for discordance among gene trees. This approach is conceptually similar to the ABBA-BABA test (Green et al. 2010; Durand et al. 2011).

Since the tree with ILS was rejected (see Results) and because reticulate evolution can generate strong discrepancies among individual gene trees, we specifically tested for hybridization events during the history of Panicoideae. This was done using SNaQ (Solis-Lemus & Ane 2016) as implemented in Julia v.0.6.0 (Bezanson et al. 2017) in the package PhyloNetworks v0.6.0 (Solis-Lemus et al. 2017). This program estimates unrooted networks with a maximum pseudo-likelihood approach using the multi-species coalescent model from multi-locus data. The networks created by SNaQ include estimates of inheritance probabilities *(γ)*, which is the proportion of genes that were contributed by each parental lineage during a hybridization event. Branch lengths are given in the networks in coalescent units, which reflect population sizes, ploidy levels and/or generation times in addition to divergence times. A Table IV.of quartet CFs and their estimation errors was produced by BUCKy and used in SnaQ. The best-fit network for each number of possible hybridization events was found using a heuristic approach and a maximum pseudolikelihood criterion. The topology corresponding to the Quartet MaxCut analysis was used as the starting topology for estimates assuming zero hybridizations (hmax=0). The best fit network of ten SnaQ independent runs was then used as the starting topology for the next ten SNaQ runs with hmax+1, until five hybridization events. The optimal number of hybridization events was obtained as the maximal value still improving fit improvements. Best-fit networks were plotted using Dendroscope v3.5.9 (Huson & Celine, 2012). To estimate confidence intervals, the analyses were repeated 100 times using quartet CF values randomly sampled from their respective confidence intervals, starting with the best-fit network at optimal Hmax-1 as the starting topology in the 80% of the bootstrap analyses and Hmax-2 in the other 20%. The major tree representing the signal of the highest number of genes was obtained by removing minor hybrid edges with $\gamma<0.5$, and its support values were calculated. All minor branches were subsequently summarized and the times each species appears as a descendent of a hybridization events was counted.

Results

One-to-one orthologs and species trees

A total of 2,866 groups of Panicoideae one-to-one orthologs were represented by all 16 species. The total alignment was 3.6 Mbp long after trimming, with an average pairwise identity across the alignment of 91.2% (SD = 2.4). The average length of alignments for individual genes was 1,239.6 bp (SD = 662.2). After removing invariants and sites with gaps, the concatenated alignment was 636 Kbp long.

All nodes in the maximum likelihood tree inferred from the concatenated alignment were highly supported (Fig. IV.1), and the relationships were congruent with those inferred with a different set of nuclear markers (Moreno-Villena et al. 2018). However, discrepancies were observed with Paniceae between the new tree and previously inferred relationships based on plastid markers (Fig. IV.1; GPWG II). Specifically, for the placement of *Digitaria ciliaris*, *Homopholis proluta*, *Echinochloa stagnina*, and *Panicum pygmaeum* differed.

The phylogenetic tree summarizing individual gene trees is poorly resolved (Fig. IV.2a). While the monophyly of the three tribes within Panicoideae is supported by the consensus of the individual gene trees, the relationships within Paniceae were largely unresolved, with a large polytomy. The only exception in the congeners *S. italica/S. barbata*, which form a highly supported monophyletic group (Fig. IV.2b). The observed patterns were confirmed by juxtaposing the individual topologies with DensiTree (Fig. IV.3; Fig. IV.S1). Again, the three tribes were recovered, but numerous topologies exist within Paniceae, with alternative sister groups existing for most terminal taxa (Fig. IV.3). The phylogenetic tree obtained from the most frequent quartets show lack of support across gene trees for nodes within Paniceae, again with the exception of the two congeners *S. barbata*/*S. italica* (Fig. IV.2a).

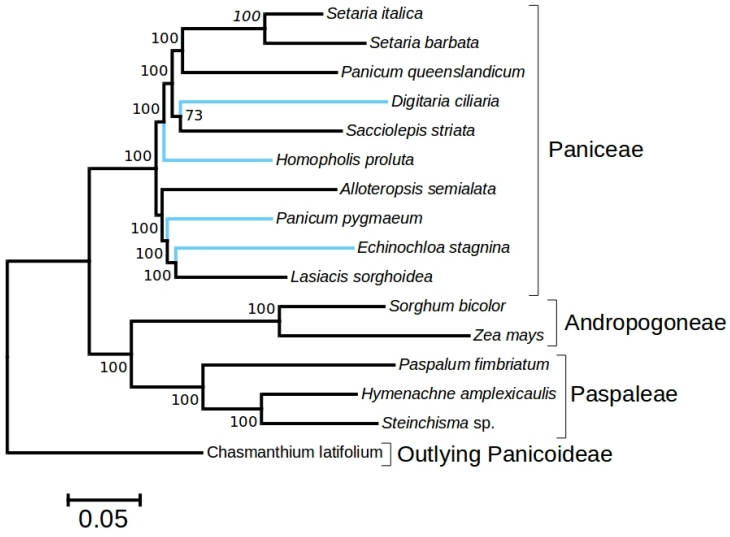


Figure IV.1. Maximum likelihood tree of Panicoideae based on concatenated nuclear markers.
This tree was constructed with PhyML under a GTR+I+G model on the concatenated nucleotide sequences of 2,866 one-to-one orthologs. Bootstrap support values are indicated near branches, and tribes are delimited on the root.

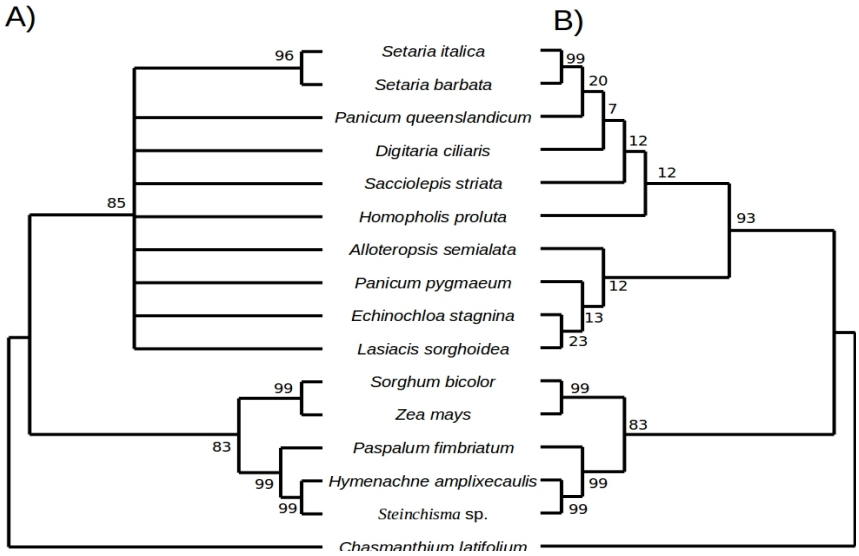


Figure IV.2. Panicoideae trees summarizing individual gene trees.

A) Consensus of individual gene trees. B) Quartet MaxCut tree based on the most frequent quartets for each four-set taxon across gene trees. For both trees, support values indicate the number of genes supporting the grouping among 2,866 nuclear genes.

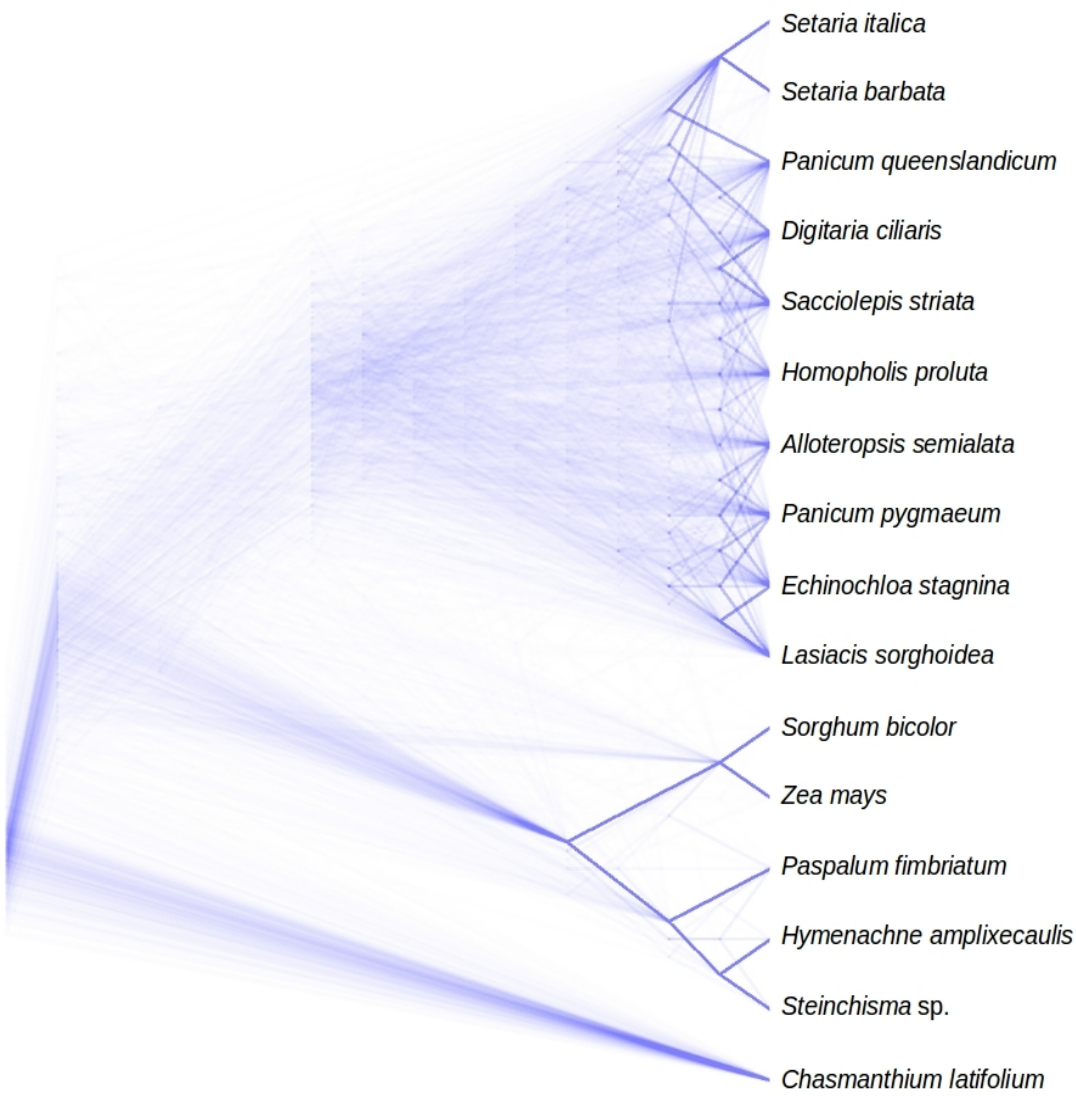


Figure IV.3. Overlap of topologies for individual nuclear markers using DensiTree.
The figure presents the juxtaposition of 510 randomly selected gene trees.

Together, these results indicate that the concatenated alignment results in high support values for most branches, but individual gene trees vary tremendously within Paniceae, so that methods summarizing gene trees result in large polytomies for this group. This could stem from low phylogenetic information in these individual markers, incomplete lineage sorting, patterns of reticulate evolution, or a combination of those.

Tests for Hybridizations

The TICR test identified more four-taxon sets with one minor quartet over-represented than expected under a pure ILS model (p-value < 0.0001; 30 outliers with p-value <0.01; 88 outliers with p-value <0.05). The species most frequently found among these outliers were *Digitaria ciliaris, Panicum quenslandicum, Sacciolepis striata,* and *Lasiascis sorghoidea,* followed by *Alloteropsis semialata, Homopholis proluta, Panicum pygmaeum,* and *Echinochloa stagnina* (Table IV.2). The rest of the species were included in a lower number of outliers, partially due to chance and/or other minor topological shifts.

Table IV.2. Number of outlier from the TICR test in which a species is included.

outliers	DCIL	PQUE	SSTR	LSOR	ASEM	HPRO	PPIG	ESTA	Sitalica	SBAR	CLAT	HAMP	OSSP	PFIM	Sbicolor	Zmays
P<0.01	21	18	9	11	17	5	5	3	6	5	4	4	4	4	2	2
P<0.05	66	50	54	47	37	32	29	26	21	19	16	16	16	15	14	14

To infer putative hybridization events, we used SNaQ (Solis-Lemus & Ane 2016), which estimates species networks based on a multi-species coalescence model. The slope heuristic technique resulted in an optimal network with three hybridizations (hmax=3, -loglik 460.002; Fig. IV.4). The best network with this number of hybridization events retrieved the major grass tribes within Panicoideae (Fig. IV.4). Within Paniceae, a first clade was composed of three taxa traditionally placed within the Boivinellinae subtribe (*P. pygmaeum, L. sorghoideae,* and *E. stagnina*) that appear as the descendant of a hybridization between a relative of the ancestor of A. semialata and a lineage sister to all sampled Paniceae (Fig. IV.4). The same reticulation event could be interpreted as the lineage leading to A. semialata being produced by reticulate evolution between the ancestor of the Boivinellinae and the ancestor of the other taxa within Paniceae (Fig. IV.4), since directionality in networks is uncertain. The second hybridization event suggests contribution from a group sister to the two *Setaria* and another group sister to *Setaria* and *Panicum* to the lineage leading to *D. ciliaris* and *S. striata* (Fig. IV.4; Fig. IV.S2). The alternative topologies for these two putative hybridization events are often found in the replicates based on different quartet CFs (Fig. IV.4). A third, less supported putatived hybridization event involves gene flow from a relative of the common ancestor of *Lasiacis* plus *P. pygmaeum* into the ancestor of *Echinochloa* placed as sister to *Lasiacis* (Fig. IV.4).

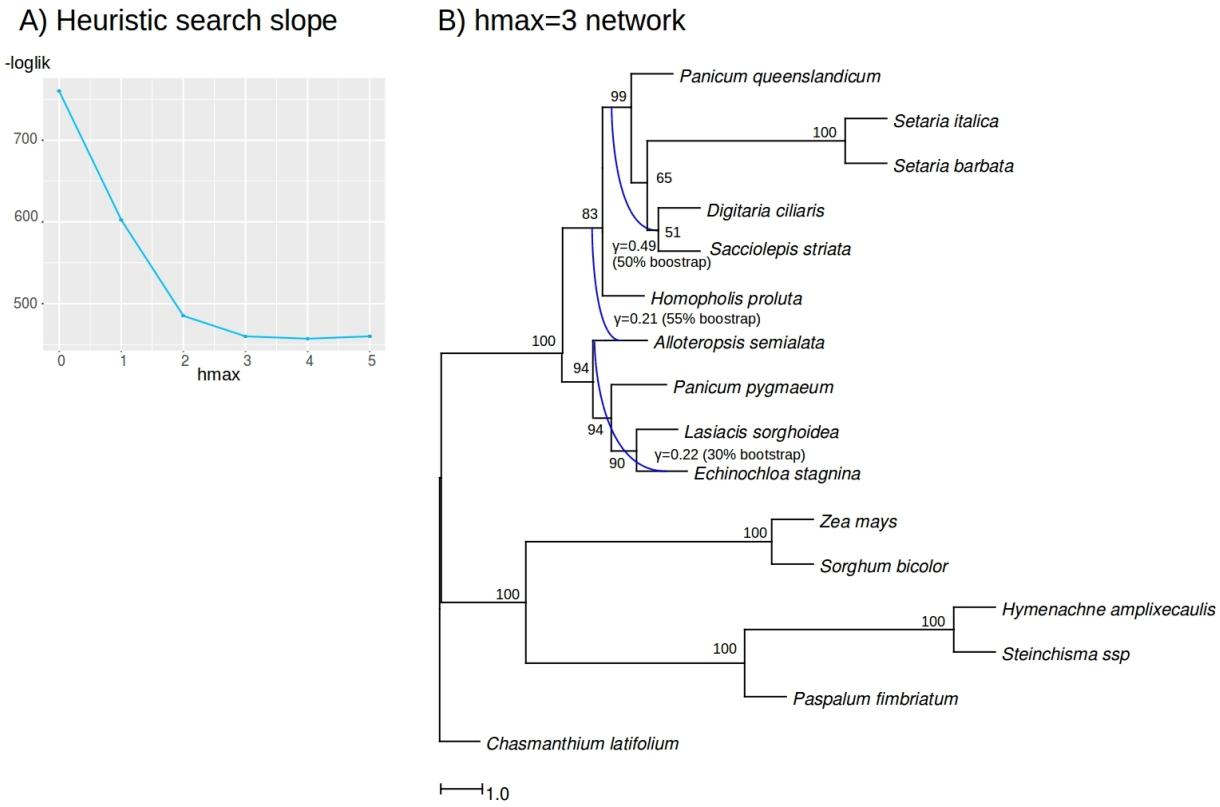


Figure IV.4. Evidence of ancient hybridization events.
A) The values of the pseudo-likelihood are indicated for different numbers of hybridization events. B) For the best-fit model, which includes three hybridization events, the inferred network is indicated. For each reticulation event, the percentage of the pseudoreplicates retrieving the donor branch is indicated is indicated near its base. The fraction of genes contributed by each parent is indicated along the branch.

Discussion

Extensive incomplete lineage sorting

To our knowledge, we present here the first study attempting to infer a species tree from multiple nuclear genes treated as separate markers for the subfamily Panicoideae, an ecologically important group of grasses spread across the tropics and subtropics, which contains multiple crops of economical importance. As previously reported (Moreno-Villena et al. 2018; Washburn et al. 2017), the relationships among taxa are strongly supported when concatenating the different nuclear markers (Fig. IV.1). However, the relationships differ from previous phylogenies inferred from plastid markers (e.g. GPWGII 2012) or individual nuclear markers (e.g. Christin et al. 2007, 2009a, 2009b, 2012). The juxtaposition of individual gene trees reveals strong variation in the branching order within Paniceae (Fig. IV.3), a pattern also seen in the lack of support of

phylogenetic trees summarizing individual gene topologies (Fig. IV.2). This could reflect a lack of resolution in the phylogenetic trees based on single genes, in which short branches and politomies are frequently found between Paniceae taxa. This pattern can also be partially explained by incomplete lineage sorting. Both would be exacerbated by the shortness of the branches near the base of Panicoideae observed on plastid markers (GPWGII 2012), which suggest a rapid early diversification of the group. Concatenating nuclear markers, as performed in Fig. IV.1 and previous studies (Moreno-Villena et al. 2018; Washburn et al. 2017) would solve this issue, as longer alignments would increase phylogenetic informativeness and the joint analyses of multiple markers would decrease problems linked to incomplete lineage sorting. However, the concatenation approach fails to acknowledge the possibility that reticulate evolution events can create strong discrepancies among gene trees, independently of incomplete lineage sorting (Pardo-Diaz et al. 2012; Cui et al. 2013; Dupuis & Sperling 2015). Furthermore, some polymorphisms can have an unbalanced weight over the final tree, introducing phylogenetic bias (Christin et al. 2012b; Shen et al. 2017).

Evidence of reticulate evolution among Paniceae

While incomplete lineage sorting very likely contributes to the discrepancies among topologies, assuming reticulate evolution significantly improves the explanatory power of the model (Fig. IV.4). We therefore conclude that several events of reticulate evolution happened during the evolutionary history of Paniceae. However, the limited species sampling considered here prevents a precise identification of the timing of such events, as well as the exact lineages that were involved. In addition, the inferred networks represent a simplified model that summarizes very complex datasets, and it is possible that multiple events involving small amount of gene flow would be better summarized in the network as fewer events involving larger amounts of gene transfer. We consequently discuss the exact nature of the hybridization events with care, but the data clearly shows that some groups possess genes with different phylogenetic position.

The group receiving the highest support for a putative hybridization event involves *Sacciolepis*, *Panicum*, and *Digitaria*. The optimal network suggests that the common ancestor of *Digitaria* and *Sacciolepis* received genes from a lineage sister to *Panicum* plus *Setaria* and one sister to *Setaria*. While this scenario is probably an

oversimplification, there is evidence in the literature for reticulate evolution involving these taxa. Indeed, plastid markers always support *Panicum* as sister to the Melinidinae (not sampled here) and Cenchrinae (represented here by *Setaria*) subtribes, while *Digitaria* consistently assumes a more basal position (GPWGII 2012; Washburn et al. 2015; Soreng et al. 2017). It was however noted that the position of *Digitaria* changed when nuclear markers were considered (Christin et al. 2009 Plant Physiol; Vicentini et al. 2008), and phylogenetic analyses using individual nuclear genes resulted in different placements of *Panicum* and/or *Digitaria* with respect to Cenchrinae/Melinidinae (e.g. Christin et al. 2007, 2009a, 2009b, 2012; Moreno-Villena et al. 2018). This problem bears consequences for our understanding of C_4 evolution. Indeed, *Digitaria* represents an isolated C_4 group in the plastid lineage, while *Panicum* is part of the same C_4 group as *Setaria*. These observations led to the widespread consideration that they represent two distinct C_4 lineages (Sage et al. 2011; GPWGII 2012). However, phylogenetic analyses of some C_4-specific genes positioned *Panicum* away of *Setaria*, while others placed *Digitaria* next to *Setaria* (e.g. Christin et al. 2007, 2009a, 2009b, 2012; Moreno-Villena et al. 2018). It is therefore possible that ancient reticulate evolution moved some of the C_4 genes contributing to the spread of this complex trait, similar to what have been shown within the grass genera *Alloteropsis* (Dunning et al. 2017) and *Neurachne* (Christin et al. 2012) and the sedge *Eleocharis* (Besnard et al. 2009).

The other inferred hybridization events involve *Alloteropsis* and *Echinochloa*, two C_4 species of the subtribe Boivinellinae. While the support for the group was strong on plastid markers, nuclear markers produced varying phylogenetic relationships (Fig. IV.2; Christin et al. 2012). While the exact order of events is unknown, our analyses bring further support to the idea that recurrent reticulate evolution moved genes among Paniceae lineages during the evolutionary history of this group. However, the mechanisms responsible for these events cannot be known with confidence. One possibility is recurrent allopolyploidization events, a phenomenon that is frequent in plants, and also has been reported for several groups of Panicoideae (Estep et al. 2014). However, the chromosome number is conserved among Paniceae, and there is no evidence of ancient polyploidization within this group of grasses. We therefore suggest that exceptional hybridization and occasional lateral gene transfers, as reported for *Alloteropsis* (Christin et al. 2012), underlie the observed patterns of extensive reticulate

evolution. Associated with extensive incomplete lineage sorting, these explain discrepancies among nuclear markers and between nuclear and plastid genomes.

Conclusions

The Panicoideae subfamily of grasses contains ecologically and economically important species, as well as the highest concentrations of C_4 origins across flowering plants. It has consequently been the subject of numerous and ongoing phylogenetic investigations. While previous evidence suggested discrepancies among different markers within the Paniceae tribe, formal tests of incomplete lineage sorting and reticulate evolution were lacking. Here, we present the first multi-gene phylogenetic tests of these processes for the subfamily. Using different approaches, we show important variation across nuclear genes in the inferred phylogenetic relationships for Paniceae. We show that these can largely be explained by incomplete lineage, which probably results from a rapid diversification at the base of the group. However, assuming reticulate evolution on top of this incomplete lineage sorting significantly improves the fit of the model, and we conclude that hybridization events have punctuated the history of the group. While the limited number of species included in our analyses hampers an exact identification of the timing and direction of these events, they involve several taxa previously thought as belonging to independent C_4 lineages. If the hybridization events involved movements of C_4-related genes, as suggested for individual markers, they could have contributed to the multiplicity of C_4 lineages in this group of grasses. This highlights the importance of reticulate evolution for the adaptive diversification of organisms.

General discussion

This thesis presents the history of genes, focusing on changes in coding sequences and expression patterns, underlying the evolutionary adaptation of grasses, with a special focus on the emergence of C_4 plants from C_3 ancestors. My approach extended on analyses based on species trees only, and instead investigated evolutionary trajectories of C_4 biochemistry after decomposing the pathway into its constituent enzymes and their transcriptional modifications studied at different evolutionary time scales. This was accomplished via high-throughput sequencing coupled with phylogenomic tools, which allowed investigation of transcriptome-wide sequences and expression information.

Comparing transcriptomes from several C_4 origins across the grass family and their C_3 relatives allowed me to infer the ancestral transcriptional states and test their effects on gene co-option (Chapter I). This first systematic test of the factors dictating C_4 gene co-option showed that the most highly expressed genes were preferentially co-opted for a C_4 function, independent of their tissue specificity. My work also showed that massive changes in both expression patterns and coding sequences happened post co-option, including several levels of gene upregulation and an unprecedented degree of parallel adaptive amino acid substitutions across multiple genes (Chapter I).

The work presented in Chapter I sheds new light on the factors that dictate gene co-option for C_4 photosynthesis, and therefore some of the properties that might increase C_4 evolvability in some clades. However, comparisons among distantly related C_3 and C_4 plants revealed abundant differences, which are not necessarily representative of the changes that allowed photosynthetic transitions. Indeed, these differences might also include changes that happened after the phenotypic transitions or those that were not directly linked to the phenotypic state. To circumvent this problem, I decided to compare the transcriptomes of individuals representing multiple photosynthetic types, but belonging to the same species (Chapter II). This first intraspecific C_3/C_4 transcriptome comparison showed that C_4 enzymes have been sequentially co-opted, and that the initial transition to new photosynthetic types involved changes in very few genes. Evidence of variation among C_4 populations suggests that the high level of specialization associated with most C_4 plants evolves after the initial photosynthetic transitions (Chapter II).

Because reconstructing past transitions can be complicated when relying solely on species trees, we decided to use analyses of individual genes to reconstruct the order of phenotypic changes within the genus *Alloteropsis* (Chapter III). This work combined anatomical data provided and analysed by Dr Marjorie Lundgren, with transcriptome data co-analysed with Dr Luke Dunning. Through comparative analyses of expression levels and selective pressures on coding sequences, we were able to reveal multiple C_4 origins that recurrently used some C_4-like components present in their common ancestor (Chapter III). In addition, we showed that some adaptive loci had been transferred across species, leading to the lateral spread of C_4 adaptations, and disconnecting the number of origins of the C_4 phenotype and those of the underlying genes (Chapter III).

My transcriptome comparisons, presented in Chapters I, II, and III, showed notable levels of up-regulation in C_4-related genes of C_4 taxa, together with a down-regulation of photorespiratory genes. Unexpectedly, phylogenetic analyses also repeatedly showed topological discordance among gene trees. Therefore, I decided to use the generated sequence data to specifically test for reticulate evolution during the history of Panicoid grasses, the group of grasses with the highest number of C_4 origins (Chapter IV). Using different methods, I obtained evidence of incomplete lineage sorting, but also evidence of reticulate evolution during the diversification of the group. Because some of the hybridization events involved C_4 groups, I suggest that reticulate evolution might have globally played a role in the origins of C_4 plants (Chapter IV).

Overall, the results obtained in this thesis offer new insights into the evolutionary trajectories that led to C_4 origins, and provide some insight into the molecular origins of complex biochemical traits and the genetic mechanisms driving evolutionary innovations. In the following sections, I will discuss these points, going beyond what is discussed for each individual chapter.

Evolutionary trajectories to C_4 Photosynthesis

From a biochemical pathway point of view, the evolution of C_4 photosynthesis involves the co-option of multiple genes, which existed in non-C_4 ancestors, but were responsible for different, non-photosynthetic functions (Monson 2003; Aubry et al. 2011). The genomic factors enabling C_4 evolution have been recurrently discussed, with a special focus on the importance of gene duplication and neo-functionalization (Sage 2001;

Monson 2003). The accumulation of gene and genome data show that the origins of C_4-specific enzymes were not consistently directly preceded by gene duplications, and the number of C_4-related genes does not differ between C_4 and non-C_4 genomes (Christin et al. 2007, 2009; Besnard et al. 2009; Wang et al. 2009; Gowik et al. 2011 ; Williams et al. 2013). However, C_4-related enzymes are encoded by multigene families, with multiple lineages generated via multiple rounds of gene and/or genome duplications during the history of plants (De Bodt et al. 2005; Cui et al. 2006; Jiao et al. 2014). It is possible that the accumulation of a reservoir of duplicated genes eased the subsequent evolution of C_4 photosynthesis (Sage 2004). First, duplicates might generate genetic redundancy, enabling later neofunctionalization even among distant duplicates. Second, the multiple gene lineages evolve independently, accumulating distinct expression properties and catalytic characteristics of the encoded enzymes (Chapter I). Having multiple duplicates might therefore increase the chance of some of them reaching a state that is favourable to a function in C_4 photosynthesis. However, the nature of a favourable state remains speculative. Comparative analyses have shown that some gene lineages were co-opted more often than others, suggesting that these genes were better suited for the C_4 function, without knowing which aspect made them better suited (Christin et al. 2010, 2013, 2015). It has been noted that the genes most highly expressed in some C_3 species were the ones generally co-opted when a few C_4 species were considered (Christin et al. 2013; Emms et al. 2016). Using a wide sample of multiple C_4 origins and their C_3 relatives, I showed that the genes that are most abundant in the leaves of non-C_4 ancestors were preferentially co-opted for C_4 photosynthesis, independently of their tissue specificity (Chapter I). Other properties might also help to determine which genes are co-opted, including the catalytic properties of the encoded enzyme, but these remain unknown. However, the fact that the most highly expressed genes are preferentially co-opted sheds new light into the early origins of the C_4 cycle. We can propose that some genes reach high leaf transcription abundance for a variety of reasons unrelated to C_4 photosynthesis, including selection for other functions or chance. Once the C_4-related genes are abundant in the leaves of non-C_4 plants that possess anatomical traits compatible with a C_4 pathway (Christin et al. 2013; Lundgren et al. 2014), a weak, rudimentary C_4 cycle might emerge through relatively few changes (Chapter II). The emergence of such a rudimentary C_4 cycle is extremely important in

evolutionary terms, because it changes the adaptive landscape, making any mutation increasing the strength of the C_4 cycle highly advantageous. Indeed, models suggest that once a C_4 cycle exists, any increase in its strength will translate into strong fitness advantages (Heckmann et al. 2013). Therefore, once a tipping point is reached where enough C_4-related enzymes are present to sustain a weak C_4 pathway, plants might be on an inevitable evolutionary highway to a full C_4 state. This might be helped by the fact that C_4 genes are upregulated to rebalance nitrogen in C_3-C_4 intermediates (Mallmann et al. 2014) or when CO_2 levels are low (Li et al. 2014), which both represent accepted preconditions to C_4 evolution (Sage 2001).

Once sufficient C_4-related enzymes are present in the leaves, a weak C_4 cycle might emerge through the upregulation of a few key genes, as shown in *Alloteropsis semialata* (Chapter II). Of course, additional modifications to leaf properties might be required, including modifications in the cells or organelles. However, plants with a weak C_4 cycle ('C_3+C_4 plants'; Chapters II and III) have similar photorespiration activity and leaf anatomy to their close C_3 relatives (Morgan & Brown1979; Morgan et al. 1980; Brown et al. 1983; McKown and Dengler 2007; Lundgren et al. 2016). Moreover, the transition from a C_3+C_4 state to a full C_4 state was also shown to require the upregulation of very few genes in *A. semialata* (Chapter II). As before, additional changes to leaf anatomy might be required, but the transition from a C_3+C_4 to a C_4 state might still represent a relatively small step. However, the C_4 *A. semialat*a plants studied lack most of the transporters and regulators classically associated with the C_4 type (Chapter II), and exhibit only partial C_4 traits, such as an incomplete segregation of photosynthetic reactions among mesophyll and bundle sheath cells (Ueno and Sentoku 2006). We propose that most properties observed in traditional C_4 systems emerge later, once plants are already in a C_4 state. Indeed, some C_4-related genes were associated with only some populations of *A. semialata*, suggesting that these emerged after the initial evolution of a C_4 type (Chapters II and III).

Because gene expression patterns enabling C_4 evolution exist in some groups (Chapter I) and because different photosynthetic stages can then be reached via few genetic changes (Chapter II), the evolution of C_4 photosynthesis might be relatively easy for some plants. Coupled with anatomical enablers characterizing some phylogenetic lineages (Christin et al. 2013), the findings of this thesis help to explain the impressive

number of origins of this seemingly complex trait. Indeed, if most of the changes associated with the C_4 pathway occur once the plants are already C_4, the initial transitions might be easy, at least for groups possessing the required preconditions. And we have shown that a set of predispositions, such as higher gene expression, can be recurrently co-opted to evolve C_4 by multiple descendants (Chapter III), a phenomenon also suggested for other groups (Christin et al. 2011). This would contribute to the observed bursts of C_4 origins in some lineages, which would be further exacerbated by the lateral spread of C_4-adaptive loci (Chapter III).

Revisiting the C_4 adaptive landscape

The evolution of C_4 photosynthesis has been recently discussed in the context of adaptive landscapes. Indeed, Heckmann et al. (2013, 2015) modelled the fitness landscape connecting C_3 and C_4 states. They suggest that C_4 photosynthesis is accessible through a series of successive physiological intermediates, along a steep hill where each mutation increases fitness. Using a different approach, Williams et al (2013) compared the phenotypes of tens of species representing different types of C_3-C_4 intermediates to conclude that the evolution of C_4 photosynthesis can follow a number of different and flexible evolutionary trajectories, which are determined by the traits present in the ancestor. They also conclude that this flexibility of trajectories likely facilitated the repeated evolution of C_4 photosynthesis (Williams et al. 2014).

These visions can be revisited and reconciled based on the results obtained in my work. First, we do suggest that the initial changes that enable the subsequent transitions to C_4 happened in a flat adaptive landscape. Indeed, the high transcript abundance that then facilitates gene co-option (Chapter I) likely evolved for a variety of reasons, none of which may be related to the C_4 trait, since it would not yet exist. However, once a tipping point is reached, a weak C_4 cycle might emerge. This event, while negligible in physiological terms, would dramatically alter the adaptive landscape. Indeed, adaptive landscapes are not fixed, but are condition dependant. Putting ecological variation aside, we can assume that these events happened in an environment that promoted photorespiration. In this condition, once a weak C_4 cycle exists, any mutation that improves its strength will be selected, and plants will be moved toward adaptive peaks corresponding to the C_4 state, following the model of Heckmann et al. (2013). However,

because the adaptive landscape changes once a weak C_4 cycle emerges, it will depend on the ancestral condition, as suggested by Williams et al. (2013). Since ancestral gene abundance dictates gene co-option (Chapter I), the trajectory will depend on the ancestral transcriptome composition, and some C_4 subtypes might never evolve in groups where the corresponding genes are not ancestrally expressed. In addition, each step will have successfully affected the subsequent adaptive landscape, effectively leading to a number of possible trajectories that reflect evolutionary constraints and opportunities.

Genomic factors promoting functional diversification

Adaptive changes represent responses to selective pressures that increase the frequency of existing or novel mutations. By definition, novel adaptations requiring multiple changes must therefore evolve in a stepwise manner, where individual mutations are either neutral and fixed by chance or slightly beneficial. The genomic material at disposition will therefore affect the likelihood of evolving novel traits. Genomes tend to be robust against variability, as most deleterious mutations are deleted by selection (Sawyer et al. 2007; reviewed in Barrick & Lenski 2013). This robustness however allows the accumulation of neutral mutations that do not affect the phenotype (de Visser et al. 2003) and can therefore represent cryptic changes enabling later modifications with an effect on the phenotype. For instance, in a long experimental evolution trial with *E. coli*, mutations without any apparent effect have been shown to be needed for subsequent adaptive metabolic transitions (Blount et al. 2012). In addition, mutations might be non neutral, but beneficial for a variety of unrelated pathways. Such mutations might create a number of elements that can later be co-opted to evolve a different, novel pathway. This process of exaptation apparently explains C_4 origins, since evolving high leaf abundance of multiple enzymes likely eased transitions to a C_4 cycle (Chapters I and II).

The process of exaptation means that random changes and contingency likely affect future evolutionary trajectories (Barve & Wagner 2013). However, there might be general rules that dictate evolvability among genomes. First, possessing a variety of biochemical pathways would increase the likelihood of possessing a set of enzymes suitable for a different, novel pathway. In addition, genetic redundancy might relax

purifying selection and therefore enable neofunctionalization (Conant & Wolfe 2008). Both of these are likely to be increased by frequent gene duplications, whether these concern single gene, whole chromosomal segments, or complete genomes. In plants, whole genome duplication is frequent, and has punctuated the history of land plants (Wendel 2015; Panchy et al. 2016). Interestingly, it has been proposed that these events do not immediately increase diversification. Instead, there is a time lag, before the beneficial effects of whole genome duplication are observed (Tank et al. 2015; Soltis et al. 2015). This view is compatible with the idea that constituting a large reservoir of similar genes favours functional diversification and therefore success. Indeed, multiple gene copies with slightly different properties increase genetic redundancy, but also the diversity of genes and enzymes available for novel pathways (e.g. Kassahn et al. 2009). I propose that this reservoir of genes facilitated C_4 evolution in flowering plants, as suggested previously (Sage 2001; Monson 2003). However, the impact of gene duplication is probably indirect, creating a diversity of similar enzymes instead of the classical genetic redundancy. Therefore, the changes that happened on each copy after the duplication will likely have an effect, highlighting the importance of contingency in evolution.

Because C_4 photosynthesis evolved multiple times independently, my research sheds new light on the repeatability of evolution. Among the species studied, the same genes tend to be co-opted more often than expected by chance, so that gene co-option is to some extent repeatable (Chapter I; Christin et al. 2013, 2015). This likely explains the fact that the non-C_4 ancestors of each C_4 group within the same family tend to be similar due to shared ancestry. Because they do express the same genes in their leaves, these genes then get recurrently co-opted (Chapter I). Therefore, the repeatability of evolution in this case is explained by contingency. This view is reinforced by the fact that distantly related C_4 origins co-opted different genes, so that the ancestral state dictates subsequent evolutionary trajectories (Chapter I; Christin et al. 2015). Besides this biased gene co-option, sequence comparisons have shown recurrent adaptive amino acid changes (Chapters I and III; Christin et al. 2007, 2009). Again, this likely reflects a combination of constraints driven by the ancestral state and a limited number of evolutionary answers to the same problem. Because the same genes tend to be co-opted, enzyme adaptation acts on similar coding sequences. The recurrent amino acid changes

being fixed probably stem from the fact that most mutations will be deleterious, and only a few confer beneficial kinetic changes while maintaining the overall stability and functionality of the enzyme (Studer et al. 2014).

Reticulate evolution and the spread of adaptations

My work has contributed to the accumulating evidence that gene flow happens among closely, and even distantly, related species (Chapters III and IV). Importantly, some of the transferred genes contributed to the adaptation of the C_4 pathway in the studied species (Chapter III). This suggests that introgression of C_4 loci can allow the lateral spread of C_4 adaptations, a phenomenon also reported in other systems, such as the weedy sunflower *Helianthus* (Whitney et al. 2006), mimetic *Heliconius* butterflies (Pardo-Diaz et al. 2012), and rodents (Song et al. 2011). In the case of C_4 genes, the coding sequences are adapted for the C_4 context during multiple rounds of adaptive amino acid changes (Chapters I and III). Receiving genes that have already undergone such changes therefore represents an evolutionary shortcut, allowing the transition to more efficient C_4 types without needing long periods of selection on random changes. In addition, the transfer of C_4 genes might in some cases spread a completely novel component. For instance, the enzyme phosphoenolpyruvate carboxykinase (PCK) was not used by members of *Alloteropsis* until a gene was laterally transferred from a distantly related C_4 group (Chapter III). In this case, the new gene likely provided a novel function, representing a gene with both regulatory and coding sequences already adapted for the C_4 context.

Because the many genes that contribute to the complex C_4 trait lie on different chromosomes, they are unlikely to be simultaneously introgressed. The C_4 trait can therefore not be transferred in one step. However, it is possible that genes that allow key transitions are introgressed, and then place an unrelated different lineage on an evolutionary trajectory toward C_4, after a modification of the adaptive landscape. Indeed, only a few changes are needed to reach a tipping point where a weak C_4 cycle is triggered (Chapter III). If these genes are introgressed, the whole pathway would then follow, leading to the partial spread of the adaptation. It is however difficult to know whether the genes are transferred to non-C_4 or C_4 lineages. This hypothesis could therefore be tested using experiments. Past attempts to integrate C_4 genes into C_3 plants

have never triggered a C_4 pathway (reviewed in Schuler et al. 2016), however the targeted C_3 plants did likely not possess all of the required predispositions. One possible experiment could target C_3 plants that are closely related to C_4 lineages, such as the C_3 populations of *Alloteropsis semialata* (Chapters II and III). The genes that we have identified to be linked to the emergence of a C_3+C_4 type might be engineered in these plants, and the consequences of the transfer might be monitored to determine whether the complete pathway can then be rapidly triggered.

My analyses have already identified multiple potential episodes of reticulate evolution in *Alloteropsis* and other Panicoideae (Chapters III and IV). Future investigations might determine whether these included transfers of adaptations other than C_4 components. Indeed, a number of adaptive traits are present among unrelated groups of Panicoideae, such as cold tolerance (Humphreys & Linder 2013). While convergent evolution should be the primary hypothesis, we cannot exclude the possibility that gene exchanges lead, directly or indirectly, to the adaptive diversification across large taxonomic groups.

General Conclusions

Determining how organisms diversify and acquire new adaptations is of paramount importance to understand why they are different and how they cope with changing environments. The vast amount of genomic, phenotypic, and population dynamic data being constantly produced, together with expanding computer resources and analytical tools, are slowly revealing the complexity and intricacy of evolutionary process. Early evolutionary biologists would have dreamt to see this level of resolution realized one day. However, the scientific community is still very far from understanding the whole spectrum of evolutionary processes. Thanks to fifty years of investigations, the biochemistry, physiology, and genomes properties associated to C_4 photosynthesis are relatively well understood. It was therefore possible in this dissertation to decompose the complex trait into its genetic components, and analyse each of them separately to track the history of changes underlying major phenotypic transitions. By doing so, my work contributed to a better resolution of the evolutionary trajectories to C_4 photosynthesis. However, my investigations were limited to coding sequences and their expression patterns. Other genomic features might affect the evolutionary trajectories to

C_4, including genomic architecture, chromosomal localization of C_4-related genes, methylation patterns, and distribution of transposable elements with respect to the C_4-related genes. The impact of these factors on C_4 evolvability should be studied once genomes, methylomes, and other descriptions of the genomes (e.g. 3D structures) become available for a large number of C_3 and C_4 species.

The comparative transcriptomics approach adopted throughout my work allowed identifying the changes in gene expression and coding sequences responsible for the evolution of C_4 photosynthesis in grasses. In particular, my research showed that the early emergence of C_4 photosynthesis was triggered by changes in the expression patterns of few genes without adaptation of the enzymes (Chapter II), which was facilitated by the existence of transcriptome preadaptations in some taxonomic groups (Chapter I). Because important changes in expression patterns and coding sequences occurred after this initial emergence (Chapters I and III), the repeated origins of C_4 photosynthesis remain an exceptional example of parallel transcriptional changes and parallel enzyme adaptation across many genes and many species. My work therefore explains how C_4 photosynthesis could have evolved so easily in some groups, but also reconciles this novel view with previous reports highlighting the complexity of the C_4 trait compared to the ancestral C_3 state. This leads to a counter-intuitive view that the C_4 trait was easy to evolve, yet required massive changes at multiple levels. The repeated origins of C_4 photosynthesis in some groups were likely further fuelled by the transfer of C_4-adaptive loci across species boundaries, as reported in Chapter III and to some extent in Chapter IV. Therefore, my dissertations contributes to understanding both the early events leading to a rudimentary C_4 system and the changes that followed, providing an improved picture of the evolutionary trajectories to C_4 photosynthesis across a complex adaptive landscape.

Overall, I conclude that the evolvability of C_4 photosynthesis was determined at multiple levels; taxonomic groups, ecological conditions, and life-history traits. Random processes and/or selection for other purposes provided some clades with C_4 preadaptations, including leaf anatomical properties (Christin et al. 2013; Lundgren et al. 2014) and abundance of transcripts for C_4 enzymes in leaves (Chapter I). Therefore, the probability of evolving C_4 varied among taxonomic groups, explaining the phylogenetic clustering of C_4 origins (Sage et al. 2011). After atmospheric conditions

changed and CO_2 decreased in the early Oligocene, having a C_4 pathway became advantageous in some environments promoting photorespiration. When minor changes of time pushing the biochemistry of the leaf to a state where a weak C_4 cycle emerges (Chapter II), the new C_4 trait would have been selected for only in those plants inhabiting such environments, with subsequent selection for a strengthened C_4 cycle. Ecology therefore determined which plant lineages could evolve C_4, explaining the predominance of C_4 origins in warm regions (Edwards and Smith 2010). The efficiency of selection will vary among species, especially as a function of population sizes, and highly compartmentalized metapopulations facilitated the fixation of C_4 mutations (Olofsson et al. 2016). In addition, while these remain unidentified, life-history traits promoting gene exchanges across species boundaries will increase the chance of introgression of C_4-adaptive loci (Chapters III and IV). The demographics and life-history traits will therefore further contribute to determining the probability of making the initial transition to a rudimentary C_4 physiology, and later continuously adapting it. These different evolutionary processes interacted to lead to the recurrent origins of the major ecological innovation represented by C_4 photosynthesis.

References

Agnarsson, I. (2004). Morphological phylogeny of cobweb spiders and their relatives (Araneae, Araneoidea, Theridiidae). *Zoological Journal of the Linnean Society*, *141*(4), 447–626.

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., & Tawfik, D. S. (2005). The “evolvability” of promiscuous protein functions. *Nature Genetics*, *37*(1), 73–6.

Albert, V. A., Oppenheimer, D. G., & Lindqvist, C. (2002). Pleiotropy, redundancy and the evolution of flowers. *Trends in Plant Science*, *7*(7), 297–301.

Altenhoff, A. M., & Dessimoz, C. (2012). Inferring Orthology and Paralogy. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 855, pp. 259–279).

Altenhoff, A. M., Gil, M., Gonnet, G. H., Dessimoz, C., Mitros, T., Weinmaier, T., … Peng, J. (2013). Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. *PLoS ONE*, *8*(1), e53786.

Andersson, J. O. (2005). Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences*, *62*(11), 1182–1197.

Artyukhin, E. N. (2006). Morphological Phylogeny of the Order Acipenseriformes. *Journal of Applied Ichthyology*, *22*(s1), 66–69.

Atkinson, R. R. L., Mockford, E. J., Bennett, C., Christin, P.-A., Spriggs, E. L., Freckleton, R. P., … Osborne, C. P. (2016). C_4 photosynthesis boosts growth by altering physiology, allocation and size. *Nature Plants*, *2*(5), 16038.

Aubry, S., Brown, N. J., & Hibberd, J. M. (2011). The role of proteins in C_3 plants prior to their recruitment into the C_4 pathway. *Journal of Experimental Botany*, *62*(9), 3049–3059.

Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D., & Hibberd, J. M. (2014). Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C_4 photosynthesis. *PLoS Genetics*, *10*(6), e1004365.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., … Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, *444*(7116), 171–178.

Barrett, R., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*(1), 38–44.

Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, *14*(12), 827–839.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., … Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, *461*(7268), 1243–1247.

Barve, A., & Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, *500*(7461), 203–206.

Beaulieu, J. M., O'Meara, B. C., & Donoghue, M. J. (2013). Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Systematic Biology*, *62*(5), 725–37.

Bellasio, C., Burgess, S. J., Griffiths, H., & Hibberd, J. M. (2014). A high throughput gas exchange screen for determining rates of photorespiration or regulation of C_4 activity. *Journal of Experimental Botany*, *65*(13), 3769–3779.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. In *Plant Molecular Evolution* (pp. 251–269). Dordrecht: Springer Netherlands.

Bergthorsson, U., Andersson, D. I., & Roth, J. R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(43), 17004–9.

Besnard, G., Muasya, A. M., Russier, F., Roalson, E. H., Salamin, N., & Christin, P.-A. (2009). Phylogenomics of C_4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Molecular Biology and Evolution*, *26*(8), 1909–1919.

Bielawski, J. P., Baker, J. L., Mingrone, J., Bielawski, J. P., Baker, J. L., & Mingrone, J. (2016). Inference of episodic changes in natural selection acting on protein coding sequences via CODEML. In *Current Protocols in Bioinformatics* (p. 6.15.1-6.15.32). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Blasing, O. E., Westhoff, P., & Svensson, P. (2000). Evolution of C_4 phosphoenolpyruvate carboxylase in *Flaveria* conserved serine residue in the carboxyterminal part of the enzyme is a major determinant for C_4-specific characteristics. *Journal of Biological Chemistry*, *275*(36), 27917–23.

Blount, Z. D., Barrick, J. E., Davidson, C. J., & Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*, *489*(7417), 513–518.

Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*.

Proceedings of the National Academy of Sciences of the United States of America, *105*(23), 7899–906.

Bock, W. J. (1959). Preadaptation and multiple evolutionary pathways. *Evolution*, *13*(2), 194–211.

Bohley, K., Joos, O., Hartmann, H., Sage, R. Liede-Shumann, S., Kadereit, G. (2014). Phylogeny of Sesuvioideae (Aizoaceae) – Biogeography, leaf anatomy and the evolution of C_4 photosynthesis. *Perspectives in Plant Ecology, Evolution and Systematics*, 17(2), 116-130.

Bork, P., Sander, C., & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Science : A Publication of the Protein Society*, *2*(1), 31–40.

Bornberg-Bauer, E., & Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(19), 10689–94.

Brown, W. V. (1975) Variations in anatomy, associations, and origins of Kranz tissue. *American Journal of Botany*, 62(4), 395-402

Ellis, R. P. (1974). Anomalous vascular bundle sheath structure in *Alloteropsis semialata* leaf blades. *Bothalia*, 11(3), a1460.

Boudreaux, H. B. (1979). *Arthropod phylogeny, with special reference to insects.* John Wiley & Sons Inc.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, *247*(4948), 1306–10.

Bräutigam, A., & Gowik, U. (2016). Photorespiration connects C_3 and C_4 photosynthesis. *Journal of Experimental Botany*, *67*(10), 2953–2962.

Bräutigam, A., Kajala, K., Wullenweber, J., Sommer, M., Gagneul, D., Weber, K. L., … Weber, A. P. M. (2011). An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. *Plant Physiology*, *155*, 142–156.

Bräutigam, A., Schliesky, S., Külahoglu, C., Osborne, C. P., & Weber, A. P. M. (2014). Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C_4 species. *Journal of Experimental Botany*, *65*(13), 3579–93.

Brischoux, F., & Shine, R. (2011). Morphological adaptations to marine life in snakes. *Journal of Morphology*, *272*(5), 566–572.

Brockington, S. F., Walker, R. H., Glover, B. J., Soltis, P. S., & Soltis, D. E. (2011). Complex pigment evolution in the Caryophyllales. *New Phytologist*, *190*(4), 854–864.

Brockington, S. F., Yang, Y., Gandia-Herrero, F., Covshoff, S., Hibberd, J. M., Sage, R. F., … Smith, S. A. (2015). Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. *New Phytologist*, *207*(4), 1170–1180.

Brown, N. J., Newell, C. A., Stanley, S., Chen, J. E., Perrin, A. J., Kajala, K., & Hibberd, J. M. (2011). Independent and parallel recruitment of preexisting mechanisms underlying C_4 photosynthesis. *Science*, *331*(6023).

Brown, R. H., Bouton, J. H., Rigsby, L., & Rigler, M. (1983). Photosynthesis of grass species differing in carbon dioxide fixation pathways : VIII. Ultrastructural characteristics of Panicum species in the Laxa group. *Plant Physiology*, *71*(2), 425–31.

Brunet, F. G., Roest Crollius, H., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., … Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*, *23*(9), 1808–16.

Brunet, F. G., Lorin, T., Bernard, L., Haftek-Terreau, Z., Galiana, D., Schartl, M., & Volff, J.-N. (2017). Case studies of seven gene families with unusual high retention rate since the vertebrate and teleost whole-genome duplications. In *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts* (pp. 369–396). Cham: Springer International Publishing.

Budde, R. J. A., Holbrook, G. P., & Chollet, R. (1985). Studies on the dark/light regulation of maize leaf pyruvate, orthophosphate dikinase by reversible phosphorylation. *Archives of Biochemistry and Biophysics*, *242*(1), 283–290.

Byrt, C. S., Grof, C. P. L., & Furbank, R. T. (2011). C_4 Plants as biofuel feedstocks: optimising biomass production and feedstock quality from a lignocellulosic perspective. *Journal of Integrative Plant Biology*, *53*(2), 120–135.

Cameron, E. Z., & du Toit, J. T. (2007). Winning by a neck: tall giraffes avoid competing with shorter browsers. *The American Naturalist*, *169*(1), 130–5.

Cantalapiedra, J. L., Prado, J. L., Hernández Fernández, M., & Alberdi, M. T. (2017). Decoupled ecomorphological evolution and diversification in Neogene-Quaternary horses. *Science*, *355*(6325), 627–630.

Cao, X., Sun, Y.-B., Irwin, D. M., Wang, G.-D., & Zhang, Y.-P. (2015). Nocturnal to diurnal transition in the common ancestor of haplorrhines: evidence from genomic-scan for positively selected genes. *Journal of Genetics and Genomics*, *42*(1), 33–37.

Carmo-silva, E., Scales, J. C., Madgwick, P. J., & Parry, M. A. J. (2015). Optimizing Rubisco and its regulation for greater resource use efficiency. *Plant, Cell & Environment*, *38*(9), 1817–1832.

Carroll, S. B. (2008). Evo-Devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, *134*(1), 25–36.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–374.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–52.

Cebra-Thomas, J., Tan, F., Sistla, S., Estes, E., Bender, G., Kim, C., … Gilbert, S. F. (2005). How the turtle forms its shell: a paracrine hypothesis of carapace formation. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *304B*(6), 558–569.

Chang, J.-M., Di Tommaso, P., & Notredame, C. (2014). TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Molecular Biology and Evolution*, *31*(6), 1625–37.

Cheatle Jarvela, A. M., & Hinman, V. F. (2015). Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. *EvoDevo*, *6*(1), 3.

Christensen, C. B., Christensen-Dalsgaard, J., & Madsen, P. T. (2015). Hearing of the African lungfish (*Protopterus annectens*) suggests underwater pressure detection and rudimentary aerial hearing in early tetrapods. *The Journal of Experimental Biology*, *218*(3), 381–7.

Christin, P.-A., Arakaki, M., Osborne, C. P., & Edwards, E. J. (2015). Genetic enablers underlying the clustered evolutionary origins of C_4 photosynthesis in angiosperms. *Molecular Biology and Evolution*, *32*(4), 846–58.

Christin, P.-A., & Osborne, C. P. (2014). The evolutionary ecology of C_4 plants, 765–781.

Christin, P.-A., Boxall, S. F., Gregory, R., Edwards, E. J., Hartwell, J., & Osborne, C. P. (2013). Parallel recruitment of multiple genes into C_4 photosynthesis. *Genome*

Biology and Evolution, *5*(11), 2174–87.

Christin, P.-A., Spriggs, E., Osborne, C. P., Strömberg, C. A. E., Salamin, N., & Edwards, E. J. (2014). Molecular dating, evolutionary rates, and the age of the grasses. *Systematic Biology*, *63*(2), 153–65.

Christin, P.-A., & Osborne, C. P. (2013). The recurrent assembly of C_4 photosynthesis, an evolutionary tale. *Photosynthesis Research*, *117*(1–3), 163–75.

Christin, P.-A., Osborne, C. P., Chatelet, D. S., Columbus, J. T., Besnard, G., Hodkinson, T. R., … Edwards, E. J. (2013). Anatomical enablers and the evolution of C_4 photosynthesis in grasses. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(4), 1381–6.

Christin, P. A., Boxall, S. F., Gregory, R., Edwards, E. J., Hartwell, J., & Osborne, C. P. (2013). Parallel recruitment of multiple genes into C_4 photosynthesis. *Genome Biology and Evolution*, *5*, 2174–2187.

Christin, P.-A., Edwards, E. J., Besnard, G., Boxall, S. F., Gregory, R., Kellogg, E. … Osborne, C. P. (2012). Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer. *Current Biology : CB*, *22*(5), 445–9.

Christin, P.-A., Besnard, G., Edwards, E. J., & Salamin, N. (2012b). Effect of genetic convergence on phylogenetic inference. *Molecular Phylogenetics and Evolution*, *62*(3), 921–927.

Christin, P.-A., Osborne, C. P., Sage, R. F., Arakaki, M., & Edwards, E. J. (2011). C_4 eudicots are not younger than C_4 monocots. *Journal of Experimental Botany*, *62*(9), 3171–3181.

Christin, P.-A., Sage, T. L., Edwards, E. J., Ogburn, R. M., Khoshravesh, R., & Sage, R. F. (2011). Complex evolutionary transitions and the significance of C_3-C_4 intermediate forms of photosynthesis in Molluginaceae. *Evolution; International Journal of Organic Evolution*, *65*(3), 643–60.

Christin, P.-A., Weinreich, D. M., & Besnard, G. (2010). Causes and evolutionary significance of genetic convergence. *Trends in Genetics*, *26*(9), 400–5.

Christin, P. A., Freckleton, R. P., & Osborne, C. P. (2010). Can phylogenetics identify C_4 origins and reversals? *Trends in Ecology and Evolution*, *25*(7), 403–409.

Christin, P.-A., Petitpierre, B., Salamin, N., Büchi, L., & Besnard, G. (2009a). Evolution of C_4 phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. *Molecular Biology and Evolution*, *26*(2), 357–65.

Christin, P.-A., Samaritani, E., Petitpierre, B., Salamin, N., & Besnard, G. (2009b). Evolutionary insights on C_4 photosynthetic subtypes in grasses from genomics and

phylogenetics. *Genome Biology and Evolution*, *1*, 221–230.

Christin, P. A., & Besnard, G. (2009). Two independent C_4 origins in Aristidoideae (poaceae) revealed by the recruitment of distinct phosphoenolpyruvate carboxylase genes. *American Journal of Botany*, *96*, 2234–2239.

Christin, P. A., Salamin, N., Muasya, a. M., Roalson, E. H., Russier, F., & Besnard, G. (2008). Evolutionary switch and genetic convergence on rbcL following the evolution of C_4 photosynthesis. *Molecular Biology and Evolution*, *25*, 2361–2368.

Christin, P.-A., Salamin, N., Savolainen, V., Duvall, M. R., & Besnard, G. (2007). C_4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. *Current Biology*, *17*(14), 1241–7.

Coen, E. S., Carpenter, R., & Martin, C. (1986). Transposable elements generate novel spatial patterns of gene expression in *Antirrhinum majus*. *Cell*, *47*(2), 285–96.

Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., … Sorek, R. (2016). Comparative transcriptomics across the prokaryotic tree of life. *Nucleic Acids Research*, *44*(W1), W46–W53.

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, *9*(12), 938–950.

Conway-Morris, S. (2003). The Cambrian 'explosion'; of metazoans and molecular biology: would Darwin be satisfied? *The International Journal of Developmental Biology*, *47*(7–8), 505–15.

Cooney, C. R., Bright, J. A., Capp, E. J. R., Chira, A. M., Hughes, E. C., Moody, C. J. A., … Thomas, G. H. (2017). Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature*, *542*(7641), 344–347.

Cooper, T. F., Rozen, D. E., & Lenski, R. E. (2003). Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 1072–7.

Cronquist, A. (1968). The evolution and classification of flowering plants. *The Evolution and Classification of Flowering Plants.*

Crozat, E., Philippe, N., Lenski, R. E., Geiselmann, J., & Schneider, D. (2005). Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics*, *169*(2), 523–32.

Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M., & Drummond, D. A. (2015). Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLOS Genetics*, *11*(5), e1005206.

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., … dePamphilis, C. W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research*, *16*(6), 738–49.

Danforth, B. N., Conway, L., & Ji, S. (2003). Phylogeny of eusocial *Lasioglossum* reveals multiple losses of eusociality within a primitively eusocial clade of bees (Hymenoptera: Halictidae). *Systematic Biology*, *52*(1), 23–36.

Danforth, B. N., Cardinal, S., Praz, C., Almeida, E. A. B., & Michez, D. (2013). The impact of molecular data on our understanding of bee phylogeny and evolution. *Annual Review of Entomology*, *58*(1), 57–78.

Darwin, C., & Wallace, A. R. (1958). Evolution by natural selection.

Darwin C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., … Jiggins, C. D. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*(7405), 94.

Daugaard, M., Rohde, M., & Jäättelä, M. (2007). The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS Letters*, *581*(19), 3702–3710.

Dawkins, R. (1976). *The selfish gene*. Oxford University press.

Dawkins R. (1986). *The blind watchmaker*. Norton, New York.

De Bodt, S., Maere, S., & Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, *20*(11), 591–7.

de Visser, J. A. G. M., & Lenski, R. E. (2002). Long-term experimental evolution in *Escherichia coli*. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evolutionary Biology*, *2*, 19.

de Visser, J. A. G. M., Hermisson, J., Wagner, G. P., Meyers, L. A., Bagheri-Chaichian, H., Blanchard, J. L., … Whitlock, M. C. (2003). Perspective: evolution and detection of genetic robustness. *Evolution*, *57*(9), 1959.

de Visser, J. A. G. M., & Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, *15*(7), 480–490.

Dehal, P., & Boore, J. L. (2005a). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, *3*(10), e314.

DePristo, M. A., Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*,

6(9), 678–687.

Desai, M. M., & Fisher, D. S. (2007). Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics*, *176*(3), 1759–98.

Desirò, A., Rimington, W., Jacob, A., Pol, N., & Smith, M. (2017). Multigene phylogeny of Endogonales, an early diverging lineage of fungi associated with plants. *Fungus*, 8(2), 245-257.

Dever, L. V., Blackwell, R. D., Fullwood, N. J., Lacuesta, M., Leegood, R. C., Onek, L. A., … Lea, P. J. (1995). The isolation and characterization of mutants of the C_4 photosynthetic pathway. *Journal of Experimental Botany*. Oxford University Press.

Dever, L. V., Bailey, K. J., Leegood, R. C., & Lea, P. J. (1997). Control of photosynthesis in *Amaranthus edulis* mutants with reduced amounts of PEP Carboxylase. *Australian Journal of Plant Physiology*, *24*(4), 469.

Dieckmann, U., & Doebeli, M. (1999). On the origin of species by sympatric speciation. *Nature*, *400*(6742), 354–357.

Ding, Z., Weissmann, S., Wang, M., Du, B., Huang, L., Wang, L., … Li, P. (2015). Identification of photosynthesis-associated C_4 candidate genes through comparative leaf gradient transcriptome in multiple lineages of C_3 and C_4 species. *PloS One*, *10*(10), e0140629.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214.

Duboule, D., & Wilkins, A. S. (1998). The evolution of “bricolage”. *Trends in Genetics : TIG*, *14*(2), 54–9.

Dunn, C. W., Howison, M., & Zapata, F. (2013). Agalma: an automated phylogenomics workflow.

Dunning, L. T., Hipperson, H., Baker, W. J., Butlin, R. K., Devaux, C., Hutton, I., … Savolainen, V. (2016). Ecological speciation in sympatric palms: 1. Gene expression, selection and pleiotropy. *Journal of Evolutionary Biology*, *29*(8), 1472–87.

Dunning, L. T., Lundgren, M. R., Moreno-Villena, J. J., Namaganda, M., Edwards, E. J., Nosil, P., … Christin, P.-A. (2017). Introgression and repeated co-option facilitated the recurrent emergence of C_4 photosynthesis among close relatives. *Evolution*, *71*(6), 1541–1555.

Dunning Hotopp, J. C., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Muñoz Torres, M. C., … Werren, J. H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, *317*(5845), 1753–6.

Duret, L., & Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, *17*(1), 68–70.

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, *9*(1), 157.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. h

Edwards, E. J., & Donoghue, M. J. (2013). Is it easy to move and easy to evolve? Evolutionary accessibility and adaptation. *Journal of Experimental Botany*, *64*(13), 4047–4052.

Edwards, E. J., Osborne, C. P., Strömberg, C. A. E., Smith, S. A., & C_4 Grasses Consortium (2010). The origins of C_4 grasslands: integrating evolutionary and ecosystem science. *Science*, 328, 587–591.

Edwards, E. J., & Still, C. J. (2008). Climate, phylogeny and the ecological distribution of C_4 grasses. *Ecology Letters*, *11*, 266–276.

Ehleringer, J. R., Cerling, T. E., & Helliker, B. R. (1997). C_4 photosynthesis, atmospheric CO_2, and climate. *Oecologia*, *112*(3), 285–299.

Elena, S. F., & Lenski, R. E. (2003). Microbial genetics: Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, *4*(6), 457–469.

Ellis, R.P. (1974). The significance of the occurrence of both Kranz and non-Kranz leaf anatomy in the grass species *Alloteropsis semialata*. *South African Journal of Science*, 70,169–173.

Emms, D. M., Covshoff, S., Hibberd, J. M., & Kelly, S. (2016). Independent and parallel evolution of new genes by gene duplication in two origins of C_4 photosynthesis provides new insight into the mechanism of phloem loading in C_4 species. *Molecular Biology and Evolution*, *33*(7), 1796–806.

Engelmann, S., Wiludda, C., Burscheidt, J., Gowik, U., Schlue, U., Koczor, M., … Westhoff, P. (2008). The gene for the P-subunit of glycine decarboxylase from the C_4 species *Flaveria trinervia*: analysis of transcriptional control in transgenic *Flaveria bidentis* (C_4) and *Arabidopsis* (C_3). *Plant Physiology*, *146*(4), 1773–85.

Evans, J., Caemmerer, S., Setchell, B., & Hudson, G. (1994). The relationship between CO_2 transfer conductance and leaf anatomy in transgenic tobacco with a reduced content of rubisco. *Australian Journal of Plant Physiology*, *21*(4), 475.

Fahnenstich, H., Saigo, M., Niessen, M., Zanor, M. I., Andreo, C. S., Fernie, A. R., …

Maurino, V. G. (2007). Alteration of organic acid metabolism in *Arabidopsis* overexpressing the maize C_4 NADP-malic enzyme causes accelerated senescence during extended darkness. *Plant Physiology*, *145*(3), 640–52.

Fares, M. A. (2015). The origins of mutational robustness. *Trends in Genetics*, *31*(7), 373–381.

Fisher, A. E., McDade, L. A., Kiel, C. A., Khoshravesh, R., Johnson, M. A., Stata, M., … Sage, R. F. (2015). Evolutionary history of *Blepharis* (Acanthaceae) and the origin of C_4 photosynthesis in section *Acanthodium*. *International Journal of Plant Sciences*, *176*(8), 770–790.

Fisher, A. E., Hasenstab, K. M., Bell, H. L., Blaine, E., Ingram, A. L., & Columbus, J. T. (2016). Evolutionary history of chloridoid grasses estimated from 122 nuclear loci. *Molecular Phylogenetics and Evolution*, *105*, 1–14.

Fisher, R.A. (1930). *The genetical theory of natural selection*. Oxford University Press.

Flagel, L. E., & Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*, *183*(3), 557–564.

Floris, M., Mahgoub, H., Lanet, E., Robaglia, C., & Menand, B. (2009). Post-transcriptional regulation of gene expression in plants during abiotic stress. *International Journal of Molecular Sciences*, *10*(7), 3168–85.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–45.

Ford, A. G. P., Dasmahapatra, K. K., Rüber, L., Gharbi, K., Cezard, T., & Day, J. J. (2015). High levels of interspecific gene flow in an endemic cichlid fish adaptive radiation from an extreme lake environment. *Molecular Ecology*, *24*(13), 3421–3440.

Fracasso, A., Trindade, L. M., & Amaducci, S. (2016). Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. *BMC Plant Biology*, *16*(1), 115.

Frean, M. L., Barrett, D. R., Ariovich, D., Wolfson, M., & Cresswell, C. F. (1983). Intraspecifíc variability in *Alloteropsis semialata* (R. Br.) Hitchc. *Bothalia*, *14*(3/4), 901–913.

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist*, *160*(6), 712–26.

Furbank, R. T., Chitty, J. A., Jenkins, C. L. D., Taylor, W. C., Trevanion, S. J.,

Caemmerer, S. von, & Ashton, A. R. (1997). Genetic Manipulation of Key Photosynthetic Enzymes in the C_4 Plant *Flaveria bidentis*. *Australian Journal of Plant Physiology*, *24*(4), 477.

Furumoto, T., Yamaguchi, T., Ohshima-Ichie, Y., Nakamura, M., Tsuchida-Iwata, Y., Shimamura, M., … Izui, K. (2011). A plastidial sodium-dependent pyruvate transporter. *Nature*, *476*(7361), 472–475.

Gallie, D. R. (1993). Posttranscriptional regulation of gene expression in plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, *44*(1), 77–105.

Gamble, T., Greenbaum, E., Jackman, T. R., Russell, A. P., & Bauer, A. M. (2012). Repeated origin and loss of adhesive toepads in geckos. *PLoS ONE*, *7*(6), e39429.

Gauthier, J. A., & Padian, K. (1989). The origin of birds and the evolution of flight. *Short Courses in Paleontology*, *2*, 121–133.

Gehring, W. J., & Ikeo, K. (1999). Pax 6: mastering eye morphogenesis and eye evolution. *Trends in Genetics*, *15*(9), 371–7.

Gerlee, P. (2015). Directional variation in evolution: consequences for the fitness landscape metaphor. *bioRxiv* 015529.

Gerrish, P. J., & Lenski, R. E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica*, *102–103*(1–6), 127–44.

Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., … Waterston, R. (2014). Comparative analysis of the transcriptome across distant species. *Nature*, *512*(7515), 445–448.

Ghannoum, O., Evans, J. R., Chow, W. S., Andrews, T. J., Conroy, J. P., & von Caemmerer, S. (2005). Faster Rubisco is the key to superior nitrogen-use efficiency in NADP-malic enzyme relative to NAD-malic enzyme C_4 grasses. *Plant Physiology*, *137*(2), 638–50.

Ghannoum, O., Evans, J. R., & von Caemmerer, S. (2010). *Nitrogen and water use efficiency of C_4 plants*. Chapter 8, 129–146.

Gibbs Russell, G. E., & E., G. (1983). The taxonomic position of C_3 and C_4 *Alloteropsis semialata* (Poaceae) in southern Africa. *Bothalia*, *14*(2), 205–213.

Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, *319*(6055), 618–618.

Gillespie, J. H. (1994). *The causes of molecular evolution*. Oxford University Press.

Glover, B. J., Airoldi, C. A., Brockington, S. F., Fernández-Mazuecos, M., Martínez-Pérez, C., Mellers, G., … Taylor, L. (2015). How have advances in comparative floral development influenced our understanding of floral evolution? *International*

Journal of Plant Sciences, *176*(4), 307–323.

Goldberg, E. E., & Igić, B. (2008). On phylogenetic tests of irreversible evolution. *Evolution*, *62*(11), 2727–2741.

Gould, S. J., & Vrba, E. S. (1982). Exaptation—a Missing Term in the Science of Form. *Paleobiology, 8(1), 4-15*

Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. M., & Westhoff, P. (2011). Evolution of C_4 photosynthesis in the genus *Flaveria*: how many and which genes does it take to make C_4? *The Plant Cell*, *23*(6), 2087–105.

Gowik, U., Burscheidt, J., Akyildiz, M., Schlue, U., Koczor, M., Streubel, M., & Westhoff, P. (2004). *cis* -regulatory elements for mesophyll-specific gene expression in the C_4 plant *Flaveria trinervia*, the promoter of the C_4 phosphoenolpyruvate carboxylase gene. *The Plant Cell*, *16*(5), 1077–1090.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. *New Phytol*, 193,304–312.

Griffiths, H., Weller, G., Toy, L. F. M., & Dennis, R. J. (2013). You're so vein: bundle sheath physiology, phylogeny and evolution in C_3 and C_4 plants. *Plant, Cell & Environment*, *36*(2), 249–261.

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*(5), 696–704.

Haldane, J. B. S., Sprunt, A. D., & Haldane, N. M. (1915). Reduplication in mice (Preliminary Communication). *Journal of Genetics*, *5*(2), 133–135.

Halliday, T. J. D., Upchurch, P., & Goswami, A. (2016). Eutherians experienced elevated evolutionary rates in the immediate aftermath of the Cretaceous-Palaeogene mass extinction. *Proceedings. Biological Sciences*, *283*(1833), 20153026.

Hancock, L., & Edwards, E. J. (2014). Phylogeny and the inference of evolutionary trajectories. *Journal of Experimental Botany*, *65*(13), 3491–3498.

Hatch, M. D. (1987). C_4 photosynthesis: a unique elend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics*, *895*(2), 81–106.

Hatch, M. D., & Boardman, N. K. (1987). *The biochemistry of plants: a comprehensive treatise. Volume 10, Photosynthesis*. Academic Press.

Hatch, M. D., & Burnell, J. N. (1990). Carbonic anhydrase activity in leaves and its role in the first step of C_4 photosynthesis. *Plant Physiology*, *93*(2).

Hatch, M. D., & Osmond, C. B. (1976). Compartmentation and Transport in C_4 Photosynthesis. In *Transport in Plants III* (pp. 144–184).

Häusler, R. E., Hirsch, H., Kreuzaler, F., & Peterhänsel, C. (2002). Overexpression of C_4-cycle enzymes in transgenic C_3 plants: a biotechnological approach to improve C_3-photosynthesis. *Journal of Experimental Botany*, *53*(369), 591–607.

Häusler, R. E., Rademacher, T., Li, J., Lipka, V., Fischer, K. L., Schubert, S., … Hirsch, H. (2001). Single and double overexpression of C_4-cycle genes had differential effects on the pattern of endogenous enzymes, attenuation of photorespiration and on contents of UV protectants in transgenic potato and tobacco plants. *Journal of Experimental Botany*, *52*(362), 1785–1803.

Hayden, E. J., Ferrada, E., & Wagner, A. (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, *474*(7349), 92–95.

He, X., & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, *169*(2), 1157–64.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A. P. M., & Lercher, M. J. (2013). Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell*, *153*(7), 1579–1588.

Heers, A. M., Dial, K. P., & Tobalske, B. W. (2014). From baby birds to feathered dinosaurs: incipient wings and the evolution of flight. *Paleobiology*, *40*(3), 459–476.

Hendry, A. P., Taylor, E. B., & McPhail, J. D. (2009). Adaptive divergence and the balance between selection and gene flow: lake and stream stickleback in the misty system. *Evolution* 56(6):1199-1216

Herron, M. D., & Michod, R. E. (2008). Evolution of complexity in the volvocine algae: transitions in individuality through Darwin's eye. *Evolution*, *62*(2), 436–451.

Hibberd, J. M., & Covshoff, S. (2010a). The regulation of gene expression required for C_4 photosynthesis. *Annual Review of Plant Biology*, *61*, 181–207.

Hibberd, J. M., & Covshoff, S. (2010b). The regulation of gene expression required for C_4 photosynthesis. *Annual Review of Plant Biology*, *61*(1), 181–207.

Hibberd, J. M., & Quick, W. P. (2002). Characteristics of C_4 photosynthesis in stems and petioles of C_3 flowering plants. *Nature*, *415*(6870), 451–454.

Hibberd, J. M., Sheehy, J. E., & Langdale, J. A. (2008). Using C_4 photosynthesis to increase the yield of rice—rationale and feasibility. *Current Opinion in Plant Biology*, *11*(2), 228–231.

Holland, P. W. H., Marlétaz, F., Maeso, I., Dunwell, T. L., & Paps, J. (2017). New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1713), 20150480.

Hu, J., & Ng, P. C. (2012). Predicting the effects of frameshifting indels. *Genome Biology*, *13*(2), R9.

Huang, R., O'Donnell, A. J., Barboline, J. J., & Barkman, T. J. (2016). Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. *Proceedings of the National Academy of Sciences*, *113*(38), 10613–10618.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., … Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 16048.

Humphreys, A. M., & Linder, H. P. (2013). Evidence for recent evolution of cold tolerance in grasses suggests current distribution is not limited by (low) temperature. *New Phytologist*, *198*(4), 1261–1273.

Huston, M. A. (1985). Patterns of species diversity on coral reefs. *Annual Review of Ecology and Systematics*, *16*(1), 149–177.

Hylton, C. M., Rawsthorne, S., Smith, A. M., Jones, D. A., & Woolhouse, H. W. (1988). Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C_3-C_4 intermediate species. *Planta*, *175*(4), 452–459.

Ibrahim, D. G., Burke, T., Ripley, B. S., & Osborne, C. P. (2009). A molecular phylogeny of the genus *Alloteropsis* (Panicoideae, Poaceae) suggests an evolutionary reversion from C_4 to C_3 photosynthesis. *Annals of Botany*, *103*(1), 127–36.

Igic, B., Bohs, L., & Kohn, J. R. (2006). Ancient polymorphism reveals unidirectional breeding system shifts. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(5), 1359–63.

Igic, B., & Busch, J. W. (2013). Is self-fertilization an evolutionary dead end? *New Phytologist*, *198*(2), 386–397.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161–6.

Janke, C., & Chloë Bulinski, J. (2011). Post-translational regulation of the microtubule cytoskeleton: mechanisms and functions. *Nature Reviews Molecular Cell Biology*,

12(12), 773–786.

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, O. (2012). The global diversity of birds in space and time. *Nature*, *491*(7424), 444–8.

Jiao, Y., Li, J., Tang, H., & Paterson, A. H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell*, *26*(7), 2792–802.

Jiggins, C. D., Wallbank, R. W. R., & Hanly, J. J. (2016). Waiting in the wings: what can we learn about gene co-option from the diversification of butterfly wing patterns? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *372*(1713).

Jogesh, T., Overson, R. P., Raguso, R. A., & Skogen, K. A. (2016). Herbivory as an important selective force in the evolution of floral traits and pollinator shifts. *AoB Plants*, *9*(1), plw088.

John, C. R., Smith-Unna, R. D., Woodfield, H., Covshoff, S., & Hibberd, J. M. (2014). Evolutionary convergence of cell-specific gene expression in independent lineages of C_4 grasses. *Plant Physiology*, *165*, 62–75.

Jones, C. T., Youssef, N., Susko, E., & Bielawski, J. P. (2017). Shifting balance on a static mutation–selection landscape: a novel scenario of positive selection. *Molecular Biology and Evolution*, *34*(2), 391-407.

Jones, G. & Teeling, E. C. (2006). The evolution of echolocation in bats. *Trends in Ecology & Evoluton*.

Kadereit, G., Lauterbach, M., Pirie, M. D., Arafeh, R., & Freitag, H. (2014). When do different C_4 leaf anatomies indicate independent C_4 origins? Parallel evolution of C_4 leaf types in Camphorosmeae (Chenopodiaceae). *Journal of Experimental Botany*, *65*(13), 3499–511.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes, *20*(10), 1313–1326.

Kajala, K., Brown, N. J., Williams, B. P., Borrill, P., Taylor, L. E., & Hibberd, J. M. (2012). Multiple *Arabidopsis* genes primed for recruitment into C_4 photosynthesis. *Plant Journal*, *69*, 47–56.

Kassahn, K. S., Dang, V. T., Wilkins, S. J., Perkins, A. C., & Ragan, M. A. (2009). Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Research*, *19*(8), 1404–18.

Katju, V., & Bergthorsson, U. (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Frontiers in Genetics*, *4*, 273.

Kayal, E., Roure, B., Philippe, H., Collins, A. G., & Lavrov, D. V. (2013). Cnidarian phylogenetic relationships as revealed by mitogenomics. *BMC Evolutionary Biology*, *13*(1), 5.

Kennedy, R. A., & Laetsch, W. M. (1974). Plant species intermediate for C_3, C_4 photosynthesis. *Science*, *184*(4141), 1087–9.

Kellogg, E. A. (1999). Phylogenetic aspects of the evolution of C_4 photosynthesis. In: Sage R. F., and R. K. Monson, eds. *C_4 plant biology.* San Diego: Academic Press, 411-444.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

King, B., & Lee, M. S. Y. (2015). Ancestral state reconstruction, rate heterogeneity, and the evolution of reptile viviparity. *Systematic Biology*, *64*(3), 532–544.

Kirschner, M., & Gerhart, J. (1998). Evolvability. *Proceedings of the National Academy of Sciences*, *95*(15), 8420–8427.

Knight, K. (2015). Lungfish hear air-borne sound. *Journal of Experimental Biology*, *218*(3), 329–330. h

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38.

Kondrashov, F. A., & Kondrashov, A. S. (2006). Role of selection in fixation of gene duplications. *Journal of Theoretical Biology*, *239*(2), 141–151.

Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J., & Smith, V. A. (2015). Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Scientific Reports*, *5*, 10775.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., & Koonin, E. V. (2011). Computational methods for gene orthology inference. *Briefings in Bioinformatics*, *12*(5), 379–391.

Ku, M. S., Monson, R. K., Littlejohn, R. O., Nakamoto, H., Fisher, D. B., Edwards, G. E., & Edwards, G. E. (1983). Photosynthetic characteristics of C_3- C_4 intermediate *Flaveria* species : I. Leaf anatomy, photosynthetic responses to O_2 and CO_2 , and activities of key enzymes in the C_3 and C_4 pathways. *Plant Physiology*, *71*(4), 944–8.

Külahoglu, C., Denton, A. K., Sommer, M., Maß, J., Schliesky, S., Wrobel, T. J., … Weber, A. P. M. (2014). Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C_3 and C_4 plant species. *The Plant*

Cell, *26*(8), 3243–60.

Lagemaat, L. van de, Landry, J.,... (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Genetics*, 19(10), 530-536.

Landry, C. R., Lemos, B., Rifkin, S., Dickinson, W. J., & Hartl, D. L. (2007). Genetic properties influencing the evolvability of gene expression. *Science*, *317*(5834), 118–21.

Lange, B. M., Rujan, T., Martin, W., & Croteau, R. (2000). Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(24), 13172–7.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Lefort, V., Longueville, J.-E., & Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, *34*(9), 2422–2424.

Lehmann, C. E. R., & Parr, C. L. (2016). Tropical grassy biomes: linking ecology, human use and conservation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1703), 20160329.

Lenormand, T., Guillemaud, T., Bourguet, D., & Raymond, M. (1998). Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito C*ulex pipiens*. *Evolution*, *52*(6), 1705–1712.

Lenski, R. E. 2017 The *E. coli* long-term experimental evolution project site. http://myxo.css.msu.edu/ecoli

Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features. *Nature*, *423*(6936), 139–44.

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, *44*(W1), W242–W245.

Li, F.-W., Villarreal, J. C., Kelly, S., Rothfels, C. J., Melkonian, M., Frangedakis, E., … Pryer, K. M. (2014). Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(18), 6672–7.

Li, G., Davis, B., Eizirik, E., & Murphy, W. (2016). Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Research*, 26(1), 1–11.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome

Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, Y., Ma, X., Zhao, J., Xu, J., Shi, J., Zhu, X.-G., … Zhang, H. (2015). Developmental genetic mechanisms of C_4 syndrome based on transcriptome analysis of C_3 cotyledons and C_4 assimilating shoots in Haloxylon ammodendron. *PLOS ONE*, *10*(2), e0117175.

Li, Y., Xu, J., Haq, N. U., Zhang, H., & Zhu, X.-G. (2014). Was low CO_2 a driving force of C_4 evolution: *Arabidopsis* responses to long-term low CO_2 stress. *Journal of Experimental Botany*, *65*(13), 3657–3667.

Liberles, D. A., Teufel, A. I., Liu, L., & Stadler, T. (2013). On the need for mechanistic models in computational genomics and metagenomics. *Genome Biology and Evolution*, *5*(10), 2008–18.

Linné, C. von. (1964). *Systema naturae 1735*. Ed Coronet Books.

Liu, Y., Zhou, M., Gao, Z., Ren, W., Yang, F., He, H., & Zhao, J. (2015). RNA-Seq analysis reveals MAPKKK family members related to drought tolerance in maize. *PloS One*, *10*(11), e0143128.

Lovejoy, C. O. (1988). Evolution of human walking. *Scientific American*. Scientific American, a division of Nature America, Inc.

Lundgren, M. R., Besnard, G., Ripley, B. S., Lehmann, C. E. R., Chatelet, D. S., Kynast, R. G., … Christin, P.-A. (2015). Photosynthetic innovation broadens the niche within a single species. *Ecology Letters*, 18(10), 1021–1029

Lundgren, M. R., Christin, P.-A., Escobar, E. G., Ripley, B. S., Besnard, G., Long, C. M., … Osborne, C. P. (2016). Evolutionary implications of C_3-C_4 intermediates in the grass *Alloteropsis semialata*. *Plant, Cell & Environment*, *39*(9), 1874–1885.

Lundgren, M. R., Osborne, C. P., & Christin, P.-A. (2014). Deconstructing Kranz anatomy to understand C_4 evolution. *Journal of Experimental Botany*, *65*(13), 3357–3369.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–73.

Lyu, M.-J. A., Gowik, U., Kelly, S., Covshoff, S., Mallmann, J., Westhoff, P., … Zhu, X.-G. (2015). RNA-Seq based phylogeny recapitulates previous phylogeny of the genus *Flaveria* (Asteraceae) with some modifications. *BMC Evolutionary Biology*, *15*, 116.

Maddison, W. P., Knowles, L. L., & Collins, T. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30.

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M. J., Weber, A. P. M., Westhoff, P., & Gowik, U. (2014). The role of photorespiration during the evolution of C_4 photosynthesis in the genus *Flaveria*. *eLife*, *3*, e02478.

Marazzi, B., Ane, C., Simon, M. F., Delgado-Salinas, A., Luckow, M., & Sanderson, M. J. (2012). Locating evolutionary precursors on a phylogenetic tree. *Evolution*, *66*(12), 3918–3930.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium, … Olsen, O.-A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, *345*(6194), 1250092.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., … Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, *23*(11), 1817–28.

Mason-Gamer, R. J., Burns, M. M., & Naum, M. (2010). reticulate evolutionary history of a complex group of grasses: phylogeny of *Elymus* StStHH allotetraploids based on three nuclear genes. *PLoS ONE*, *5*(6), e10989.

Mason-Gamer, R. J., & Linder, P. (2004). Reticulate evolution, introgression, and intertribal gene capture in an allohexaploid grass. *Systematic Biology*, *53*(1), 25–37.

Masumoto, C., Miyazawa, S.-I., Ohkawa, H., Fukuda, T., Taniguchi, Y., Murayama, S., … Miyao, M. (2010). Phosphoenolpyruvate carboxylase intrinsically located in the chloroplast of rice plays a crucial role in ammonium assimilation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(11), 5226–31.

McGuire, J. A., Witt, C. C., Remsen, J. V, Corl, A., Rabosky, D. L., Altshuler, D. L., & Dudley, R. (2014). Molecular phylogenetics and the diversification of hummingbirds. *Current Biology*, *24*(8), 910–6.

McKown, A. D., & Dengler, N. G. (2007). Key innovations in the evolution of Kranz anatomy and C_4 vein pattern in *Flaveria* (Asteraceae). *American Journal of Botany*, *94*(3), 382–99.

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*, 14363.

Meléndez-Hevia, E., Waddell, T. G., & Cascante, M. (1996). The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *Journal of Molecular Evolution*, *43*(3), 293–303.

Min, X. J., Butler, G., Storms, R., & Tsang, A. (2005). OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Research*, *33*(Web Server issue), W677-80.

Moreno-Villena, J. J., Dunning, L. T., Osborne, C. P., & Christin, P.-A. (2018). Highly expressed genes are preferentially co-opted for C_4 photosynthesis. *Molecular Biology and Evolution*, 35(1), 94-106

Monson, R. K. (2003). Gene duplication, neofunctionalization, and the evolution of C_4 Photosynthesis. *International Journal of Plant Sciences*, *164*(S3), S43–S54.

Monson, R. K., Edwards, G. E., & Ku, M. S. B. (1984). C_3-C_4 Intermediate photosynthesis in plants. *BioScience*, *34*(9), 563–574.

Monson, R. K., Moore, B. d., Ku, M. S. B., & Edwards, G. E. (1986). Co-function of C_3-and C_4-photosynthetic pathways in C_3, C_4 and C_3-C_4 intermediate *Flaveria* species. *Planta*, *168*(4), 493–502.

Monteiro, A., & Podlaha, O. (2009). Wings, horns, and butterfly eyespots: How do complex traits evolve? *PLoS Biology*, *7*(2), 0209–0216.

Montiel, E. E., Badenhorst, D., Tamplin, J., Burke, R. L., & Valenzuela, N. (2017). Discovery of the youngest sex chromosomes reveals first case of convergent co-option of ancestral autosomes in turtles. *Chromosoma*, *126*(1), 105–113.

Moreau, C. S., & Bell, C. D. (2013). Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. *Evolution*, *67*(8), 2240–2257.

Morgan, C. L., Turner, S. R., & Rawsthorne, S. (1993). Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine hydroxymethyltransferase in leaves of C_3-C_4 intermediate species from different genera. *Planta*, *190*(4), 468–473.

Morgan, J. A., & Brown, R. H. (1979). Photosynthesis in Grass Species Differing in Carbon Dioxide Fixation Pathways: II. A Search for Species with Intermediate Gas Exchange and Anatomical Characteristics. P*lant physiology*, *64*(2), 257–262.

Morgan, J. A., Brown, R. H., & Reger, B. J. (1980). Photosynthesis in grass species differing in carbon dioxide fixation pathways: III. Oxygen response and enzyme activities of species in the laxa group of *Panicum*. *Plant Physiology*, *65*(1), 156–9. http://doi.org/10.1104/PP.65.1.156

Muller, H. J. (1932). Some genetic aspects of sex. *The American Naturalist*, *66*(703), 118–138.

Nadeau, N. (2014). Butterfly genomics sheds light on the process of hybrid speciation.

Molecular Ecology, *23*(18), 4441–4443.

Niemiller, M. L., Fitzpatrick, B. M., Shah, P., Schmitz, L., & Near, T. J. (2013). evidence for repeated loss of selective constraint in rhodopsin of amblyopsid cavefishes (teleostei: amblyopsidae). *Evolution*, *67*(3), 732–748.

Nilsson, D. E. (2013). Eye evolution and its functional basis. *Visual Neuroscience*, *30*(1–2), 5–20.

Nolis, I. K., McKay, D. J., Mantouvalou, E., Lomvardas, S., Merika, M., & Thanos, D. (2009). Transcription factors mediate long-range enhancer-promoter interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20222–7.

Norris, R. D. (2000). Pelagic species diversity, biogeography, and evolution. *Paleobiology,* 26(4), Supplement. *DEEP TIME: Paleobiology's Perspective*, 236-258

Ohno, S. (1970). The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*, *99*(1), 120–130.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, *23*(1), 263–286.

Oliver, J. C., Tong, X.-L., Gall, L. F., Piel, W. H., & Monteiro, A. (2012). A single origin for nymphalid butterfly eyespots followed by widespread loss of associated gene expression. *PLoS Genetics*, *8*(8), e1002893.

Olofsson, J., Bianconi, M., Besnard, G., Dunning, L., Lundgren, M., Holota, H., … Christin, P. (2016). Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology*, 25(24), 6107–6123.

Orgel, L. E. (1994). The Origin of Life on the Earth. *Scientific American*. 271(4), 76-83.

Orphanides, G. (2002). A unified theory of gene expression. *Cell*, *108*(4), 439–451.

Osborne, C. P., Edwards, E. J., & Christin, P. (2015). Genetic enablers underlying the clustered evolutionary origins of C_4 photosynthesis in angiosperms Article Fast Track, *32*(4), 846–858.

Osborne, C. P., Salomaa, A., Kluyver, T. A., Visser, V., Kellogg, E. A., Morrone, O., … Simpson, D. A. (2014). A global database of C_4 photosynthesis in grasses. *The New Phytologist*, *204*(3), 441–6.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884.

Pagel, M. (2004). Limpets break Dollo's Law. *Trends in Ecology & Evolution*, *19*(6), 278–80.

Pál, C., & Papp, B. (2017). Evolution of complex adaptations in molecular systems. *Nature Ecology & Evolution*, *1*(8), 1084–1092.

Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics*, *25*(1), 547–572.

Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, *171*(4), 2294–316.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)*, *20*(2), 289–90.

Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., … Jiggins, C. D. (2012). Adaptive introgression across species boundaries in *Heliconius* Butterflies. *PLoS Genetics*, *8*(6), e1002752.

Pearcy, R. W., & Ehleringer, J. (1984). Comparative ecophysiology of C_3 and C_4 plants. *Plant, Cell and Environment*, *7*(1), 1–13.

Pengelly, J. J. L., Tan, J., Furbank, R. T., & von Caemmerer, S. (2012). Antisense reduction of NADP-malic enzyme in *Flaveria bidentis* reduces flow of CO_2 through the C_4 cycle. *Plant Physiology*, *160*(2), 1070–80.

Pennuto, M., Palazzolo, I., & Poletti, A. (2009). Post-translational modifications of expanded polyglutamine proteins: impact on neurotoxicity. *Human Molecular Genetics*, *18*(R1), R40–R47.

Perfeito, L., Fernandes, L., Mota, C., & Gordo, I. (2007). Adaptive mutations in bacteria: high rate and small effects. *Science*, *317*(5839), 813–5.

Prendergast, H., Hattersley, P., & Stone, N. (1987). New structural/biochemical associations in leaf blades of C_4 grasses (Poaceae). *Australian Journal of Plant Physiology*, *14*(4), 403.

Protas, M. E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W. R., … Tabin, C. J. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature Genetics*, *38*(1), 107–111.

Pyron, R. A., & Burbrink, F. T. (2014). Early origin of viviparity and multiple reversions to oviparity in squamate reptiles. *Ecology Letters*, *17*(1), 13–21.

Rabosky, D. L., Santini, F., Eastman, J., Smith, S. A., Sidlauskas, B., Chang, J., & Alfaro, M. E. (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications*, *4*, 1958.

Rajendrudu, G., Prasad, J. S., & Das, V. S. (1986). C_3-C_4 intermediate species in A*lternanthera* (amaranthaceae) : leaf anatomy, CO_2 compensation point, net CO_2 exchange and activities of photosynthetic enzymes. *Plant Physiology*, *80*(2), 409–14.

Rao, X., Lu, N., Li, G., Nakashima, J., Tang, Y., & Dixon, R. A. (2016). Comparative cell-specific transcriptomics reveals differentiation of C_4 photosynthesis pathways in switchgrass and other C_4 lineages. *Journal of Experimental Botany*, *67*(6), 1649–1662.

Rawsthorne, S., Hylton, C. M., Smith, A. M., & Woolhouse, H. W. (1988a). Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C_3-C_4 intermediate species *Moricandia arvensis* (L.) DC. *Planta*, *176*(4), 527–532.

Rawsthorne, S., Hylton, C. M., Smith, A. M., & Woolhouse, H. W. (1988b). Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C_3 and C_3-C_4 intermediate species of *Moricandia*. *Planta*, *173*(3), 298–308.

Reed, R. D., Papa, R., Martin, A., Hines, H. M., Counterman, B. A., Pardo-Diaz, C., … McMillan, W. O. (2011). Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, *333*(6046), 1137–41.

Reiss, K. Z. (2001). Using phylogenies to study convergence: the case of the ant-eating mammals. *American Zoologist*, *41*(3), 507–525.

Remold, S. K., & Lenski, R. E. (2001). Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11388–93.

Renvoize, S. A. (1987). A survey of leaf-blade anatomy in grasses XI. Paniceae. *Kew Bulletin*, *42*(3), 739.

Reyna-Llorens, I., & Hibberd, J. M. (2017). Recruitment of pre-existing networks during the evolution of C_4 photosynthesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1730), 20160386.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Rosenzweig, M. L. (1995). *Species diversity in space and time*. Cambridge University Press.

Ruff, C. B. (1994). Morphological adaptation to climate in modern and fossil hominids. *American Journal of Physical Anthropology*, *37*(S19), 65–107.

Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, *8*(3), 336–352.

Rutschmann, F. (2006). Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity Distributions*, *12*(1), 35–48.

Sage, R. F. (2001). Environmental and evolutionary preconditions for the origin and diversification of the C_4 photosynthetic syndrome. *Plant Biology*, *3*(3), 202–213.

Sage, R. F. (2004). The evolution of C_4 photosynthesis. *New Phytologist*, *161*(2), 341–370.

Sage, R. F., Christin, P.-A., & Edwards, E. J. (2011). The C_4 plant lineages of planet Earth. *Journal of Experimental Botany*, *62*(9), 3155–69.

Sage, R. F., & Monson, R. K. (1999). The origins of C_4 genes and evolutionary pattern in the C_4 metabolic phenotype. In C_4 *Plant Biology* (pp. 377–410).

Sage, R. F., & Monson, R. K.,Russell K. . (1999). *C_4 plant biology*. Academic Press.

Sage, R. F., Sage, T. L., & Kocacinar, F. (2012). Photorespiration and the evolution of C_4 Photosynthesis. *Rev. Plant Biol*, *63*, 19–47.

Sage, T. L., Busch, F. A., Johnson, D. C., Friesen, P. C., Stinson, C. R., Stata, M., … Sage, R. F. (2013). Initial Events during the Evolution of C_4 Photosynthesis in C_3 Species of *Flaveria*. P*lant physiology*, *163*(3), 1266–1276.

Sawyer, S. A., Parsch, J., Zhang, Z., & Hartl, D. L. (2007). Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(16), 6504–10.

Sayou, C., Monniaux, M., Nanao, M. H., Moyroud, E., Brockington, S. F., Thévenon, E., … Dumas, R. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science*, *343*(6171), 645–8.

Sayre, R. T., & Kennedy, R. A. (1977). Ecotypic differences in the C_3 and C 4 photosynthetic activity in *Mollugo verticillata*, a C_3-C_4 intermediate. *Planta*, *134*(3), 257–62.

Schlüter, U., & Weber, A. P. M. (2016). The road to C_4 photosynthesis: evolution of a complex trait via intermediary states. *Plant and Cell Physiology*, *57*(5), 881–889.

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–5.

Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human

evolutionary history. *Nature Reviews Genetics*, *16*(12), 727–740.

Schuler, M. L., Mantegazza, O., & Weber, A. P. M. (2016). Engineering C_4 photosynthesis into C_3 chassis in the synthetic biology age. *The Plant Journal*, *87*(1), 51–65.

Schulze, S., Mallmann, J., Burscheidt, J., Koczor, M., Streubel, M., Bauwe, H., … Westhoff, P. (2013). Evolution of C_4 photosynthesis in the genus *Flaveria*: establishment of a photorespiratory CO_2 pump. *The Plant Cell*, *25*(7), 2522–35.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., … Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342.

Scroggins, B. T., & Neckers, L. (2007). Post-translational modification of heat-shock protein 90: impact on chaperone function. *Expert Opinion on Drug Discovery*, *2*(10), 1403–1414.

Shantz, H., & Piemeisel, L. (1927). *The water requirement of plants at Akron, Colo*.

Shen, C., Li, D., He, R., Fang, Z., Xia, Y., Gao, J., … Cao, M. (2014). Comparative transcriptome analysis of RNA-seq data for cold-tolerant and cold-sensitive rice genotypes under cold stress. *Journal of Plant Biology*, *57*(6), 337–348.

Shi, S., Huang, Y., Zeng, K., Tan, F., He, H., Huang, J., & Fu, Y. (2005). Molecular phylogenetic analysis of mangroves: independent evolutionary origins of vivipary and salt secretion. *Molecular Phylogenetics and Evolution*, *34*(1), 159–66.

Shimizu, K. K., Shimizu-Inatsugi, R., Tsuchimatsu, T., & Purugganan, M. D. (2007). Independent origins of self-compatibility in *Arabidopsis thaliana*. *Molecular Ecology*, *17*(2), 704–714.

Sinha, N. R., & Kellogg, E. A. (1996). Parallelism and diversity in multiple origins of C_4 photosynthesis in the grass family. *American Journal of Botany*, *83*(11), 1458.

Smith, J., & Kronforst, M. R. (2013). Do *Heliconius* butterfly species exchange mimicry alleles? *Biology Letters*, *9*(4), 20130503

Soltis, P. S., Marchant, D. B., Van de Peer, Y., & Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development*, *35*, 119–125.

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., … Kohn, M. H. (2011). Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, *21*(15), 1296–301.

Sonnhammer, E. L. ., & Koonin, E. V. (2002). Orthology, paralogy and proposed

classification for paralog subtypes. *Trends in Genetics*, *18*(12), 619–620.

Sonnhammer, E. L. L., & Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, *43*(Database issue), D234-9.

Soreng, R. J., Peterson, P. M., Romaschenko, K., Davidse, G., Zuloaga, F. O., Judziewicz, E. J., … Morrone, O. (2015). A worldwide phylogenetic classification of the Poaceae (Gramineae). *Journal of Systematics and Evolution*, *53*(2), 117–137.

Soros, C. L., & Dengler, N. G. (2001). Ontogenetic derivation and cell differentiation in photosynthetic tissues of C_3 and C_4 Cyperaceae. *American Journal of Botany*, *88*(6), 992–1005.

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, *16*(8), 472–482.

South, P. F., Walker, B. J., Cavanagh, A. P., Rolland, V., Badger, M., & Ort, D. R. (2017). Bile Acid Sodium Symporter BASS6 can transport glycolate and is involved in photorespiratory metabolism in *Arabidopsis thaliana*. *The Plant Cell*, *29*(4), 808–823.

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics*, *14*(11), 751–764.

Stiffler, M. A., Hekstra, D. R., & Ranganathan, R. (2015). Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell*, *160*(5), 882–92.

Still, C. J., Berry, J. A., Collatz, G. J., & DeFries, R. S. (2003). Global distribution of C_3 and C_4 vegetation: Carbon cycle implications. *Global Biogeochemical Cycles*, *17*(1), 6-1-6–14.

Storz, J. F. (2016). Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics*, *17*(4), 239–250.

Studer, R. A., Christin, P.-A., Williams, M. A., & Orengo, C. A. (2014). Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(6), 2223–8.

Svensson, P., Bläsing, O. E., & Westhoff, P. (2003). Evolution of C_4 phosphoenolpyruvate carboxylase. *Archives of Biochemistry and Biophysics*, *414*(2), 180–188.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729.

Tang, H., Bowers, J. E., Wang, X., & Paterson, A. H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 472–7.

Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., … Harmon, L. J. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist*, *207*(2), 454–467.

Tausta, S. L., Miller Coyle, H., Rothermel, B., Stiefel, V., & Nelson, T. (2002). Maize C_4 and non-C_4 NADP-dependent malic enzymes are encoded by distinct genes derived from a plastid-localized ancestor. *Plant Molecular Biology*, *50*(4/5), 635–652.

Tcherkez, G. G. B., Farquhar, G. D., & Andrews, T. J. (2006). Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(19), 7246–51.

Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., & Gaut, B. S. (2012). The molecular diversity of adaptive convergence. *Science*, *335*(6067), 457–61.

Terborgh, J. (1992). *Diversity and the Tropical Rain Forest.* Scientific American Library, W. H. Freeman, New York.

Tetu, S. G., Tanz, S. K., Vella, N., Burnell, J. N., & Ludwig, M. (2007). The *Flaveria bidentis* *β*-carbonic anhydrase gene family encodes cytosolic and chloroplastic isoforms demonstrating distinct organ-specific expression patterns. *Plant Physiology*, *144*(3), 1316–27.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.

Tobin, E. M., & Silverthorne, J. (1985). Light regulation of gene expression in higher plants. *Annual Review of Plant Physiology*, *36*(1), 569–593.

Tobin, E. M., & Suttie, J. L. (1980). Light Effects on the Synthesis of Ribulose-1,5-Bisphosphate Carboxylase in *Lemna gibba* L. G-3. P*lant physiology*, *65*(4), 641–647.

Tokuriki, N., & Tawfik, D. S. (2009). Protein dynamism and evolvability. *Science*, *324*(5924), 203–207.

True, J. R., & Carroll, S. B. (2002). Gene co-option in physiological and morphological

evolution. *Annual Review of Cell and Developmental Biology*, *18*(1), 53–80.

Trueman, J. W. H., Pfeil, B. E., Kelchner, S. A., & Yeates, D. K. (2004). Did stick insects *really* regain their wings? *Systematic Entomology*, *29*(2), 138–139.

Ueno, O., & Sentoku, N. (2006). Comparison of leaf structure and photosynthetic characteristics of C_3 and C_4 *Alloteropsis semialata* subspecies. *Plant, Cell and Environment*, *29*(2), 257–268.

van de Lagemaat, L. N., Landry, J.-R., Mager, D. L., & Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics*, *19*(10), 530–6.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426.

Van Tuinen, M., Butvill, D. B., Kirsch, J. A., & Hedges, S. B. (2001). Convergence and divergence in the evolution of aquatic birds. *Proceedings. Biological Sciences*, *268*(1474), 1345–50.

Vicentini, A., Barber, J. C., Aliscioni, S. S., Giussani, L. M., & Kellogg, E. A. (2008). The age of the grasses and clusters of origins of C_4 photosynthesis. *Global Change Biology*, *14*(12), 2963–2977.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., & Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, *19*(2), 327–35.

de Visser, J. A. G. M., & Lenski, R. E. (2002). Long-term experimental evolution in *Escherichia coli*. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evolutionary Biology*,

de Visser, J. A. G. M., & Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, *15*(7), 480–490.

von Caemmerer, S., & Furbank, R. T. (2003). The C_4 pathway: an efficient CO_2 pump. *Photosynthesis Research*, *77*(2/3), 191–207.

von Caemmerer, S., Evans, J. R., Cousins, A. B., Badger, M. R., & Furbank, R. T. (2008). C_4 photosynthesis and CO_2 diffusion. In *Charting New Pathways to C_4 Rice* (pp. 95–115). World Scientific.

von Caemmerer, S., Quick, W. P., & Furbank, R. T. (2012). The development of C_4 rice: current progress and future challenges. *Science* , *336*(6089), 1671–2.

Vopalensky, P., Pergner, J., Liegertova, M., Benito-Gutierrez, E., Arendt, D., & Kozmik, Z. (2012). Molecular analysis of the amphioxus frontal eye unravels the

evolutionary origin of the retina and pigment cells of the vertebrate eye. *Proceedings of the National Academy of Sciences*, *109*(38), 15383–15388.

Wagner, G. P., & Altenberg, L. (1996). Perspective: complex adaptations and the evolution of evolvability. *Evolution*, *50*(3), 967–976.

Wake, D. B. (1991). Homoplasy: the result of natural selection, or evidence of design limitations? *The American Naturalist*, *138*(3), 543–567.

Walker, J. F., Yang, Y., Moore, M. J., Mikenas, J., Timoneda, A., Brockington, S. F., & Smith, S. A. (2017). Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany*, *104*(6), 858–867.

Wang, P., Khoshravesh, R., Karki, S., Tapia, R., Balahadia, C. P., Bandyopadhyay, A., … Langdale, J. A. (2017). Re-creation of a key step in the evolutionary switch from C_3 to C_4 leaf anatomy. *Current Biology*, *27*(21), 3278–3287.e6.

Wang, X., Gowik, U., Tang, H., Bowers, J. E., Westhoff, P., & Paterson, A. H. (2009). Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. *Genome Biology*, *10*, R68.

Wang, Y., Bräutigam, A., Weber, A. P. M., & Zhu, X.-G. (2014). Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. *Journal of Experimental Botany*, *65*(13), 3567–3578.

Weinreich, D. M., Delaney, N. F., Depristo, M. A., & Hartl, D. L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, *312*(5770), 111–114.

Wendel, J. F., Jackson, S. A., Meyers, B. C., & Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biology*, 17-37.

Werner, G. D. A., Cornwell, W. K., Sprent, J. I., Kattge, J., & Kiers, E. T. (2014). A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms. *Nature Communications*, *5*, 4087.

Whiting, M. F., Bradler, S., & Maxwell, T. (2003). Loss and recovery of wings in stick insects. *Nature*, *421*(6920), 264–7.

Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2006). Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. *The American Naturalist*, *167*(6), 794–807.

Wielstra, B., Arntzen, J. W., van der Gaag, K. J., Pabijan, M., Babik, W., Funk, D., … Carstens, B. (2014). Data concatenation, bayesian concordance and coalescent-based analyses of the species tree for the rapid radiation of triturus newts. *PLoS*

ONE, *9*(10), e111011.

Wiens, J. J., Brandley, M. C., & Reeder, T. W. (2006). Why does a trait evolve multiple times within a clade? Repeated evolution of snakelike body form in squamate reptiles. *Evolution; International Journal of Organic Evolution*, *60*(1), 123–41.

Wiens, J. J. (2011). Re-evolution of lost mandibular teeth in frogs after more than 200 million years, and re-evaluating Dollo's law. *Evolution*, *65*(5), 1283–1296.

Williams, B. P., Johnston, I. G., Covshoff, S., & Hibberd, J. M. (2013). Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis. *eLife*, *2*, e00961.

Williams, B. P., Burgess, S. J., Reyna-Llorens, I., Knerova, J., Aubry, S., Stanley, S., & Hibberd, J. M. (2016). An untranslated cis-element regulates the accumulation of multiple C_4 enzymes in *Gynandropsis gynandra* mesophyll cells. *The Plant Cell*, *28*(2), 454–65.

Williams, P. D., Pollock, D. D., Blackburne, B. P., & Goldstein, R. A. (2006). Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Computational Biology*, *2*(6), e69.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, *8*(3), 206–216.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences of the United States of America*, *6*(6), 320–32.

Wright, S. (1930). Evolution in mendelian populations. *Genetics*: 16(2), 97-159

Xiao, S. (2014). Evolution: the making of ediacaran giants. *Current Biology*, *24*(3), R120-2.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, *15*(5), 568–73.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, *19*(6), 908–917.

Yang, Z., & Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, *25*(3), 568–579.

Yang, Z., & Swanson, W. J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution*, *19*(1), 49–57.

Yoshida, M., Yura, K., Ogura, A., & Furuya, H. (2015). Cephalopod eye evolution was modulated by the acquisition of Pax-6 splicing variants. *Scientific Reports*, *4*(1), 4256.

Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N., & Ma, H. (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications*, *5*, 4956.

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., … Wang, J. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology*, *30*(6), 549–554.

Zhaxybayeva, O., & Doolittle, W. F. (2011). Lateral gene transfer. *Current Biology*, *21*(7), R242-6.

Zinovieva, R. D., Piatigorsky, J., & Tomarev, S. I. (1999). O-Crystallin, arginine kinase and ferritin from the octopus lens. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *1431*(2), 512–517.

Zwieniecki, M. A., & Newton, M. (1995). Roots growing in rock fissures: their morphological adaptation. *Plant and Soil*, *172*(2), 181–187.